ulm university universität
uulm

Universität Ulm | 89069 Ulm | Germany

Fakultät für
Ingenieurwissenschaften,
Informatik und Psychologie
Institut für Neuroinformatik
Direktor: Prof. Dr. Günther Palm

Multiple Classifier Systems in Human-Computer Interaction

Dissertation zur Erlangung des Doktorgrades
Doktor der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Ingenieurwissenschaften, Informatik
und Psychologie der Universität Ulm

vorgelegt von
Martin Benedikt Schels
aus Regensburg

Ulm, 2015

Amtierende Dekanin: Prof. Dr. Tina Seufert

Gutachter: Prof. Dr. Günther Palm
Gutachter: Prof. Dr. Dr. Wolfgang Minker

Tag der Promotion: 3. Juli 2015

Abstract

Making the interaction of human subjects with technical systems more intuitive is a quickly emerging interdisciplinary field of research. One important aspect of this field is the surveillance of the user by the respective system and thus enabling it to estimate distinct user states. Thus, a technical system is enabled to adapt better to the context of the situation and also to the respective needs of the subject. A key factor for this development is the availability of affordable sensory equipment such as cameras and microphones but also physiological measuring devices together with increasing computational power and the according methods of analysis and classification.

The automatic recognition of user states in human-computer interaction poses a great challenge to statistical pattern recognition for several reasons: The measurements of the different sensors are inherently heterogeneous in their technical properties, for example considering sample rates, range of values or resolution. A further issue is that in real world scenarios, the different categories occur often in an imbalanced distribution, which makes it difficult to estimate sound models for the underrepresented classes. Another important aspect is the fact that the true state of a subject is generally not entirely observable from the outside, which makes the design of corpora that study human-computer interaction extremely difficult. This leads in many cases to weakly or subjectively defined class labels by either using human test persons that annotate the collected material manually or by using externally triggered stimuli that are designed to elicit distinct predefined states.

In order to approach the recognition of user states in human-computer interaction, the multi-modal and the temporal properties of the application are exploited in this work. For this purpose, different information fusion architectures based on multiple classifier system approaches and temporal integration techniques are introduced and discussed. Besides this, the incorporation of unlabeled data into the training of the classifier is a compelling issue since one is generally short of training data as described earlier. This work will intro-

duce a partially supervised learning approach, that combines unsupervised and supervised learning in order to extend the amount of usable data. Finally, the problem of imbalanced class distributions is tackled by a class weighting mechanism in the training of support vector machines, which increases the loss for the underrepresented class. The approaches are further extensively evaluated on publicly available multi-modal data collections.

Zusammenfassung

Die Verbesserung der Interaktion menschlicher Benutzer mit technischen Systemen ist eine interdisziplinärer wissenschaftlicher Bereich, der in der jüngeren Vergangenheit eine verstärkte Aufmerksamkeit erfahren hat. Ein wichtiger Teilbereich dieses Forschungsfeldes ist die Beobachtung des Benutzers durch das technische System und die Ableitung vordefinierter Nutzerzuständen daraus. Dadurch kann ein solches technisches System sich besser auf den situativen Kontext und die jeweiligen Bedürfnisse eines menschlichen Interaktionspartners ausrichten. Ein wichtiger Baustein zum Erreichen dieses Ziels ist die Verfügbarkeit von preisgünstiger Sensorik, wie beispielsweise Mikrophone oder Videokameras aber auch von Apparaten zur Aufnahme physiologischer Signale, sowie der gestiegenen Rechenleistung moderner Computer und des methodischen Fortschritts in der Entwicklung intelligenter Algorithmen.

Die automatische Erkennung von Nutzerzuständen in der Mensch-Maschine Interaktion ist aus folgenden Gründen eine große Herausforderung für die bestehenden Methoden der statistischen Mustererkennung: Die verschiedenen Messgrößen aus der einzelnen Sensoren sind stark heterogen in ihren jeweiligen Eigenschaften wie der Abtastrate, den Wertebereichen oder Auflösungen. Eine weitere Problemstellung bei der Bearbeitung von Problemstellungen, wie sie tatsächlich in der praktischen Anwendung auftreten, ist dass die jeweiligen Kategorien üblicherweise nicht gleich-verteilt auftreten, was die Erstellung adäquater Modelle für die unterrepräsentierten Klassen erschwert. Weiterhin ist der tatsächliche Zustand eines Benutzers in einer Interaktion nicht vollständig von außen erfassbar, was die Erstellung von Korpora zur Erforschung diesen Aspekt der Mensch-Computer Interaktion äußerst schwierig. Dieser Umstand führt in vielen Fällen zu nur schwach oder subjektiv definierten Klassenlabels indem entweder menschliche Annoteure die Daten händisch mit Kategorien versehen oder durch die Verwendung von extern getriggerten Stimuli, die bestimmte vordefinierte Zustände elizitieren sollen.

In dieser Arbeit werden die multi-modalen und temporalen Charakteristika

der Mensch-Computer Interaktion zur Erkennung von Nutzerzuständen in dieser Anwendung ausgenutzt. In diesem Rahmen werden verschiedene Architekturen der Informationsfusion, basierend auf den Technologien der Mehrklassifikatorsysteme und der temporalen Integration vorgestellt und numerisch evaluiert. Darüber hinaus ist die Verwendung ungelabelter Daten in der Trainingsphase von Klassifikatoren da in den meisten Anwendungen Trainingsdaten, wie oben beschrieben, nur knapp vorhanden ist. Im Folgenden wird ein teil-überwachtes Lernverfahren vorgestellt, das unüberwachte mit überwachten Verfahren kombiniert um die Gesamtheit der verwendbaren Daten zu vergrößern. Ein weiterer Punkt der Arbeit ist die Analyse unbalancierter Klassenverteilungen durch die Einarbeitung eines Gewichtungsverfahrens in die Trainingsphase von Support Vektor Maschinen mittels einer modifizierten Fehlerfunktion, die die Gewichtung der unterrepräsentierten Klasse erhöht. Die numerische Evaluation der Methodik wird im Folgenden auf verschiedenen öffentlich verfügbaren multi-modalen Datensätzen durchgeführt.

Contents

1 Introduction

1.1 Multi-modal Classification Architectures

The development of technical systems which are suited to make interactions of humans with technical systems more intuitive is a quickly emerging interdisciplinary field of research. One important aspect of this field is the surveillance of the user by the respective system in order to enable it to estimate distinct user states. Knowledge on these states allows the technical system to adapt better to the respective needs of a subject in the context of the situation. A key factor for these developments is the availability of affordable sensory equipment such as cameras and microphones, but also physiological measuring devices that are attached to a subject's skin like electrodes to quantify the skin conductance or an electromyograph to measure muscle contractions. This development is also pushed forward by an increase of technical capabilities for instance in an increasing miniaturization of computer hardware or in computational power, which comprises not only faster CPUs but also an increase of the capacity of storage devices and the available memory of computers. This makes it feasible to create intelligent computer programs that provide an additional value for the user, for example to better consult, monitor and teach the subject in the course of certain tasks. In order to study this subject multifaceted multi-modal corpora in the context of human-computer interaction are required and are now becoming feasible that allow to study statistical models for their analysis and to develop classifiers for the prediction of complex user states. An illustrative example for the multifaceted and comprehensive nature of possible multi-modal corpora of human-computer interaction is shown in Figure 1.1.

However, the automatic recognition of user states in human-computer interaction still poses great challenges to statistical pattern recognition techniques for several reasons. The measurements of the different sensors are inherently heterogeneous in their technical properties, for example considering sample rates

or resolution. In addition to that, the features that are computed from the raw data are very heterogeneous in their technical properties such as the respective time granularity but also in their individual informativeness for a classifier. Furthermore, the different categories are not only sparsely available but they occur often in imbalanced class distribution in real world scenarios. These circumstances make it particularly difficult to derive sound models for the underrepresented classes. Another important aspect is the fact that the true state of a subject is generally not entirely observable from the outside, which makes the design of corpora, that study human-computer interaction extremely difficult. This fact leads in many cases to weakly or subjectively defined class labels by either using human test persons, that annotate the collected material manually or by using externally triggered stimuli that are designed to elicit distinct predefined states.

In order to approach the recognition of user states in human-computer interaction, the multi-modal and the temporal properties of the application are exploited in this work. For this purpose, different information fusion architectures based on multiple classifier system techniques and temporal integration approaches are introduced and discussed. In addition to this, the incorporation of unlabeled data into the training of the classifier is a compelling issue as one is generally short of training data as described earlier.

This work will introduce a partially supervised learning approach, that combines unsupervised and supervised learning in order to extend the magnitude of usable data. Finally, the problem of imbalanced class distributions is addressed using a class weighting mechanism in the classifier training, which increases the loss for the underrepresented class.

1.2 Affective States in Human-Computer Interaction

Novel human-computer interfaces make use of the detection of human emotions as described for example by Picard et al. (2001b) in her very forward-looking book. This is also reflected in past and ongoing interdisciplinary research projects like the EU project *Semaine*[1], the *Companions Project*[2], or the Transregional Collaborative Research Centre SFB/TRR 62 *Companion-Technology for Cognitive Technical Systems*[3], that is established at the universities of Ulm and Magdeburg.

The research in the field of human emotions is very prominently emanated from Darwin (1978) and Ekman and Friesen (1978), amongst others, who de-

[1] `www.semaine-project.eu/`
[2] `www.companions-project.org`
[3] `www.sfb-trr-62.de/`

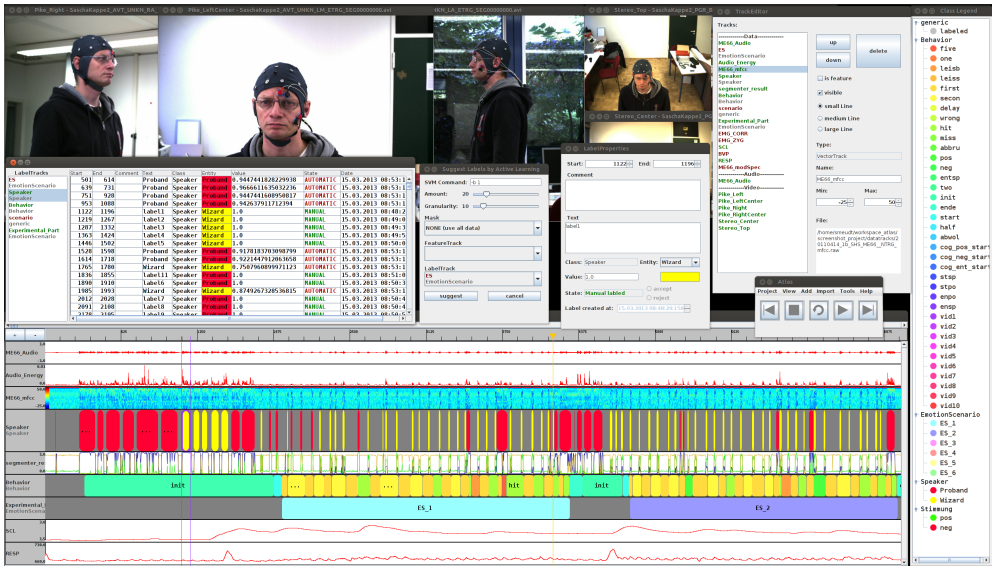


Figure 1.1: *Depiction of a multi-modal recording of human-computer interaction. It shows how multifaceted the different channels and annotations can be in this application. The picture contains multiple camera views on the subject, representations of the recorded audio signal as the respective energy and as mfcc coefficients and two different physiological signals i.e., skin conductance and respiration. Furthermore, various annotations to the interaction, for example speaker turns or certain subject behavior, are displayed in the figure as colored blocks. Taken from (Schels et al., 2013a).*

fined six prototypical basic emotions that are supposed to be universal for all human beings. These emotions, which are often called "the big six" are: "happiness", "anger", "fear, "sadness", "surprise" and "disgust". The work of Ekman and co-workers provides well defined categories, that are well established in the literature. However, they are unfortunately not likely to occur in a full blown shape in human-computer interaction — if at all.

An alternative approach is modeling the human emotions in a continuous, multivariate space, where each point in space resembles an affective state. A prominent example for a continuous emotional space is the so called PAD space (Russell, 2003; Russell and Barrett, 1999), where the upper case letters stands for "pleasure", "arousal" and "dominance". In short, the different dimensions are defined in the following way (Mehrabian, 1996): the term pleasure is described as "positive vs. negative affective states", arousal as "mental alertness and physical activity" and dominance "as a feeling of control and influence over one's surroundings and others". However, these dimensions are far from being fixed and different versions of the concepts are possible, for example Fontaine et al. (2007) define four different dimensions, that they call "arousal", "expectancy", "power" and "valence".

A categorization of human emotions, that is in between the concept of basic

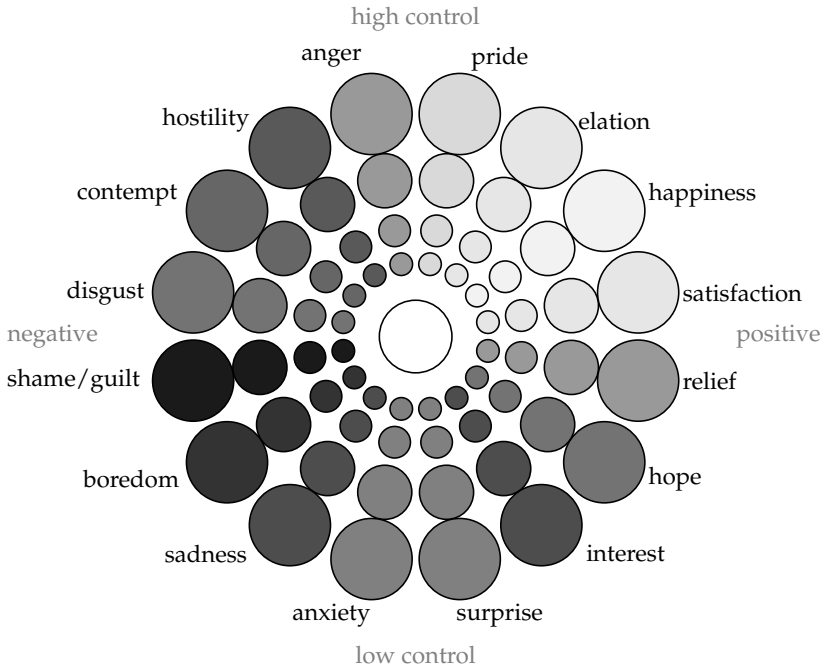


Figure 1.2: *The Geneva Emotion Wheel enables to choose categories from a circular label system with different intensities. Adapted from (Scherer, 2005).*

emotions and a continuous space by allowing intensities for a greater number of categories is the Geneva emotional wheel by Scherer (2005). Hereby the different categories are arranged in a circular layout and the intensity of a label can be encoded by the distance from the center of the circle.

An important issue for the creation of affective data collections is the elicitation of emotions in a distinct scenario. There is a huge amount of literature on this matter and different approaches of induction of emotion were implemented in the past. The most prominent of these are the acting of facial expressions, the narration of stories, the presentation of distinct movie clips or images or the difficulty of a given task to solve (Kierkels et al., 2009; Lang et al., 1993; Lisetti and Nasoz, 2004; Soleymani et al., 2008; Stemmler, 1989; Wright and Dill, 1993). A comprehensive summary of different studies in this context is provided by Lisetti and Nasoz (2004).

1.3 Acted versus Real-World Corpora

An important question in the context of the research for the recognition of human emotions or user states in human-computer interaction is the design of matching corpora, that enable the development of suitable models. There

Figure 1.3: *Sample images from the acted Cohn-Kanade Comprehensive Database for Facial Expression Analysis displaying different basic emotions (Kanade et al., 2000). The last image of a sequence resembles the full blown facial expression.*

are two general approaches in the literature for the recording of such kinds of corpora: They are either recorded in a controlled environment with many constraints for the actors, that conduct predefined actions or alternatively the recordings can be conducted with naïve test persons in an unconstrained interaction with an interlocutor.

There are various prominent examples for acted emotional corpora like the Berlin database of emotional speech (Emo-DB) introduced by Burkhardt et al. (2005), which resembles short sentences, that are spoken by professional actors or the Cohn-Kanade Comprehensive Database for Facial Expression Analysis (Kanade et al., 2000), where subjects are portrayed at a frontal view in short video clips, which are instructed to perform certain facial expressions.

An example for an acted audio visual data collection is the eNTERFACE corpus (Martin et al., 2006), where the subjects are confronted with a situation that is supposed to elicit a specific emotional reaction in the basic emotion categories. Finally, the *Magdeburger Prosodie Korpus* by Wendt and Scheich (2002) is a collection of pseudo words that are uttered by actors in the different basic emotions in order to reduce correlations with the semantics of an existing word.

The automatic recognition based on these kinds of data collections has been proven to be very successful in the literature rendering high recognition rates with many classes (compare e.g., Scherer et al., 2007; Schels et al., 2009; Schmidt et al., 2010; Schuller et al., 2010; Vlasenko et al., 2007; Wagner et al., 2007). However, as mentioned before, the applicability of these results in real world scenarios is rather limited.

Hence there are several successful attempts to design affective corpora, that capture a more realistic interaction with technical interlocutors. Many of these corpora are designed under the wizard of Oz (Woz) paradigm (Kelley, 1983), where the technical system, that interacts with the subject, is directly controlled by the experimenter, that guides the subject deliberately through the predefined experimental procedure.

An example for this kind of data collection is the EmoRec corpus that has been recorded at Ulm University (Walter et al., 2011; Hrabal et al., 2012), where a subject solved multiple rounds of a puzzle game where different stimuli are presented to the subject in order to elicit defined emotional reactions.

A further corpus that can be found in the literature is the AVEC 2011 corpus, that is derived from the SEMAINE project[4] (Schuller et al., 2011). The data has been recorded from subjects, that conduct a relatively unconstrained dialog with an artificial avatar. These avatars are designed to have different characters in order to provoke distinct reactions from the test persons.

Another corpus that comprises recordings from a Woz experiment is the so called "Last Minute" corpus which is collected at the University of Magdeburg (Rösner et al., 2012). For this corpus, the subject uses a dialog system in order to pack a suitcase for a voyage for which the final destination is revealed only at the very end of the experiment. This procedure is supposed to induce a stressful experience for the human interlocutor as the final destination does normally not match the items the user chose in his suitcase.

The PIT corpus[5], that is also recorded at the Ulm University, adds a further human interlocutor to the experimental setting, that is restricted to a two party interaction in the previous examples (Strauss et al., 2008; Scherer, 2011). The technical system is supporting one main human user to find a suitable restaurant for a meeting with the second person.

Conducting classification experiments on these corpora is a much more challenging task for a statistical classifier, and much lower classification rates are reported in the literature (Schels et al., 2012a; Krell et al., 2013; Schuller et al., 2011). The EmoRec and the AVEC 2011 corpora, that are labeled in a multivariate emotional space are investigated in this thesis.

1.4 Outline of the Thesis

An outline of the thesis is provided in this following section. Chapter 2 describes the general methodological surroundings of the work in statistical pat-

[4] http://www.semaine-project.eu/

[5] http://www.uni-ulm.de/in/pit.html

tern recognition. It starts in Section 2.1 by introducing basic unsupervised learning techniques. Both, prototypical and probabilistic approaches are described. Also, a short description of hierarchical cluster algorithms is given.

Basic supervised machine learning approaches are subsequently described in Section 2.2. This comprises linear classification approaches like the linear least squares classifier and the perceptron. Also, nonlinear extensions like the multilayer perceptron and the support vector machine with non-linear kernel approaches are briefly described. Furthermore tree-based learning techniques are introduced in this section together with the random forest algorithm as an extension based on ensemble learning techniques.

An important building block of this work is the classification under uncertainty. This technique is described in Section 2.3. This comprises a review of the techniques to measure uncertainty of classifier decisions. Further, the directions for the processing of uncertain or fuzzy decisions and their incorporation in classification architectures are outlined.

Another major block of the thesis is based on multiple classifier systems. This technique is reviewed in Section 2.4. One important question in this context, that will be addressed, is how to construct a meaningful classifier ensemble. This comprises different approaches to generate diverse classifier teams, which is an important requirement for an ensemble to improve over the best member of the team. Further, the combination of individual classifiers will be addressed: this comprises fixed rule combiners and trainable fusion mappings, that are capable to render more complex combination architectures. Another important field of research in the context of multiple classifier systems that will be addressed here, is the selection of informative classifiers.

Section 2.5 will discuss the literature on the basic techniques of the incorporation of unlabeled data into the training of a statistical classifier, i.e., the general field of partially or semi-supervised learning. The approaches, that are described there comprise semi-supervised learning from generative models, self-training and co-training. Furthermore, a brief overview over the transductive support vector machine will be given.

In Chapter 3, the applications of human-computer interaction, that are used in this work to evaluate the developed classification approaches are introduced. This comprises mainly three corpora of recordings of human subjects, that are interacting with a technical system. The first data collection is the EmoRec wizard of Oz experimental corpus, which is described in Section 3.1. In this corpus the user is instructed to solve a task in a voice controlled human-computer interaction scenario, and the experimenter guides the subject through a manifold of emotional states by presenting distinct stimuli to the subject. A distinguishable feature of this corpus is the circumstance, that not only the audio and video channels, but also a manifold of different physiological signals are

recorded.

The AVEC 2011 corpus is described as a second relevant data collection for this work in Section 3.2. In this corpus, the subject is conducting a conversation with an affectively colored virtual agent. Other than in the EmoRec corpus, the emotional labels for classification are annotated by human raters. This annotation is conducted by human labelers in four different labels, namely "arousal", "expectancy", "power" and "valence".

The last corpus, that is used in this work is the Pascal 2 mind reading competition data set, which is described in Section 3.3. Here the EEG signal of a subject is recorded while visual stimuli are presented in a quick sequence. The task for the subject is to recognize an underrepresented image class, which follows a so-called oddball paradigm. Thus, a P300 EEG pattern is elicited, which is subject for detection in this application.

In Chapter 4 the methodical contributions of this thesis are outlined. The temporal characteristics of audio-visual emotion recognition in non-acted corpora is investigated in Section 4.1. Further, the classifier fusion in temporal processes is investigated with respect to on which stage, the multi-modal classifier combination is optimally conducted. This Chapter is based on the following publications: (Schels et al., 2014a, 2013a,b, 2012a, 2009; Schels and Schwenker, 2010; Glodek et al., 2013a, 2012b, 2011; Meudt et al., 2013; Scherer et al., 2012a, 2011; Schmidt et al., 2010; Walter et al., 2011; Schwenker et al., 2010).

In Section 4.2, a novel approach for the partially supervised learning is introduced. Here an unsupervised density estimation is used to compute a new representation for the data. In this step additional unlabeled data is used to render a better estimate of the distribution of the data. Based on this, a supervised classifier is constructed. This method and its evaluation is published in (Schels et al., 2014b, 2012b, 2011).

A third contribution is given in Section 4.3. Here, a support vector machine is adapted to be able to reflect highly imbalanced class distributions by integrating class-weights into the training process. The adapted SVM and its numerical evaluation is published in (Schels et al., 2013c, 2010).

The methods are evaluated in Chapter 5 in the context of human-computer interaction using the data collections, that are mentioned earlier. For this purpose, the evaluation of a statistical classifier is an important issue, which is discussed in Section 5.1. The multi-modal fusion algorithms are evaluated using the AVEC 2011 corpus and the audio-visual part of the EmoRec data collection in Section 5.2. In Section 5.3, the physiological parts of the EmoRec corpus are used to evaluate the partially supervised approach. It is especially qualified for the evaluation of approaches using unlabeled data as the extraction of physiological features requires generally longer time scales than the audio-

visual parts. This circumstance leads to fewer feature vectors per time step and makes it even more compelling to use additional data for the training of a classifier. Finally, the adapted SVM is evaluated in Section 5.4 in the context of the detection of the P300 EEG pattern.

In Chapter 6, a summary and discussion of the work and its integration into the broader picture is given. Also, further future directions will be discussed there. Finally, the summary of the major contributions of the thesis is provided in Chapter 7.

2 Basic Methodical Principles

2.1 Basic Unsupervised Learners

In this thesis techniques of unsupervised learning are integrated in the training of a classifier. This is aiming at incorporating unlabeled data into the classification process. Under the terms unsupervised learning or clustering, different approaches to statistical learning from data without a distinct teacher signal are assembled. A crucial aspect, which has a strong influence on the outcome of the approaches is the definition of an appropriate dissimilarity measure. These measures rate the degree of alikeness of data samples and they are addressed in Section 2.1.1. The remainder of the chapter introduces different popular clustering techniques, namely the k-means algorithm for sum of squares approaches, hierarchical clustering techniques and mixture models in sections 2.1.2 – 2.1.4. Finally a brief discussion, summarizing the most important keypoints, is conducted in Section 2.1.5.

2.1.1 Dissimilarity Measures

Dissimilarity measures assess the degree of alikeness (or unlikeness) of two samples (Theodoridis and Koutroumbas, 2009). Formally a dissimilarity measure is given by a function defined on pairs of samples from a data set X, that are mapped to a real number:

$$d : X \times X \to \mathbb{R}.$$

Further a minimal value $d_0 < \infty$ is required (typically $d_0 = 0$), that is assigned for the distance for any data sample $x \in X$ to itself:

$$\exists d_0 \in \mathbb{R} : -\infty < d_0 \leq d(x, y) < \infty, \quad \forall x, y \in X$$

$$d(x, x) = d_0, \quad \forall x \in X.$$

A dissimilarity measure is called a metric dissimilarity measure if the following two equations hold: The minimal value d_0 is returned if and only if the dissimilarity is computed between a sample and itself:

$$d(x,y) = d_0 \Leftrightarrow x = y.$$

Further the measure must suffice the triangle equation, which implies that the shortest distance between two points is a straight line:

$$d(x,z) \leq d(x,y) + d(y,z) \quad \forall x,y,z \in X.$$

Prominent weighted metric dissimilarity measures for l dimensional real valued vectors are defined from the following equation by the choice of the parameter p:

$$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{l} w_i |x_i - y_i|^p}.$$

Further the parameters $w_i \geq 0$ for all $i = 1,\ldots,l$ can be chosen to allow the weighting of certain dimensions accounting to a concrete application. However in most of the literature and applications, the dimensions are equally weighted and the w_i are uniformly set to 1.

For $p = 1$, the weighted Manhattan distance is given by

$$d_1(x,y) = \sum_{i=1}^{l} w_i |x_i - y_i|.$$

This corresponds to a summation of the (weighted) distances in the different dimensions.

Setting $p = 2$ and $w_i = 1$ for all $i = 1,\ldots,l$, yields the famous Euclidean distance:

$$d_2(x,y) = \sqrt{\sum_{i=1}^{l} (x_i - y_i)^2}.$$

Most of the applications use this very intuitive measure to compute the dissimilarity of samples.

A further prominent dissimilarity measure, that is derived from the above form is the weighted l_∞ norm, which is defined for $p = \infty$ as follows:

$$d_\infty(x,y) = \max_{1 \leq i \leq l} w_i |x_i - y_i|.$$

This is equal to the largest absolute value of the differences in the individual dimensions.

Analogously to dissimilarity measures, a similarity measure is defined using a maximal value of similarity s_0. The most prominent similarity measures are the inner product and the correlation coefficient.

2.1.2 k-means Clustering

The k-means algorithm is an iterative approach to minimize the following term (Bishop, 2006; Jain et al., 1999):

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - c_k\|^2.$$

In the above equation the variable $r_{nk} \in \{0, 1\}$ equals 1 if the n-th data point x_n is assigned to the k-th cluster center c_k. The algorithm hence aims at finding values for r_{nk} and c_k, that minimize the term J.

This problem can be solved by iteratively computing assignments of the data to the nearest cluster and re-computing the clusters based on the new assignment until some stopping criterion is met.

Let $c_j(t)$ be the j-th cluster center for $j = 1, \ldots, k$ at the t-th iteration and $C_j(t)$ the set of data samples, that are members of this cluster. In the beginning of the algorithm, a suitable initialization for the clusters has to be set and hence the following computations have to be conducted:

$$c_j(t) = \frac{1}{\|C_j(t)\|} \sum_{x_\mu \in C_j(t)} x_\mu, \quad j = 1, \ldots, k. \tag{2.1}$$

Based on this, new memberships are computed by choosing the cluster represented by the prototype, that is closest to the respective samples:

$$C_j(t+1) = \left\{ x_\mu \,\middle|\, \|x_\mu - c_j(t)\| = \min_{j=1,\ldots,k} \|x_\mu - c_j(t)\| \right\}, \; j = 1, \ldots, k. \tag{2.2}$$

Equations 2.1 and 2.2 are then alternately evaluated until a stopping criterion, such as a maximal amount of iterations or the adaption of the centers is appropriately small, is met. The algorithm iteratively computes k arithmetic prototypes as means of the vectors in a cluster. An illustrative example for the partition of a two dimensional data set that is clustered into four partitions is shown in Figure 2.1.

An important issue for the k-means algorithm is the choice of the initial values for the partitions as the approach is vulnerable to local optima. A naïve initialization of the parameters is for example to sample randomly k data points from the training set as centers for the first iteration. However, the literature states that using k randomly chosen partitions of the data to compute the first cluster centers from is a better strategy (Theodoridis and Koutroumbas, 2009). Other strategies include a repeated execution of the algorithm in order to find an optimal clustering.

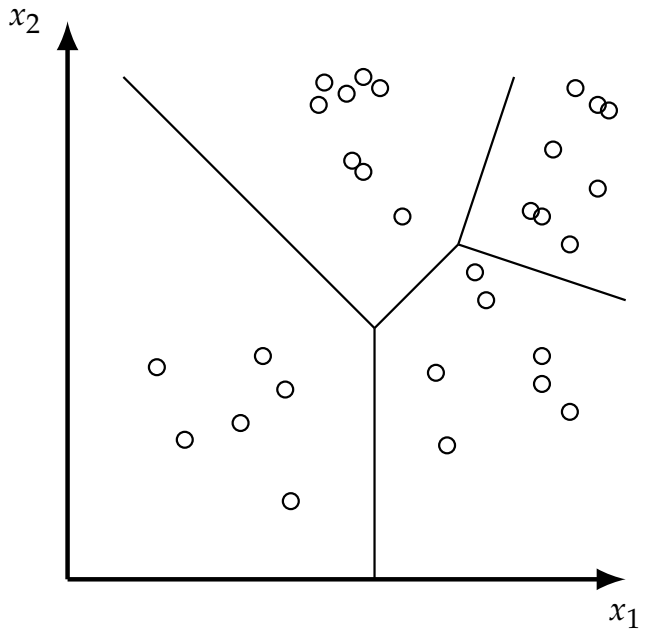


Figure 2.1: *Example of partitions that are computed with k-means for 4 centers.*

A further input to the algorithm, that is subject to an optimization is the number of cluster centers (Fraley and Raftery, 1998; Berkhin, 2002). Again, evaluating different parameter settings in multiple runs to find an optimum is a widely applied strategy (Theodoridis and Koutroumbas, 2009).

The *k*-means algorithm is a comparably elementary clustering approach, which makes it feasible in many applications. However, due to the computations of the averages, the result is sensitive for outliers in the data. This leads naturally to extensions of the algorithms, such as the *k*-medoids algorithm, where individual data points from the training set are used as prototypes. This reduces this sensitivity at the cost of an increased computationally complexity.

2.1.3 Hierarchical Clustering

Rather than computing a single partitioning of the given data as described earlier, multiple so-called nested clusterings are computed in hierarchical clustering (Johnson, 1967). Hierarchical clustering constructs a tree structured hierarchy of clusterings, that is defined the following way: Let $X = \{x_1, \ldots, x_N\}$ be a set of N data points. Based on this, a partition of the data with m clusters is defined as $\mathcal{C} = \{C_1, \ldots, C_m\}$, with $C_j \subseteq X$, $C_j \neq \emptyset$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. Further, two clusterings, say $\mathcal{C}_i$, which comprises k clusters and $\mathcal{C}_j$, having k' clusters with $k > k'$, are called nested ($\mathcal{C}_i \sqsubset \mathcal{C}_j$), if every cluster of

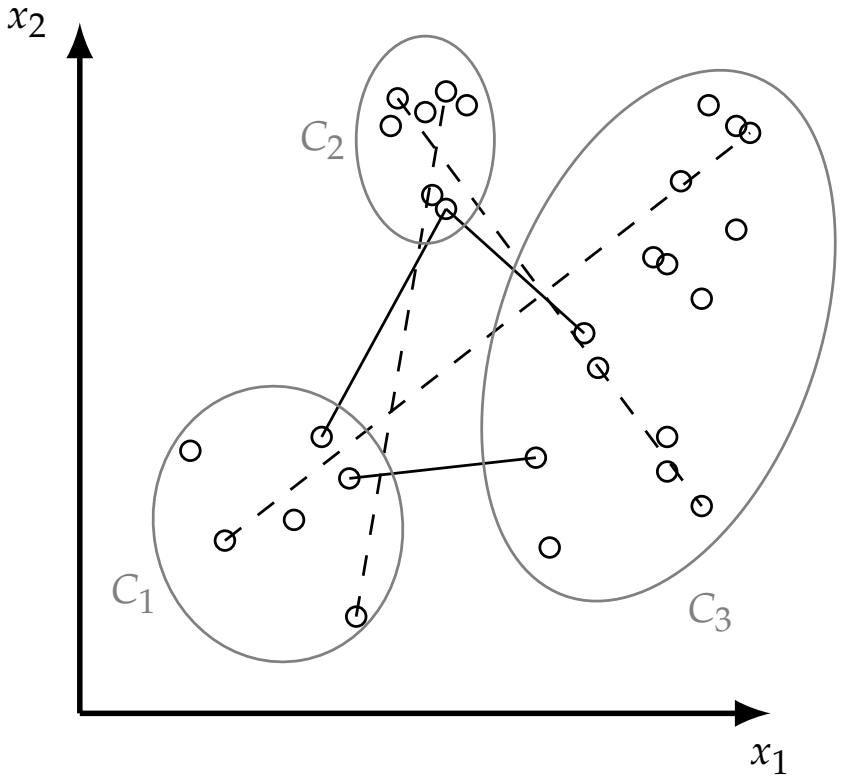


Figure 2.2: *Complete (dashed lines) and single linkage (solid lines) distances for a sample data set with three clusters.*

$\mathcal{C}_i$ is a subset of a cluster in $\mathcal{C}_j$ (Theodoridis and Koutroumbas, 2009). Hence, the hierarchical approaches construct all possible partitions using pre-defined distance measures including the trivial clusterings, where all samples are in the same cluster and where every sample has its own cluster.

The two main techniques to compute a hierarchical clustering are the agglomerative clustering and the complementary divisive clustering. In divisive clustering, the nested clusterings are computed based on the clustering, that comprises all available samples $\mathcal{C} = \{X\}$. In every iteration of the algorithm a cluster is split into two separate clusters, such that a dissimilarity measure for the resulting partition is maximized. Contrary, the agglomerative clustering versions start with partitions where every data sample is assigned to its own cluster $\mathcal{C} = \{\{x_1\}, \ldots, \{x_N\}\}$. By iteratively combining the most similar clusters, partitioning is coarsened step by step until one big cluster is formed including every data point (Theodoridis and Koutroumbas, 2009).

For the agglomerative clustering approaches the similarity between two clusters is measured with certain distance measures. A manifold of different measures are proposed in the literature that are mostly derived by choosing differ-

ent parameters a_i, a_j, b and c for the following term (Webb, 2002):

$$d_{i+j,k} = a_i d_{ik} + a_j d_{jk} + b d_{ij} + c|d_{ik} + d_{jk}|. \tag{2.3}$$

This defines the distance $d_{i+j,k}$ of the cluster C_k to the union of the clusters C_i and C_j based on the distance d_{ik} of clusters C_i and C_k, the distance d_{jk} C_j and C_k, the distance of the clusters that are to be combined d_{ij}. Based on these distances, the order of combination of the clusters is determined in different approaches.

By choosing $a_i = a_j = \frac{1}{2}$, $b = 0$ and $c = -\frac{1}{2}$ the so-called single linkage approach is rendered. Thus, the distance between two clusters equals the distance between their respective closest members. Hence only a single link between the clusters is required. Unfortunately, the method is vulnerable to so-called chaining effects: If — in principle distant — clusters are connected via a trace of chaining samples, it is possible that they are fused before other clusters are combined, even if intuitively different combinations are favorable (compare also Webb 2002, Figure 10.5). In other words elongated clusters are preferred (Theodoridis and Koutroumbas, 2009).

The somewhat complementary approach is the complete linkage algorithm (Defays, 1977). There, the distance between two clusters is computed using the samples in two clusters, that are the most distant from each other. The parameter setting for Equation 2.3 to render a complete linkage approach is $a_i = a_j = \frac{1}{2}$, $b = 0$ and $c = \frac{1}{2}$. The complete linkage approach favors more compact partitions than the single link clustering. Hence, if such a behavior is desirable, this algorithm is preferable (Theodoridis and Koutroumbas, 2009).

An illustrative example for the single linkage and the complete linkage distance calculations is sketched in Figure 2.2. The two dimensional data set is already partitioned into three clusters. The single link approach uses the distances of nearest neighbors of the respective clusters for the computation of the distances (solid line), whereas the complete linkage uses the samples, that have the farthest distance (dashed lines) (Webb, 2002).

In principle, many different approaches are thinkable and described in the literature (Hartigan, 1975). For example, the centroid of each cluster can be used for the computation of distances. This is rendered by setting $a_i = \frac{n_i}{n_i+n_j}$, $a_j = \frac{n_j}{n_i+n_j}$ and $b = -\frac{n_i n_j}{(n_i+n_j)^2}$ and $c = 0$. Here, n_i and n_j are the numbers of samples in clusters i and j, that are combined. A further prominent approach is based on minimizing the variance $a_i = \frac{n_i+n_k}{n_i+n_j+n_k}$, $a_j = \frac{n_j+n_k}{n_i+n_j+n_k}$ and $b = -\frac{n_k}{n_i+n_j+n_k}$ and $c = 0$ (Ward, 1963). Approaches like the group average and the median hierarchical clustering algorithms are described for example in (Webb, 2002) or (Theodoridis and Koutroumbas, 2009).

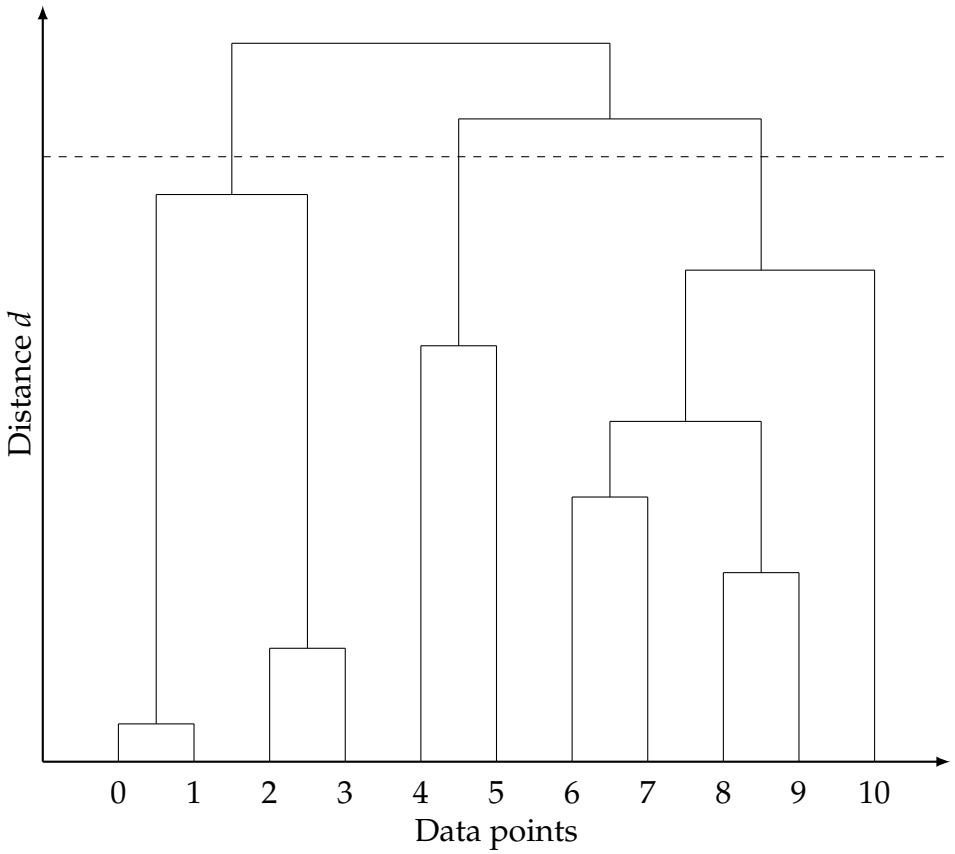


Figure 2.3: *Dendrogram for the single linkage algorithm with a cut-off level for three clusters marked with a dashed line.*

A special feature of hierarchical clustering is the possibility to visualize the cluster tree using so-called dendrograms. A dendrogram is a tree diagram, where every data sample is assigned to a leaf of the tree in the x-axis (compare Figure 2.3). On the y-axis, the distance of the clusters is displayed. If two clusters are fused at a distance d, an edge is drawn between the two clusters at $y = d$, which is the new representative for the new cluster. The number of clusters in a data set can be relatively intuitively determined. For example in Figure 2.3 the dashed horizontal line marks a clustering of eleven data points with three partitions.

2.1.4 Gaussian Mixture Models

A probabilistic approach to clustering is the Gaussian mixture model (GMM) (Bishop, 2006; Kriegel et al., 2011). The clusters are described in this approach using multiple probability density functions, that are implemented using multivariate Gaussians.

A Gaussian mixture model is formally defined as a linear superposition of K Gaussians $\mathcal{N}(\mu_k, \Sigma_k)$ with means μ_k and covariances Σ_k (Bishop, 2006):

$$p(x|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k).$$

The π_k form the weights of the individual Gaussians with the following constraints:

$$0 \leq \pi_k \leq 1; \qquad \sum_{k=1}^{K} \pi_k = 1.$$

A Gaussian distribution is completely defined by its mean and the covariance matrix:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{((2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}})} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

Here l denotes the dimensionality of the data. A simple example for a one dimensional Gaussian mixture model with two components is shown in Figure 2.4. The single Gaussians F (blue) and G (red) are combined using $0.7 \cdot F(x) + 0.3 \cdot G(x)$ to form the darkish yellow probabilistic density function in Figure 2.4.

Gaussian mixture models are commonly trained using the so-called expectation maximization (EM) algorithm (Dempster et al., 1977). As the name suggests, the algorithm is composed of alternations of the expectation and the maximization steps. In the expectation step the current model is evaluated

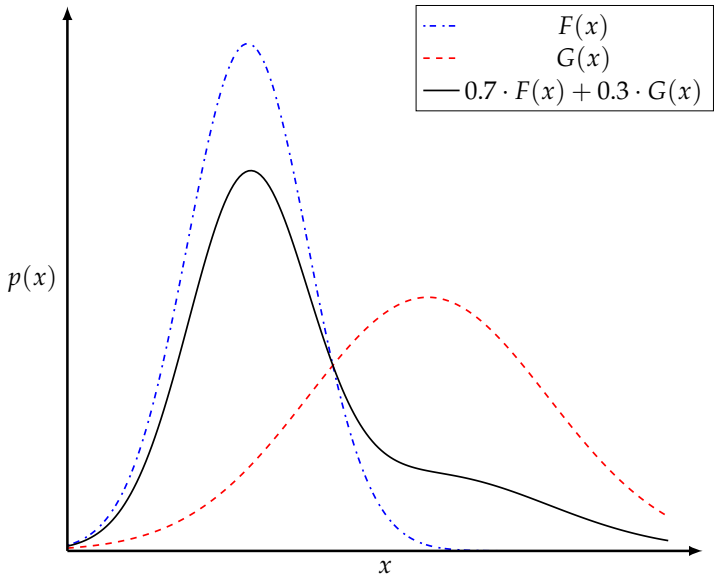


Figure 2.4: *A Gaussian mixture model with two mixtures in one dimension.*

with the given data to compute the posterior probabilities $p(x|\pi_i, \mu_i, \Sigma_i)$ for the components. The outcome is then used to iteratively re-estimate the parameters of the mixture model.

After the initialization of a model with K components, the responsibility γ for a data point x_n of the k-th component is computed (E-Step):

$$\gamma(z_{nk}) = p(z = k|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}.$$

Based on γ, the model is re-estimated using the following equations (M-Step):

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

$$\text{with} \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

Hence the new centers μ_k^{new} of the Gaussians are computed from the average of the data weighted by the responsibility, that is assigned to the particular com-

ponent. The new covariance matrix Σ_k^{new} is computed analogously by computing a standard covariance but with a weighting mechanism with respect to the responsibility of the component for the data samples. The new weights for the components π_k^{new} are computed using the over-all normalized responsibility of the individual component.

The E and M-steps are iterated until a stopping criterion is met. This is often based on the log likelihood of the data for the present model, which is computed as follows:

$$\ln p(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{n=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}.$$

If the likelihood decreases beyond a certain threshold during the iterations, the algorithm terminates. Alternatively a maximum number of runs of the approach can be defined. Combinations of both strategies are also commonly implemented.

The EM-algorithm is closely related to the *k*-means algorithm described in Section 2.1.2. The *k*-means is commonly used to find a proper initialization for the means of the Gaussian mixture components. This should mitigate the general liability to local minimums. A further issue in the context of the GMM is that the number of components have to be determined in advance (Fraley and Raftery, 1998). A popular technique to find an optimal model is to conduct multiple runs with different parameters and to choose the optimal, for example with respect to the likelihood.

In order to mitigate the computational complexity, different constraints can be applied to the covariance matrix of the models. One possible approach is to allow only entries on the diagonal of the matrix and the rest of the matrix is set to zero. Thus, the variance of the resulting Gaussians are bound to the axis and skewed probability densities are not possible. In order to further constrain the covariance matrix, all entries of the diagonal can be restricted to a single number. This constraint is called a spherical covariance matrix as the resulting Gaussian has hence the shape of a hypersphere in the feature space.

A method to improve the probability densities under the utilization of ensembles of GMM has recently been published by Glodek et al. (2013b). Different configurations for the models, for the covariance and the number of centers are used in this approach in order to integrate the individual errors out in the combination of the single models. It has been shown experimentally, that this technique can improve over other approaches for the estimation of probability density functions.

2.1.5 Discussion

In this chapter, unsupervised learning or clustering is described. Based on the definition of a dissimilarity measure in Section 2.1.1, the k-means algorithm is introduced in Section 2.1.2. Further, hierarchical clustering is discussed in Section 2.1.3 together with the concept of the dendrogram, which is a powerful means to visualize clustering results. The Gaussian mixture model is described in Section 2.1.4 as a clustering technique based on probabilistic models. The EM-algorithm is commonly used for the training of the model.

This chapter covers only a small part of the rich literature on unsupervised learning. Other prominent techniques are for example the learning vector quantization (Kohonen, 1998) and the neural gas clustering (Martinetz and Schulten, 1991). A further unsupervised technique, that is commonly used for dimensionality reduction is the principal component analysis, that is conducted by computing the eigenvectors of the covariance matrix of the data. These eigenvectors span a new basis for the given data. The different approaches are extensively discussed in the textbooks, for example in (Webb, 2002; Bishop, 2006; Theodoridis and Koutroumbas, 2009) only to mention some.

There is a manifold of different applications for the described unsupervised learning (Theodoridis and Koutroumbas, 2009).

- The grouping of entities can be used for the generation and testing of hypotheses.

- A prominent application is data reduction: For example if data is conveyed over a slow channel it can be beneficial not to transmit every data sample, but it may be enough send the index of the nearest prototype.

- A further application is the generation of so-called codebooks. Here, different clusters are formed to characterize a distinct concept. Constructing such codebooks for multiple concepts allows to conduct a classification process.

In this work unsupervised learning will be used as a means to implement a partially supervised learning process as described in Section 5.3. This comprises an unsupervised preprocessing step, that incorporates unlabeled data in order to conduct a re-encoding of the labeled data, that is henceforth used for a supervised classification.

2.2 Basic Supervised Learners

For supervised machine learning techniques a teacher signal is added to the training samples. This makes it feasible to construct classifiers, that assign categorical labels to patterns, that are not a member of the training set. A statistical classifier is a function f, that maps data samples $x \in X$ to one of C predefined categories or labels $y \in Y$. Hence, f is formally defined as follows:

$$f : X \to Y.$$

Thus, the input space X is subdivided into multiple regions Ω_i, $i = 1, \ldots, C$, where the decision for the samples within the region is made in favor for class y_i. The boundaries of the Ω_i are called decision boundary (Webb, 2002).

In order to obtain such a function, labeled data samples (x_i, y_i) are presented to the model. For the optimization to a given data set, a so called loss function is defined, that is minimized during the training of the classifier. The loss function is defined via the so called loss matrix (Webb, 2002; Bishop, 2006):

$$\lambda_{ji} = \left(\text{ cost of } f(x) = y_i, \text{ when } x \in y_j \right)_{j,i=1,\ldots,C}.$$

Typically, the cost for correct classifications is zero and the cost for a misclassification is greater than zero. In principle different confusions can be weighted differently in this context.

Based on this, the conditional risk to assign the pattern x to the class y_i is defined as follows (Webb, 2002):

$$l^i(x) = \sum_{j=1}^{C} \lambda_{ji} p(y_j|x).$$

Here $p(y_j|x)$ is the conditional probability of class y_j given the data point x and C is the number of classes. Integrating over the regions $\Omega_1, \ldots \Omega_C$ returns the over-all expected costs:

$$r = \sum_{i=1}^{C} \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(y_j|x) p(x) dx.$$

A typical process for the construction and evaluation of a statistical classifier is depicted in Figure 2.5. The input samples are conducted through an application-specific preprocessing and feature extraction step (compare Chapter 3). Based on the respective feature representations, the classifier is constructed for example based on the techniques described in the following sections. Above the training procedure the testing protocol is depicted, which is separated from the training classifier by the means of the input samples. Its specific steps are however determined in the respective training steps of the figure.

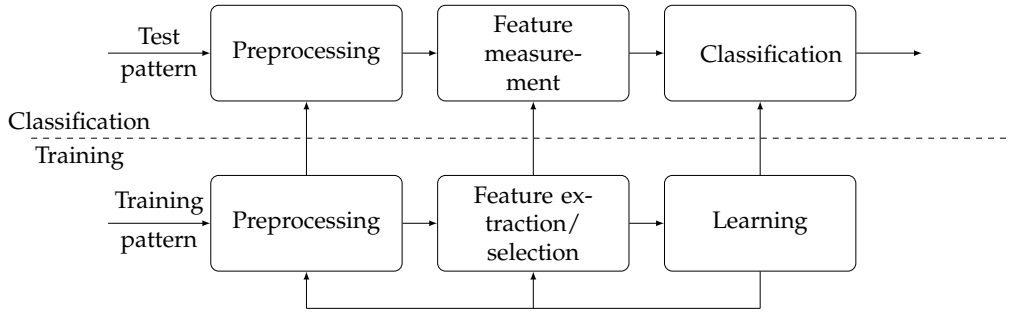


Figure 2.5: *The principle process of statistical pattern recognition: The training procedure defines the preprocessing, the feature extraction and the learning of the classifier. The testing of the resulting classifier is separated from the training procedure in order to asses its performance correctly. Adapted from Jain et al. (2000).*

2.2.1 Linear Models for Classification

A simple and well established technique to construct a classifier is the utilization of a linear discriminant function (Bishop, 2006):

$$y(x) = w^{\mathrm{T}}x + b,$$

where the parameter b is called bias and w is the weight vector. The vector w is multiplied with the data sample x using the inner product. The parameter b is called bias. Classification is conducted in this context by comparing the $y(x)$ to 0. If $y(x) \geq 0$, x is classified as C_1 and as C_2 otherwise. An illustrative example for such a linear classifier is given in Figure 2.6 as a single unit with input x, weight vector w and bias b producing the output y.

The model is only applicable, if the problem is linearly separable, which is a major limitation of this type of classifier as this is generally not the case in real world applications. The most prominent example for a problem, that is not feasible using a single linear discriminant function is obviously the well-known Xor-problem (Minsky and Papert, 1972).

The two class linear classifier can be extended to problems with C classes by either constructing $C - 1$ classifiers, which are separating the samples of each class from the data points, that are not in this class (one-vs.-rest), or alternatively a discriminative function could be created for every pair of classes and a voting is conducted to make a decision (one-vs.-one). A further variant is to use a one out of C coding of the data using a C dimensional vector with an entry unequal to zero at the respective position.

There are different approaches in the literature to find suitable parameters w and b described here. The following sections will introduce two important techniques to learn them from a given training set. This comprises pseudo inverse algorithms, that creates the least squares solution, which is a fast and effi-

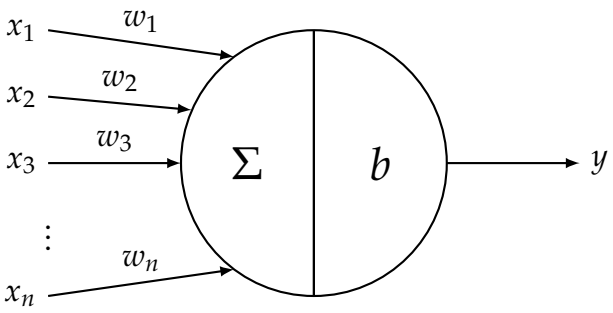


Figure 2.6: *A linear model for classification.*

cient way to compute a linear classifier, and the perceptron learning algorithm, that is a historically important approach for the machine learning community.

2.2.1.1 Least Squares Solution

For the linear least squares solution a classifier is constructed by solving the following linear equation (Bishop, 2006):

$$y(x) = \widehat{W}^{\mathrm{T}}\hat{x}.$$

This equation pools the $k = 1, \ldots, C$ individual linear equations for each class: $y_k = w_k^{\mathrm{T}} x + w_{k0}$. As usual x is the training data and the w_k comprises the weights for the dimensions, that are stored in the k-th column of W. In the matrix notation, the bias w_{k0} is set as an additional entry in this column, that finds its corresponding entry in the new input vector $\hat{x}$ with a 1 at the respective entry. The matrix X is unfortunately not invertible in most cases and hence the so called Moore-Penrose pseudo-inverse (Penrose, 1955) is applied.

$$W = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = X^{+}Y$$

For the computation of the pseudo-inverse X^{+}, the matrix X is defined with the training data samples $\hat{x}_n$ in its rows. Further, Y stores the respective labels in the one out of C coding. This results in the following classifier (Bishop, 2006):

$$y(x) = W^{\mathrm{T}}\hat{x} = T^{\mathrm{T}}\left(X^{+}\right)^{\mathrm{T}}$$

2.2.1.2 Perceptron

A further interesting way to construct a linear classifier for a two class problem is the perceptron learning algorithm (Rosenblatt, 1962). The leaning of the linear model is conducted by iteratively presenting the training data. Again,

the function $y(x) = w^{\mathrm{T}}x + b$ is used to create a classifier and the target classes T for a sample x are defined to govern in $\{-1, 1\}$.

Thus, the weights at iteration t, i.e., $w^{(t)}$, are adapted to $w^{(t+1)}$ by adding the sample vector to the weight vector, discounted by a learning rate η.

$$w^{(t+1)} = w^{(t)} + \eta(T - y)x.$$

The bias parameter b is analogously augmented as follows:

$$b^{(t+1)} = b^{(t)} - \eta(T - y).$$

A classification of a new sample is hence conducted with this model with the so called Heaviside transfer function:

$$f(y) = \begin{cases} 1, & y \geq 0 \\ 0, & y < 0. \end{cases}$$

An illustration of the learning process is shown in Figure 2.7. The encircled white data sample is presented to the perceptron and a misclassification is detected (1). The normal of the separating hyperplane is adapted according to the presented sample (2). Finally, the new linear classifier is obtained (3) and the whole process is repeated.

It has been proven that the perceptron learning algorithm converges if the underlying problem is linearly separable. If this is not the case it can be forced to converge, for example by discounting the learning rate with the number of iterations.

2.2.2 Multilayer Perceptron

The multilayer perceptron (MLP) overcomes the limitation of the single perceptron of depending on linearly separable problems to find a decision boundary by using a layered structure of multiple units (Bishop, 2006; Webb, 2002). An MLP with two layers is shown in Figure 2.8. It comprises one hidden layer, an output layer and an input layer, that is commonly not counted as an independent layer because no computation is conducted here. Each unit in a layer receives its input from every unit of the preceding layer and distributes its output to every unit in the subsequent layer.

The j-th individual unit in the first hidden layer is represented by the weights $w_{ij}^{(1)}$ for $i = 1, \ldots, D$. In this context, D equals the dimension of the input vector x. Further, w_{j0} is additive bias value. Thus, the intermediate output of the unit is given by the following term (Bishop, 2006):

$$a_j^{(1)} = \sum_{i=0}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}.$$

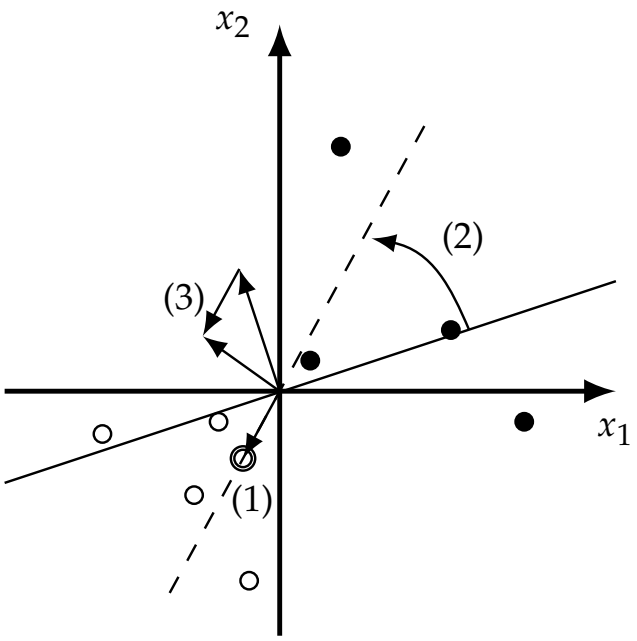


Figure 2.7: *An illustrative example for the perceptron learning rule. The black and white circles resemble the data samples, that are labeled in two classes. The encircled white sample is presented to the learning algorithm (1) and the normal vector of the decision boundary is shifted with respect this point (2). This renders the new boundary (3).*

The value of $a_j^{(1)}$ is then further processed by a squashing function σ:

$$y_j = \sigma(a_j^{(1)}).$$

Unlike the perceptron, where the Heaviside function is applied, the units in the MLP use a differentiable transfer function for example the logistic function:

$$\sigma(a) = \frac{1}{1 + \exp(-\beta a)}, \beta > 0.$$

This can be interpreted as the probability of the respective neuron to fire. In principle, other differentiable functions can be used such as the hyperbolic tangent or a linear function, as it is often used for the output layer.

Following the structure given in Figure 2.8, the output of the k-th neuron of the output layer of a neural network with one hidden layer is given by the following term (Bishop, 2006):

$$y_k(x, w) = \sigma\left(\sum_{j=0}^{M} w_{jk}^{(2)} h\left(\sum_{i=0}^{D} w_{dj}^{(1)} x_i\right)\right).$$

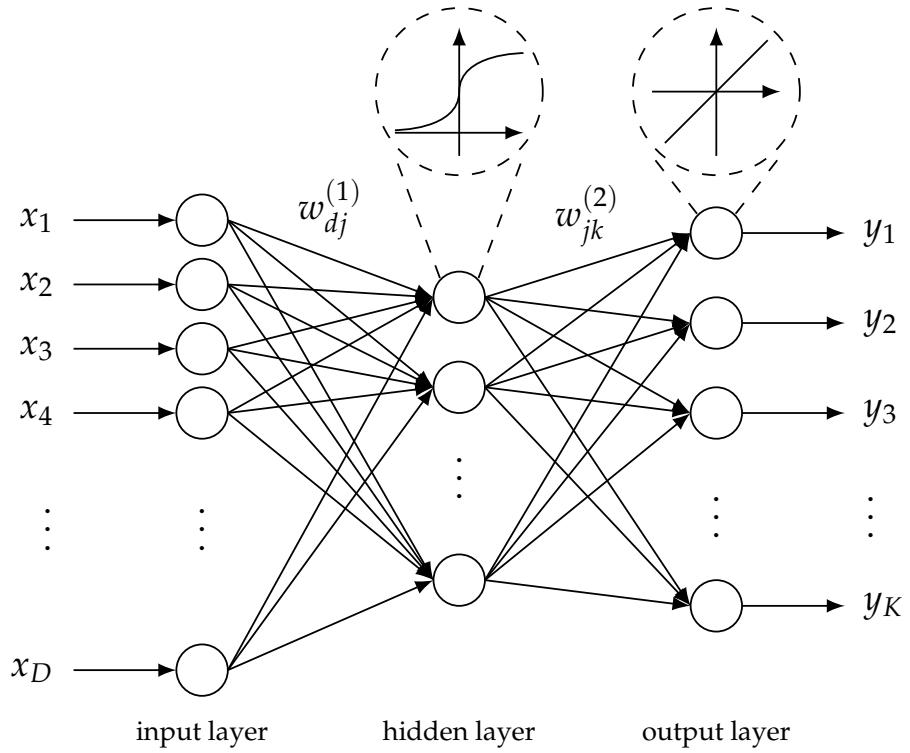


Figure 2.8: *Multilayer perceptron having one hidden layer. The input $x = (x_1, \ldots, x_D)$ to the network is provided via the input layer. The hidden layer is implemented using sigmoid neurons with weights $w_{dj}^{(1)}$. The output layer is resembled of linear units, that incorporate the weight vectors $w_{jk}^{(2)}$. The respective bias parameters are omitted in this image for simplicity. Adapted from (Webb, 2002).*

Here h and σ are the transfer functions for the hidden layer and the output layer. The dimensionality of the input data equals D, the size of the hidden layer equals M and the weight for the units in the first and second layers are denoted with a superscript index.

The training of a MLP is usually conducted by minimizing the mean square error between the actual output of the network y for the data sample x_n and the teacher signal t_n:

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, w) - t_n \right)^2 .$$

This minimization is normally carried out using a gradient descent and error back-propagation (Rumelhart et al., 1986). That means, that the weights of the network at iteration i are adjusted to $w^{(i+1)}$ by traversing into the direction of the negative derivative of the error with respect to the value of w with a learning rate η:

$$w^{(i+1)} = w^{(i)} - \eta \nabla E \left(w^{(i)} \right) .$$

As the transfer functions in all layers are chosen to be differentiable, this approach is feasible and the deduction of the learning rules can be found in most of the textbooks on neural networks and pattern recognition such as (Bishop, 2006) or (Webb, 2002).

The MLP has been proven to be a universal function approximator if enough units are provided in the hidden layers (Hornik, 1991; Cybenko, 1992). However, as the gradient descent is a greedy optimization technique, it is very likely that the gradient descent stops in a local optimum. This makes a proper initialization of the weights crucial. Furthermore, there are several important parameters that have to be configured externally, such as the learning rate η, the number of neurons in the hidden layer and the number of hidden layers. Another issue considering the gradient descent is that the learning algorithm could converge slowly when the derivatives are close to zero, for example when the error curve has the shape of a plateau or the sigmoid functions are in the saturation areas.

There are many approaches to improve over the plain gradient descent algorithm. For example a momentum term can be integrated into the learning formula in order to also use the Δw of the preceding iteration in the current repetition (Qian, 1999; Rumelhart et al., 1986). A further prominent extension is RPROP by Riedmiller (1994), implementing an approach, that aims at making the step size independent of the current magnitude of the gradient by using only the sign of the gradient together with the learning rate.

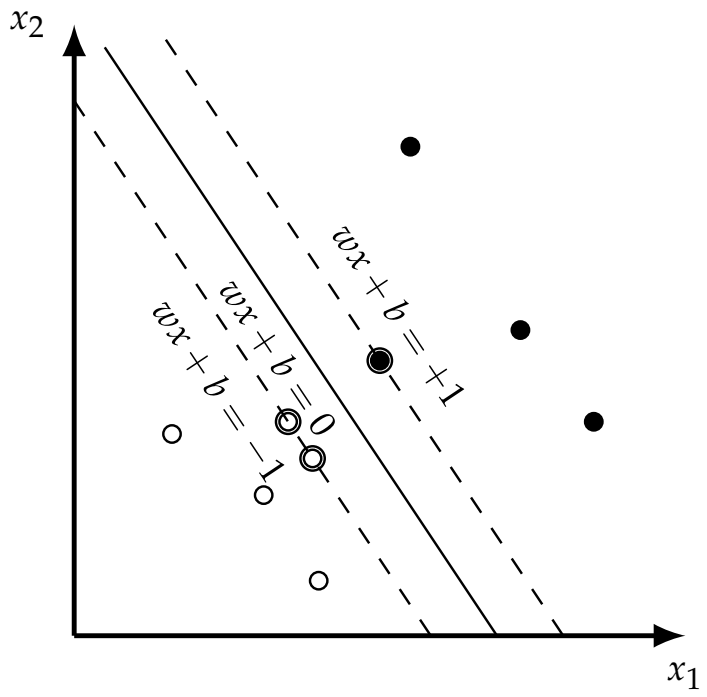


Figure 2.9: *Definition of the margin of a linear SVM. The decision boundary (solid black line) is defined by maximizing the margin (dashed lines), which is defined by the support vectors (SV) denoted as encircled samples.*

2.2.3 Support Vector Machine

The support vector machine (SVM) is one of the recently most frequently used classification technique (Bennett and Campbell, 2000; Schölkopf et al., 2000; Schölkopf and Smola, 2001). The key idea for this approach is to find a separating hyperplane in the feature space, where the margin is maximized. The margin is defined as the distance of the decision hyperplane of the SVM to the nearest data sample as it is sketched in Figure 2.9. The definition of the margin is motivated by the structural risk considerations in statistical learning theory (Vapnik, 1998, 1999), which is defined in order to assert a good generalization of the classifier. This margin is formed by the so called support vectors (SV), that are used to describe the classifier, the SV are denoted as encircled data points in Figure 2.9.

In principle, the SVM is defined as a linear classifier in the feature space, that can be augmented to nonlinear problems by using a nonlinear feature space mapping $\phi(x)$ (Cortes and Vapnik, 1995)

$$y(x) = w^{\mathrm{T}}\phi(x) + b. \tag{2.4}$$

The parameters w and b are defined as described earlier as weights and bias for

linear classifiers. Further, is x the input sample and y the classification result. The classifier is defined for two-class problems, that are given as training data $(x_1, \ldots, x_n)$ with target labels $(t_1, \ldots, t_n)$. The labels are defined to be in the set $\{-1, +1\}$. The extension to multiple classes can be conducted by one-vs.-one or one-vs.-rest approaches. In the following the SVM will be theoretically introduced first in the linear case and the extension to nonlinear problems will be discussed afterwards.

The separating hyperplane in Equation 2.4 is defined by setting $y(x) = 0$. Further, the margin is formally defined by $1/\|w\|$. Hence, maximizing the margin can be conducted by minimizing $\|w\|$ under the following constraint (Webb, 2002):

$$t_i(w^{\mathrm{T}} x_i + b) \geq 1.$$

This can be conducted using the Lagrange multipliers $a = a_1, \ldots, a_n$ with $a_i \geq 0$, resulting in the following objective function (Webb, 2002):

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n(w^{\mathrm{T}} x_n + b) - 1 \right\}.$$

This equation is commonly called the primal form. It is solved by differentiating L with respect to w and b. This returns the following constraints (Bishop, 2006):

$$w = \sum_{n=1}^{N} a_n t_n x_n$$

$$0 = \sum_{n=1}^{N} a_n t_n.$$

Substituting these results in the primal form yields the dual representation of the SVM (Webb, 2002):

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m x_n^{\mathrm{T}} x_m.$$

The dual form is maximized subject to $a_i \geq 0$ and $\sum_{i=1}^{N} a_i t_i = 0$. This is a quadratic optimization problem with N variables, that can be solved using standard techniques such as quadratic programming. The dual representation is in many cases easier to optimize (Bishop, 2006). Further, this formulation using the inner product can be used to extend the SVM to nonlinear problems as we will see later in this section. The support vectors are thus given by those samples x_i, where the respective Lagrange multiplier is not equal to 0. The classifier is then constituted by

$$y(x) = \sum_{n=1}^{N} a_n t_n x^{\mathrm{T}} x_n + b.$$

The above described SVM obviously relies on linearly separable classification problems, that may not be the case in most applications. Hence the definition of the margin is normally augmented in order to allow a certain amount of margin violations. Thus, new parameters, the so called slack variables, $\xi_1, \ldots, \xi_N$ are introduced, reflecting the degree of violation of the respective data sample. This is done by setting $\xi_i = 0$ if x_i is classified correctly, $0 < \xi_i \leq 1$ if the sample x_i is on the correct side of the hyperplane but violates the margin and $\xi_i > 1$ if the respective sample x_i is misclassified. Hence the equation for the margin is altered to

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}\|w\|^2.$$

The new parameter C is used to adjust the impact of margin violations and has to be set externally (Thiel, 2010). This new formulation also yields a modified primal form of the SVM

$$L(w, b, a) = \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} a_n \left\{ t_n(w^{\mathrm{T}} x_n + b) - 1 + \xi_n \right\} - \sum_{n=1}^{N} \mu_n \xi_n.$$

This is transformed into the dual form by also differentiating with respect to $\xi_i, i = 1, \ldots, N$. The respective deductions can be found in pattern recognition and machine learning textbooks, for example (Webb, 2002; Bishop, 2006).

As mentioned before, the SVM can be further extended to solve nonlinear classification problems. This is conducted by mapping the original data into a new feature space by a fixed transformation ϕ:

$$y(x) = w^{\mathrm{T}} \phi(x) + b.$$

This new feature space can easily have a higher dimensionality than the original data and thus it is more likely, that the data set is linearly separable (Cover, 1965). Hence, the dual representation is transformed to the following equation (Bishop, 2006):

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \phi(x_n)^{\mathrm{T}} \phi(x_m).$$

The scalar product $\phi(x_n)^{\mathrm{T}} \phi(x_m)$ in not necessarily computed explicitly, but can be replaced by a so called kernel function (Schölkopf and Smola, 2001). These kernels $k(x_n, x_m)$ implicitly compute the inner product in the feature space, which is commonly called "kernel trick". In order to construct a valid kernel function, the so called Mercer's conditions have to be fulfilled (Vapnik, 1998):

$$k(x_n, x_m) = k(x_m, x_n)$$

$$\int k(x_n, x_m) f(x_n) f(x_m) dx_i dx_j \geq 0$$

with

$$\int f^2(x) dx < \infty.$$

Prominent kernels, beside the linear kernel $x_n^{\mathrm{T}} x_m$, are the polynomial kernel $(1 + x_n^{\mathrm{T}} x_m)^p$ and the RBF kernel $\exp(-\|x_m - x_n\|^2 / \sigma^2)$ with $\sigma \neq 0$ (Webb, 2002). These kernels incorporate additional parameters, i.e., the degree of the polynomial p and the variance of the RBF function σ^2, that have to be chosen externally. The optimization of these parameters has to be adequately incorporated into the training procedure such that over-fitting is avoided.

In the standard version, the SVM is only able to process crisp class labels in both, the training procedure and the class retrieval for test samples. However, there exist a variety of different extensions to the SVM, that soften that constraint and incorporate probabilistic or fuzzy labels in the training of the classifier and the classification of an unseen sample (Thiel, 2010). The most prominent approach to obtain a probabilistic class assignment for a test sample is introduced by Platt (1999a):

$$p(t = 1|y) = \frac{1}{1 + \exp(Ay + B)}.$$

A and B are usually optimized with respect to the least squared error (Platt, 1999b) of the estimates to the true (probabilistic) labels and the distance to the hyperplane.

There are also attempts to incorporate fuzzy memberships into the deduction of the hyperplane (Lin and D., 2002; Huang and Liu, 2002). A rather intuitive method is proposed by Thiel et al. (2007), who incorporate the information about fuzzy labels as an additional factor into the slack term

$$C \sum_{n=1}^{N} (\xi_n^+ m_n^+ + \xi_n^- m_n^-) + \frac{1}{2} \|w\|^2.$$

The new variables m_n^+ and m_n^- store the fuzzy memberships for the n-th data sample. The punishment of margin violations are thus regulated proportional to the membership of the respective class.

This kind of fuzzy SVM has been very successfully applied in several difficult applications such as the voice quality classification (Scherer et al., 2013), discrimination of facial expressions (Thiel, 2010; Schels and Schwenker, 2010) and emotional spoken utterances Thiel et al. (2007).

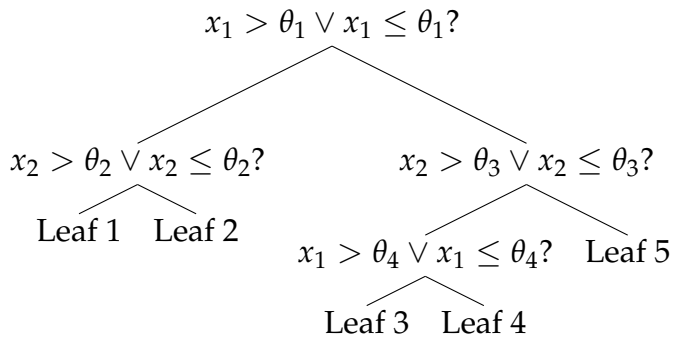


Figure 2.10: *A sample decision tree. The threshold values $\theta_1, \ldots, \theta_4$ are used to partition the data space in a tree based structure. The decision for a new sample is made in the leafs of one of the five leafs. Adapted from (Bishop, 2006).*

2.2.4 Decision Trees

Decision trees construct a binary tree classifier, where each node represents a threshold decision for a dimension of the data. These thresholds define the path to different leaf nodes of the tree, that represent the various categories. Such a decision tree is depicted in Figure 2.10, where the variables x_1 and x_2 are processed in a tree with four nodes and five leafs. In every node a threshold θ_i is evaluated for the respective variable and the left or right path is taken as denoted by the symbol $\vee$. The decision tree resembles a set of rules, which are used to implement a classifier. This makes this approach appealing for some applications since these rules are easier to interpret than, for example, the weight vectors in neural networks or SVM. On the other hand the decision tree is restricted to due to its construction to a tile-like structure as depicted in Figure 2.11. There, the representation of leafs in a decision tree based coarsely on Figure 2.10 is depicted. The class regions are rectangular areas in a two dimensional space, that mark the areas for the different categories.

The decision tree is constructed by iteratively determining the optimal feature split by defining the decrease in impurity (Webb, 2002):

$$\Delta Q = Q(t) - \left(\frac{N_{\text{left}}}{N} Q(t_{\text{left}}) + \frac{N - N_{\text{left}}}{N} Q(t_{\text{right}}) \right)$$

where Q is an impurity measure and N and N_{left} are the numbers of samples in the parental node and the left child node. In order to construct a valid impurity measure, the following constraints must hold (Breiman et al., 1984; Raileanu and Stoffel, 2004): It is maximal if the relative frequency of the classes for the samples in the partition is equal to $1/C$, with C being the number of classes. Further it has to be minimal if all samples of the respective partition are of one class. And finally, it has to be a symmetric function of its inputs.

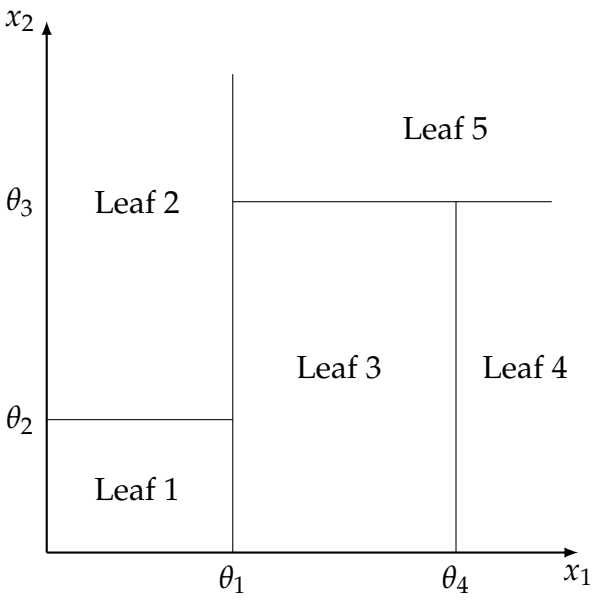


Figure 2.11: *Partition of a 2-dimensional space into five different regions for the decision tree depicted in Figure 2.10. Adapted from (Bishop, 2006).*

The simplest impurity measure, that is used for decision trees is the misclassification index:

$$Q_{\text{mis}}(p) = 1 - \max_j p_j.$$

Here $p = p_1, ..., p_C$ denotes the relative frequencies of the C classes in the data set. A further measure is the well-known Gini index, that is used in Breiman's classification and regression trees (CART) (Breiman et al., 1984):

$$Q_{\text{Gini}}(p) = 1 - \sum_{k=1}^{C} p_k^2.$$

The entropy impurity measure is applied in the well-known ID3 (Quinlan, 1986) and its derivative the C4.5 (Quinlan, 1993) algorithm:

$$Q_{\text{ent}} = -\sum_{k=1}^{C} p_k \log_2 p_k.$$

However, growing the tree until full purity in the nodes is achieved may return an over-trained classifier. Hence so called pruning techniques have been developed, that revert some steps of the training of a tree in order to obtain a proper generalization. Most of these approaches use a validation set on which the generalization is tested for the decision tree. However, we will see that pruning is not necessary when constructing a random forest (Breiman, 2001) as an ensemble approach composed of multiple decision trees, which will be discussed in Section 2.4.

2.2.5 Discussion

Supervised learning has proven to render reliable classifiers in many applications (Webb, 2002; Bishop, 2006; Theodoridis and Koutroumbas, 2009). However, the approaches in this chapter rely on a carefully engineered training set with thoroughly labeled data in well defined classes (Bache and Lichman, 2013). But in many real world applications, it is relatively easy to collect data, for example with cameras and microphones, but the annotation of well-defined categories is very expensive or even not directly possible (Walter et al., 2011; Scherer et al., 2012a), which is circumvented for example by subjective annotation (Schuller et al., 2011) or global labeling (Walter et al., 2011). This typically renders a very weak classification performance. In order to still obtain robust classifier outputs, the plain statistical model has to be augmented towards the combination of information from multiple sources, which will be reviewed in Section 2.4, partially supervised learning, which will be addressed in Section 2.5 and the temporal integration of intermediate decisions in a time sequence, that will be discussed in Section 5.2.

A further prominent challenge for statistical pattern recognition is the fact that the distribution of classes is in many cases unbalanced, i.e., there are many samples available for some classes and only few for others. This is sometimes addressed by using under or oversampling techniques (Kubat and Matwin, 1997; Chawla et al., 2002). A further approach is to incorporate weights into the loss function for the training process such that misclassification of the minority class is punished more severely than the majority class. This is further addressed in Section 5.4 of this thesis.

2.3 Learning under Uncertainty

Uncertainty is a powerful concept, which becomes increasingly important in computer science (Bishop, 2006; Thiel, 2010) and will be addressed in the following. It will be used for the combination of multiple classifier systems and the stabilization of intermediate classification results. In this chapter, the basic concepts for the classification under uncertainty are presented. In Section 2.3.1, the basic uncertainty calculi are introduced. The estimation of an uncertainty value from a classifier is discussed in Section 2.3.2. An outlook for popular applications of uncertainty measures is given in Section 2.3.2. A discussion of the matter is given in Section 2.3.4.

2.3.1 Uncertainty Calculi

In this section probability theory and the Dempster-Shafer theory of evidence are described exemplarily for the different ways to use uncertainty measures.

2.3.1.1 Probability Theory

The most frequent way to model uncertainty is to use probability theory. The triplet $(\Omega, \mathbf{A}, p)$ is called probability space for a universe Ω, which is the set of possible outcomes, $\mathbf{A}$ is a so called σ algebra and p is a probability measure (Spies, 1996). A σ algebra is defined as a set of subsets of Ω, that is not empty and is closed under complementation and countable unions. The probability measure p has to fulfill the so called Kolmogorov axioms:

- Non-negativity: $p(A_i) \geq 0 \; \forall A_i \in A$
- Unitary: $p(\Omega) = 1$
- σ-additivity: $p(\cup_{i \in I} A_i) = \sum_{i \in I} p(A_i)$ for any countable many events $A_i \in A$, that are pairwise disjoint.

The conditional probability of an event A given the event B is then defined as:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \text{ with } p(B) > 0$$

A and B are statistically independent if the probability of A given B equals the probability of A solely: $p(A|B) = p(A)$ holds. Hence, the joint probability equals the product of the individual probabilities of the events $p(A \cap B) = p(A)p(B)$.

Further, the so called sum and product rules are as follows:

$$\text{sum rule: } p(A) = \sum_B p(A \cap B)$$

$$\text{product rule: } p(A, B) = p(B|A) \cdot p(A)$$

Based on the product rule and the symmetry of p, the fundamental Bayes theorem is derived:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}.$$

In the context of the Bayes theorem, $p(A|B)$ called probability of A given B or a posterior probability and $p(A)$ and $p(B)$, which is used as a normalization factor, are called prior probabilities.

2.3.1.2 Dempster-Shafer Theory of Evidence

A further theory for the incorporation of uncertainty and information fusion, that is often considered an extension of Bayesian theory is the theory of evidence by Dempster (1968) and Shafer (1976). The basics of the theory are based on the so called frame of discernment Ω, which is the set of elementary events, that are disjoint and complete. For every element of the power set of Ω, a basic probability assignment $m : 2^{\Omega} \rightarrow [0, 1]$ is defined. The following constraints hold on m:

$$m(\emptyset) = 0 \text{ and } \sum_{A \in 2^{\Omega}} m(A) = 1,$$

where $\emptyset$ is the empty set. Based on this the belief in a hypothesis A can be expressed based on the probability mass of the supporting subsets $B \subseteq A$ with $B \neq \emptyset$. Thus the belief function bel : $2^{\Omega} \rightarrow [0, 1]$ is defined as the summation over the probability mass of B:

$$\text{bel}(B) = \sum_{A \subseteq B} m(A).$$

In order to combine two basic probability assignments m_1 and m_2, different rules of combination have been proposed. The most prominent is the orthogonal sum or Dempster's rule of combination is defined as follows:

$$m_1 \oplus m_2 = m_{12}(C) = K \sum_{A,B:A \cap B = C} m_1(A) \cdot m_2(B).$$

The variable K is a normalization factor, that is defined by

$$K^{-1} = 1 - \sum_{A,B:A \cap B = \emptyset} m_1(A) \cdot m_2(B) = \sum_{A,B:A \cap B \neq \emptyset} m_1(A) \cdot m_2(B).$$

The theory of evidence is able to distinguish between having no knowledge about a hypothesis and not supporting a hypothesis. Further, a believe assignment can not only be given to elementary hypothesizes but also to coarser sets of individual hypotheses.

2.3.2 How to Measure Uncertainty

An important issue in the context of classification is to derive a suitable fuzzy or probabilistic result from a statistical classifier. A standard approach to derive a valid probabilistic classifier answer is to normalize the outputs of a continuous classifier such as a neural network (Kuncheva, 2004). One way to that is the so called soft-max function:

$$g'_j(x) = \frac{\exp(g_j(x))}{\sum_{k=1}^{C} \exp(g_k(x))}.$$

Here, $g_j, j = 1, \ldots, C$ is the output for the j-th class for a sample x and C is the number of classes. This asserts, that all new class memberships g'_j are in $[0, 1]$ and the values sum up to one.

A variant for two-class problems is to use the logistic link function to normalize the output. This yields for class 1 the following squashed probabilistic results for a sample x:

$$g'_1(x) = \frac{1}{1 - \exp(-g_1(x))}.$$

Hence for class 2, the new probabilistic value g'_2 is given by

$$g'_2(x) = 1 - g'_1(x).$$

The outputs of the original classifier can also be treated as a further optimization problem for a statistical learner. For example the above described method can be extended using a fermi function, that can be optimized for a given problem:

$$g'_1(x) = \frac{1}{1 - \alpha \exp(-\beta g_1(x))}.$$

Platt (1999b) proposed an effective method to optimize α and β, that is frequently used for SVM. Here the distance to the separating hyperplane is used to rate the confidence of a classification.

A further possibility, that is often used in the literature is to create multiple diverse classifiers, which is often called classifier ensemble (compare Section 2.4 for an elaborate discussion) and use the agreement of the ensemble as confidence measure. Krogh and Vedelsby (1995) use the variance of individual classifiers around a weighted mean value as confidence, which they call ambiguity a:

$$a(x) = \sum_{\alpha} w_\alpha \left(V^\alpha(x) - \bar{V}(x)\right)^2.$$

Here, V^α is the α-th individual classifier and the mean classifier answer for sample x is denoted as $\bar{V}(x)$. Hansen et al. (1997) use a voting approach to compute a classification uncertainty, which they call consensus value. Concretely, they normalize the number of votes for the winner class with the number of all classifiers.

There are also attempts to model the confidence of a classification on a lower level from the quality of the data. For example, Poh et al. (2007) use global parameters in a human face verification system by explicitly modeling states, where the classification performance will be good or bad. Such a parameter could be "good illumination" versus "bad illumination". Further, Thiel (2010) describes an approach to weight the decision of a system for the recognition of facial expressions by the over-all movement in the individual images of a video.

2.3.3 Applications

Incorporating uncertainty values in a classification process has been studied and evaluated in many different real word scenarios. Especially for challenging problems and larger classification architectures, it is beneficial to postpone making a crisp decision as long as possible.

Another common application for probabilistic class memberships is the field of classifier fusion (Kuncheva, 2004; Schels et al., 2009). For the combination of information of multiple sources, it is crucial to have an accurate probability estimate. This enables to correct decisions of individual classifiers, that are possibly wrong (compare Section 2.4 for an introduction in MCS).

A widely used technique using uncertainty measures is the classification using reject options (Chow, 1970; De Stefano et al., 2000). This means, that if the classifier shows uncertainty values over a distinct threshold, no decision is drawn. This implies, that making no classification at all is a better option than triggering a false alarm. Such approaches have been successfully evaluated by Glodek et al. (2012a,b) on the AVEC 2011 data collection for the recognition of affective states from audio-visual data. Further a smoothing technique has been applied in order to still return results for all time steps.

A variant of the multiple classifier fusion is the integration of intermediate classification results of the same model over a time period. This often enables one to still render a good classification result, even if the direct results are very weak (Schels et al., 2012a; Glodek et al., 2012a). This follows the assumption, that the respective phenomena, i.e., class labels, in a non-stationary process do not change quickly over time. A systematical evaluation of the temporal integration in the context of audio-visual emotional data collections is conducted in Section 5.2.

Another application is the integration of unlabeled data into a supervised classification process in the framework of partially or semi-supervised learning (Zhu, 2005). Measuring confidences is also crucial in this process: Often, the unlabeled data is classified by some pre-trained model and according to the confidence, the data is added to a training set of the same or a different classifier (Blum and Mitchell, 1998; Hady, 2011). In active learning, the most uncertain classification is selected to conduct a query to an expert.

However, at a certain step in an information processing system, a crisp decision is normally mandatory. This can be the case, when an action has to be conducted in a bigger architecture or for computing a profane classification accuracy.

2.3.4 Discussion

In this section, different techniques for the incorporation of uncertainty into a statistical classification process are discussed. Besides the two probabilistic calculi, that are presented here, namely probability theory and the Dempster-Shafer theory of evidence, additional approaches are possibility theory (Zadeh, 1999) or fuzzy logic (Hajek, 2010), amongst others.

An important issue is to estimate a valid uncertainty value for a classification result. In many cases, a normalization approach is conducted to compute a value, that can be interpreted as a probability from a continuous output. It is also feasible to optimize a function with respect to the given training data as it is often utilized with the probabilistic SVM (Platt, 1999b). But there are also techniques to independently estimate the reliability of the classification of a data point, for example by judging the quality of the underlying feature.

Even though there are intuitive techniques to estimate the uncertainty of a decision it is not completely sure if a meaningful value is returned. That means that a classifier can easily make a confident decision, that is nevertheless wrong. However it is crucial for many applications like multiple classifier fusion and semi-supervised learning to have reliable predictions of the probability of class memberships. One of the main future challenges in computer science is to incorporate uncertainty and probabilistic principles not only into pattern recognition approaches but also into artificial intelligence architectures (Bishop, 2006).

2.4 Learning From Multiple Sources

Information from multiple sources is naturally provided in many pattern recognition applications. This can be originated from using different sensors or feature extraction approaches. In this section the combination of different channels or classifier outputs is discussed in the context of multiple classifier systems.

2.4.1 Fusion Methods

In the following, different popular techniques for the combination of individual sources of information are discussed. One possible very coarse taxonomy of fusion approaches is the principle of late, where the combination is conducted after the classification of the individual channels, and early fusion, where the classification is conducted after the combination on a feature or data level. A further interesting point in the context of feature and decision combi-

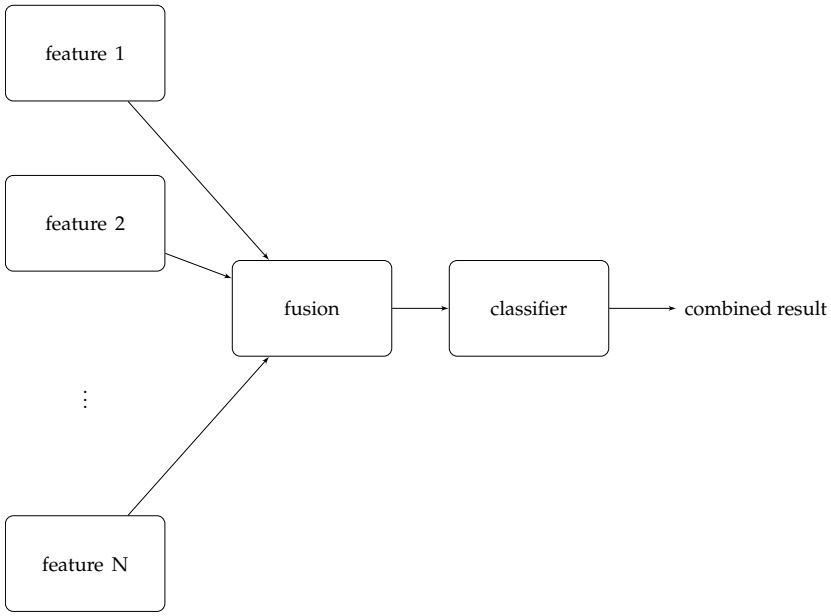


Figure 2.12: *Early fusion: The combination of the inputs is conducted before the classification.*

nation arises from the temporal nature of many real world applications. This can improve the classification for problems, where the alternation of the classes is relatively slow.

2.4.1.1 Early Fusion

In early fusion, the combination of the data or the features, that are provided by the individual information sources, for example different modalities or different feature views on the data, is conducted beforehand the learning of the classifier. This setting is depicted in Figure 2.12. The most straightforward approach to implement an early fusion is to concatenate the individual feature vectors. This setting is studied for example by Wimmer et al. (2008) and Wagner et al. (2011) in the context of human-computer interaction.

A further prominent approach to implement an early fusion of feature vectors is the GMM supervector (Campbell et al., 2006; Schels and Schwenker, 2010). Here, a so called universal background model, which is basically a GMM, is trained class-indifferently using all available data. This model is then adapted for every sequence using maximum a posteriori adaptation using the features of the respective time series. The concatenation of the centers of the adapted background model form then the combined representation. Bocklet et al. (2010) use a combination of both techniques, where supervectors are computed for multiple features and these results are then concatenated.

Early fusion can help to improve the classification performance. However,

there are several drawbacks, that have to be considered for early fusion. One issue is that the resulting concatenated feature vector may have an extensive length, which may not be reflected in the size of the training set and as a result may also lead to over-fitting, runtime or memory issues. Further, especially when different feature extraction approaches are used, the concatenation may lead to different scalings of the entries in the over-all vector. Consider for example, when in image processing color and orientation histograms are combined, then the same value of different entries may have divergent meanings.

Simple concatenation may not work when sequences of different sample rates are considered. This makes pooling approaches, i.e., combining multiple features of a kind with one other, necessary in order to overcome these differences. Wagner et al. (2011) argue, that early fusion does not allow the incorporation of missing data. If a feature vector is absent, the whole classification is nullified.

2.4.1.2 Late Fusion

The other possibility to conduct the fusion of multiple sources is to carry out the classification on every feature type separately and combining the individual decisions afterwards. This setting is displayed in Figure 2.13. The combination of classifiers is a vivid field of research in the pattern recognition community (Kuncheva, 2004; Sansone et al., 2011). Different names for this kind of architecture are “ensemble learning”, “multiple classifier systems” (MCS) or “classifier combination”.

This approach is generally more flexible than early fusion, mitigating many of the drawbacks that are mentioned there. The combination is performed on a very compact representation, i.e., the respective labels. The classification can in principle still be conducted in parts, when a particular feature fails. On the other hand, the need for separate classifiers imposes additional computational efforts.

Dietterich (2000) sketches in his influential paper three main reasons, why multiple classifier systems can improve classification over the most accurate individual classifier.

- One argument for MCS are of statistical nature: In many applications, there are too few training data to compute the true hypothesis. Hence it is likely, that multiple hypotheses that are of equal accuracy on the given data are estimated. Hence it may be safer to combine different models to average out the classification error.
- Further, computational reasons in the training of the individual classifiers are as follows: Many machine learning techniques use greedy optimization approaches, that suffer from local minimums. Prominent exam-

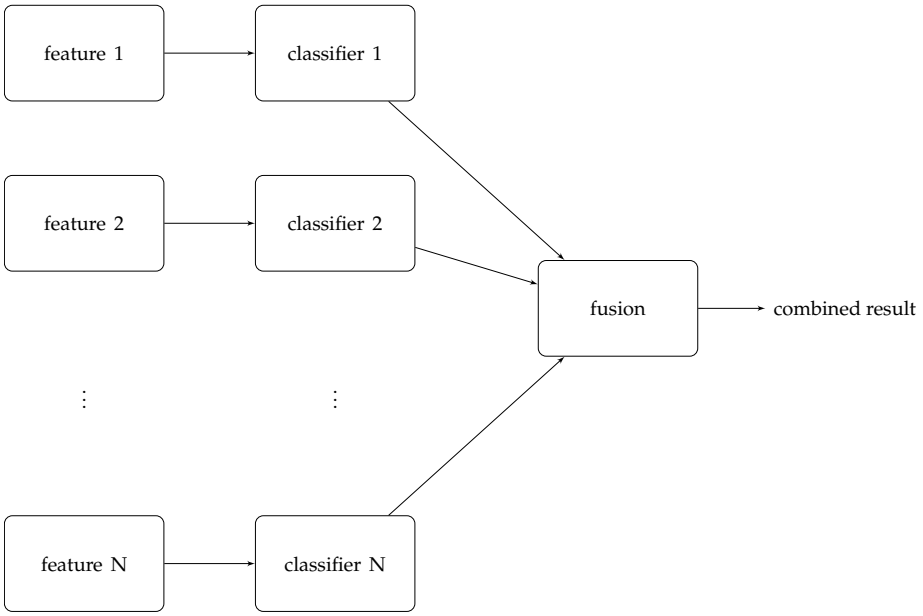


Figure 2.13: *Late fusion: Each feature is classified separately and the individual results are then combined adequately.*

ples for such learning algorithms are the gradient decent approaches for neural networks and the greedy splitting techniques for decision trees. Again, the combination of the locally optimal hypotheses can help to approach towards the global optimum.

- Lastly, there are representational arguments for MCS: The true hypothesis for a given application might not be in the respective hypothesis space. For example a nonlinear decision boundary cannot be captured by a linear classifier. However the combination of multiple linear models may accomplish the task. In principle, many theoretical models are able to approximate every arbitrary function, such as neural networks with enough hidden units. However, as training data is only available in finite quantities, the models are not able to reach the global optimum. An example for this setting is depicted in Figure 2.14, where two linear classifiers can reflect the decision border for the data better than each individually.

In (Dietterich, 2000), decent graphical illustrations of these reasons are delivered, which are often reproduced, for example in (Kuncheva, 2004). Kuncheva (2004) further states that there is of course no guarantee, that the combination of classifiers can outperform the best individual classifier.

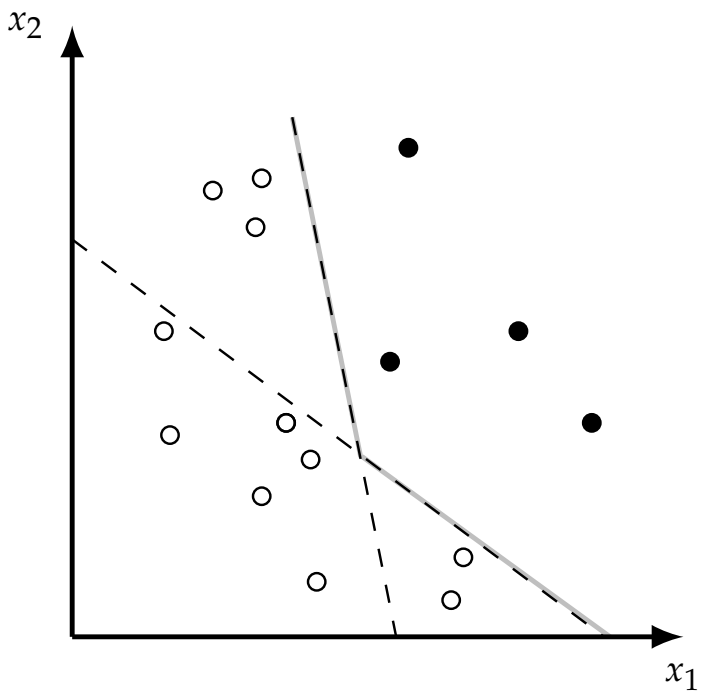


Figure 2.14: *Using a combination of decision borders with few degrees (dashed lines) of freedom can render classification of data with complexer structures (as denoted by the gray solid line), for example by classifying a sample as "black" if both models classify it "black" and "white" otherwise.*

2.4.2 Techniques of Classifier Combination

In the following, different approaches of classifier combination are discussed. The major difference for these approaches is whether a fixed combination rule is used for the fusion (Kittler et al., 1998) or the combiner is adapted by an additional training (Schwenker et al., 2006).

2.4.2.1 Combination with a Fixed Rule

The most straightforward and rather popular rule for the combination of discrete classes is the majority voting. Kuncheva (2004) discusses different variants of this rule, but the most popular is defined as follows:

$$\text{combined decision} = \arg\max_c \sum_{i=1}^{L} d_{ic}.$$

Here, L is the number of individual classifiers and C is the number of classes. Further, d_{ij} equals 1 if the i-th classifier is voting for class j and 0 otherwise.

Kuncheva (2004) provides an expectation value for the accuracy of majority voting for odd numbers of individual classifiers L.

$$\text{Acc} = \sum_{m=\lfloor L/2 \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m}.$$

Here, p is the probability of a correct decision for every classifier on all data points x. The combined output will be correct, if more than $\lfloor L/2 \rfloor$ classifiers are correct. This holds true for two class problems. For more classes, the combined accuracy may be even higher.

Other approaches rely on probabilistic class memberships $d_c(x) = \mathrm{P}(\omega_c|x)$, which is the probability of class ω_c given the data point x. They are estimated from continuous outputs of the individual classifiers as uncertainty values. Hence, let the output of the j-th classifier for a data point x be defined as $D_c(x) = (d_{1c}(x), \ldots, d_{Lc}(x))$. Based on this, the most prominent fixed rule combination rules are defined as follows (Dietterich, 2000):

- Product rule:

  $$\mu_c(x) = \prod_{i=1}^{L} d_{ic}(x).$$

  For this rule, the individual classifiers are regarded as statistically independent, i.e., conditionally independent given the class label. This can be achieved, for example by using features, that are as independently as

possible. An intuitive example is to use fingerprint and facial images for identity verification. An important issue of the approach is that if one classifier returns a fuzzy value of zero, the final decision may be biased disproportional to a distinct result.

- Average rule:
$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^{L} d_{ij}(x).$$
When using this rule, the decisions of the individual classifiers are also supposed to be independent. Additionally, a random noise on the computation of the confidence values of the classification is accounted to by the average. Such a setting is present, when the individual classifiers are trained on different subsets of a training set or when diverse features are sampled for the construction of the classifiers.

- Maximum rule:
$$\mu_j(x) = \max d_{ij}(x).$$
This rule chooses the decision of the most confident classifier for a class as the combined result. It is generally intuitive, but it depends on reliable confidence estimates for the combination. For example, classifiers that are able to reflect more subtle confidence distributions may be overruled by others that are for example only capable of returning zero or one as a result. A prominent example is the combination of multiple binary classifiers to adapt to a multi-class problem, for example one-against-rest combination of SVM (Tax and Duin, 2002). The maximum rule is equal to the logical "or" in the fuzzy logic.

- Minimum rule:
$$\mu_j(x) = \min d_{ij}(x).$$
Analogous to the maximum rule, the minimum rule is equal to the "and" in fuzzy logic. This rule is hardly ever used in applications, but commonly described in the literature.

The product and average rule are investigated theoretically by Kittler et al. (1998) from a Bayesian perspective. Further a formal description of the boundaries of the classification errors for these combiners are given by Kuncheva (2002b).

2.4.2.2 Combination using Trainable Mappings

Beside the fixed rule combiners, the classifier combination can be conducted using trainable fusion mappings (Duin, 2002; Kamel and Wanas, 2003). Kuncheva (2004) further divides these approaches into class conscious combiners,

that computes the combined result only within the different classes and class indifferent combiners. The simplest class conscious combiner is the weighted average:

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^{L} w_i d_{ij}(x).$$

Here, it is commonly assumed that the weights are $w_i \geq 0$ and that they add up to one: $\sum_{i=1}^{L} w_i = 1$.

Class indifferent combiners use also computations between different labels in order to compute the over-all result. A very prominent variant of this class of combiners are the decision templates (Kuncheva, 2004), which are defined via the so called decision profile of the classifiers:

$$DP(x) = \begin{pmatrix} d_{11}(x) & \dots & d_{1j}(x) & \dots & d_{1c}(x) \\ \vdots & & \vdots & & \vdots \\ d_{i1}(x) & \dots & d_{ij}(x) & \dots & d_{ic}(x) \\ \vdots & & \vdots & & \vdots \\ d_{L1}(x) & \dots & d_{Lj}(x) & \dots & d_{Lc}(x) \end{pmatrix}$$

Here, the $d_{ij}(x)$ are the output for the $j = 1, \dots, c$ classes of the $i = 1, \dots, L$ classifiers for a data sample x. The decision template for class j is then defined as follows:

$$DT_j = \frac{1}{N_j} \sum_{\substack{z_k \in \omega_j \\ z_k \in Z}} DP(z_k).$$

In this equation, ω_j is defined as the set of the samples of class j that are members of the training set Z and N_j is its magnitude.

For classification of a new sample x, a distance measure, for example the Euclidean distance, between the sample's decision profile $DP(x)$ and the decision templates for the classes is computed. The over-all classification result is defined by the smallest distance decision template (Kuncheva et al., 2001). Thus, a linear mapping for the classifier fusion function is defined.

Further linear mappings, for example the least squares pseudo inverse solution, associative memory and naïve Bayes are described by Schwenker et al. (2006). They showed that these mappings and also the decision templates incorporate the confusion matrices of the individual classifiers.

In principle, any classification approach can be used for the fusion layer. This reflects the viewpoint of Kuncheva (2004), that the classifier fusion is only a new pattern recognition in an intermediate feature space, i.e., the outputs of the individual classifiers.

An important meta-technique in the literature for the construction of classifier ensembles is stacking (Wolpert, 1992). The individual classifiers are trained on different subsets of the training data. The combining layer is then trained on the data that is held out of the training process for the classifiers.

Duin (2002) gives recommendations for the usage of trainable classifiers: If the individual classifiers are not over-trained and valid confidences are available, then a fixed combining rule should be preferred. Hence the full training set can be used for the individual classifiers. If the base classifiers are trained weakly, i.e., so that the classifier output is only weakly correlated with the true label, on the whole available training set it is suggested to construct a further trained mapping based on the same data. If the training data is split into two pieces, the individual classifiers can be trained as good as possible. One part of the data is hence used for the training of the classifiers and the other one for the construction of the combiner.

2.4.3 Construction of Meaningful Ensembles

The concept of diversity is crucial for the construction of working classifier ensembles. In Section 2.4.3.1, the term diversity is defined and measure for it are introduced. In Section 2.4.3.2, techniques for the creation of diverse ensembles are explained.

2.4.3.1 Classifier Diversity

In order to improve an ensemble of classifiers over the best individual classifier, the members of the classifier team do not only have to show high individual accuracies, but the individual classifiers also have to be diverse according to (Ruta and Gabrys, 2005). This means that it is beneficial for the individual classifiers to not all agree on a distinct result, especially when the classification is wrong.

In the literature, a manifold of possible diversity measures (Kuncheva and Whitaker, 2003) are described, of which the most important are the following three:

- A measure of diversity can be defined based on the entropy

$$E = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{(L - \lceil L/2 \rceil)} \min \left(l(z_j), L - l(z_j) \right),$$

$l(z_j) = \sum_{i=1}^{L} y_{ji}$, the number of correct classifiers for a data sample z_{ji}.

- The Kohavi-Wolpert variance (Kohavi and Wolpert, 1996) is accordingly defined as

$$KW = \frac{1}{NL^2} \sum_{j=1}^{N} l(z_j)(L - l(z_j)).$$

- The κ statistics of inter-rater agreement (Fleiss, 1981) is a popular measure in psychology and related fields of research. It is defined as follows:

$$\kappa = 1 - \frac{\frac{1}{L} \sum_{j=1}^{N} l(z_j) \left(L - l(z_j)\right)}{N(L-1)\bar{p}(1-\bar{p})}.$$

  Here, $\bar{p} = \frac{1}{NL} \sum_{j=1}^{N} \sum_{i=1}^{L} y_{ji}$ is the average accuracy of the individual classifiers. The variable $l(z_j)$ is defined above.

The relationship between the diversity measures and the ensemble accuracy is not so straight forward (Kuncheva, 2003b). Kuncheva and Whitaker (2003) conducted experimental evaluations for different measures with respect to the ensemble accuracy. They use scatter plots of the accuracy of the majority voting fusion against various diversity measures. One finding of these evaluations is that when using a pairwise diversity measure, the values for the classifier team should show similar pairwise dependencies in order to render a feasible ensemble. Furthermore there is a threshold for the measures, where the authors found an improvement of the classification for all cases. However, finding the classifier team with the highest diversity does not necessarily result in an optimal accuracy.

Nevertheless, there exist a variety of approaches, that construct ensembles using diversity measures. For example Giacinto and Roli (2001) use a diversity measure as distance matrix for a hierarchical clustering approach. The clusters with the least pairwise diversity are joined iteratively until a predefined number of L individual classifiers are reached. The center of a cluster consists for example of the most accurate individual classifier in the partition.

Brown and Kuncheva (2010) conducted an error decomposition of the ensemble error for majority voting into the individual error component and terms for a "good" and a "bad" diversity. These terms improve or degrade the combined classification. The decomposition is stated as follows:

$$E_{\text{maj}} = E_{\text{ind}} - \frac{1}{T} \sum_{v^{\mathrm{T}}\mathbf{1} \leq \frac{T+1}{2}} (T - v^{\mathrm{T}}\mathbf{1}) p(v) + \frac{1}{T} \sum_{v^{\mathrm{T}}\mathbf{1} < \frac{T+1}{2}} v^{\mathrm{T}}\mathbf{1} p(v).$$

In this equation, E_{maj} and E_{ind} are the combined and the individual errors, T is the number of individual classifiers, which is assumed to be odd. The t-th entry of the binary vector v is set to one, when the classification of the data sample

x by the t-th classifier is correct and zero otherwise. Further, $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}}$ of length S. Hence the ensemble is correct if $\frac{S+1}{2}$ individual classifiers decide correctly and the good diversity is hence modeled with $S - v^{\mathrm{T}}\mathbf{1}$, which is the number of incorrect classifications. The bad diversity is the amount of correct classifications when the ensemble is wrong. The summation over all possible cases of v corresponds to an integration over the data space.

2.4.3.2 Approaches to Create Diverse Classifier Teams

In the literature, there are four main techniques to create diverse classifier ensembles, that are summarized by Kuncheva (2003a).

- The most popular approach is to manipulate the training set by restricting the data, that is used to construct the respective individual classifier. The easiest way is to split the data randomly into subsets. A more elaborate technique is boosting (Freund and Schapire, 1997), where an iterative classifier training of the individual classifiers is conducted, where more and more weight is given to the samples that are in a sense harder to classify. A further prominent technique is bagging (Breiman, 1996), where a uniform sampling with replacement is conducted to choose the samples for the respective individual classifier. A variant of this is adaptive resampling (Bauer and Kohavi, 1999), where a distribution is used for the sampling of the data, that is adapted to focus on the samples, that are harder to classify such that these data points will occur more often in the training sets.

- The manipulation of the feature set is also used to construct ensembles. An examples for that is the random subspace method (Ho, 1998), where random sampling of features is conducted to construct independent decision tree classifiers. Breiman (2001) combines random feature selection with bagging for the well established random forest algorithm. In some cases there are naturally different feature views on the data, for example by using different extraction techniques. Using these features for individual classifiers can form a successful ensemble (compare e.g., Schels et al., 2009).

- As a third technique to create a diverse classifier team, different base classifier approaches can be used. For example in (Woods et al., 1997), four different (locally) optimal classifiers, i.e., an artificial neural network, a decision tree and a Bayes classifier amongst others, are used to form a successful ensemble.

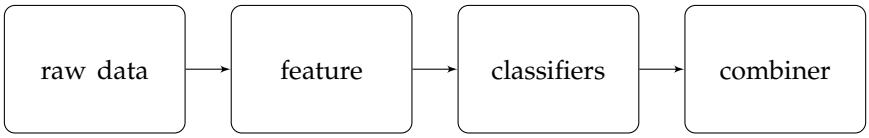


Figure 2.15: *The construction of a successful classifier ensemble can be conducted on the features, the training set, the individual classifiers and the combiner.*

- Kuncheva (2003a) states a fourth approach, where the individual classifiers are given in advance and the challenge is to pick a proper combination scheme, that ought to be adapted properly.

2.4.4 Classifier Selection

Classifier selection is the process to choose the most adequate individual classifiers rather than a combination of them (Kuncheva, 2004). This selection can also be based on a so called regional competence. This competence is for example computed using a validation set by evaluating the accuracy for the k nearest neighbors of a sample (Woods et al., 1997). Note that the selection of the best classifier can be easier affected by over-training of the individual classifiers.

An example for such classifier selection techniques is the well known mixture of experts algorithm (Jordan and Jacobs, 1994). Here, a gating network that reflects the local competence of the individual classifiers, is constructed during the training of the over-all classifier system. Further combinations of classifier selection and classifier combination are for example described in (Kuncheva, 2002a).

2.4.5 Discussion

Combining multiple classifiers is a powerful concept for pattern recognition. This chapter introduced this technique in several aspects. Fixed rule and trainable combiners were introduced and discussed to some extent. Further, different arguments were given why a combination of classifiers can improve over the best individual classifier.

A very important concept in this context is the classifier diversity. Several well-known measures of diversity are discussed and a theoretical analysis for a simpler classifier fusion scheme is given.

2.5 Partially Supervised Learning

For statistical training of a classifier, it is crucial to have carefully labeled example samples at hand. Unfortunately, this annotation is often very time consuming and expensive as human experts have to make a thorough label assignment. Hence it is an interesting research question how unlabeled data can be appropriately used in this context. The process of incorporating unlabeled data into a supervised classification process is here called partially or semi-supervised learning.

Active learning is a technique, that still includes a human expert into the training of a classifier. This expert is queried by the algorithm only for the most informative samples, that are not yet labeled, in order to keep the annotation overhead low. Other approaches aim at replacing the expert by labeling the data for the classifier itself (self-training) or teaching other classifiers by labeling data in turns (co-training).

Further semi-supervised approaches rely on the cluster assumption: The respective classes are clustered together in areas of high sample density. The decision border is then assumed to be located in regions of the feature space, where only few samples occur. Examples for that are the semi-supervised learning with generative models, the EM algorithm and the transductive SVM. This SVM tries to find the optimal large margin classifier by determining labels for previously unlabeled data, which is often resembled by the test data. For a comprehensive overview, the reader is referred to (Zhu, 2005).

2.5.1 Active Learning

Active learning is a popular technique for the learning of a classifier using unlabeled data (Settles, 2009), where the most informative sample from the unlabeled data is selected by the algorithm and passed to a human expert (Cohn et al., 1996) and the classifier is adapted accordingly. For the algorithm to work properly and to minimize the labeling efforts, it is crucial in this area to conduct a reliable sampling of the unlabeled data.

One approach in the literature is uncertainty sampling, where the sample, that inherits the least confident classification result is selected. One rule to select a sample for the new training set x^* is as follows:

$$x^* = \arg\max_x \left(1 - P_\theta(y_{\max}|x)\right)$$

for all data points x and a probabilistic model θ and $y_{\max} = \arg\max_y P_\theta(y|x)$. This is commonly related to the distance of the samples to the decision problem. This setting is displayed in Figure 2.16 for a two class problem.

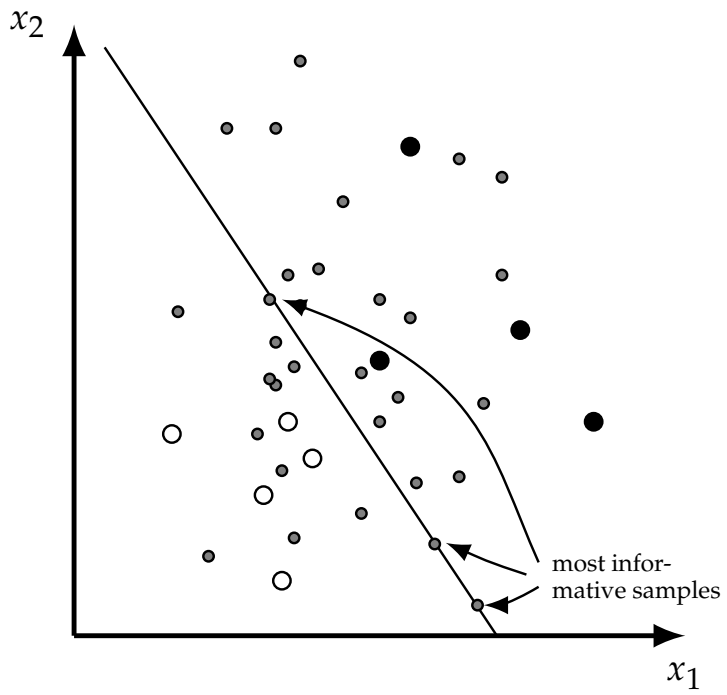


Figure 2.16: *Selecting the most informative unlabeled samples, that are located closest to the decision border for active learning.*

In order not to lose the information for the classification result of other classes the above definition is extended to the so called margin sampling:

$$x^* = \arg\min_x P_\theta(y_1|x) - P_\theta(y_2|x).$$

It measures the confidence using the difference of the most confident class y_1 and the second most confident class y_2.

A further possible approach to measure the confidence for active learning in the literature is based on ensemble learning. As mentioned before, the agreement of the individual classifiers is used for confidence measure. This is in the semi-supervised learning literature often called query by committee (QBC). Further, well known ensemble techniques such as boosting and bagging are adapted to implement an efficient strategy for the selection of the data (Abe and Mamitsuka, 1998).

2.5.2 Generative Models

One of the early approaches to partially supervised learning is to utilize additional unlabeled data in the well established EM algorithm (Baluja, 1999;

Nigam et al., 2000) and thus iteratively estimate generative models more accurately than using only the labeled data (Cooper and Freeman, 1970; Ganesalingam, 1989). The resulting model is hence used as a classifier. In Figure 2.17, an illustrative example for this setting with two Gaussian components is given: The model can be estimated more accurately using all available data, labeled and unlabeled. Most commonly it is conducted by the following steps described in Algorithm 2.1.

Algorithm 2.1: Semi supervised learning using generative models (Nigam et al., 2006).

Input:

- a set L of labeled training examples
- a set U of unlabeled examples

Output: Learnt classifier model θ
Estimate an initial model using the labeled data, solely L;
while *log likelihood improves* **do**

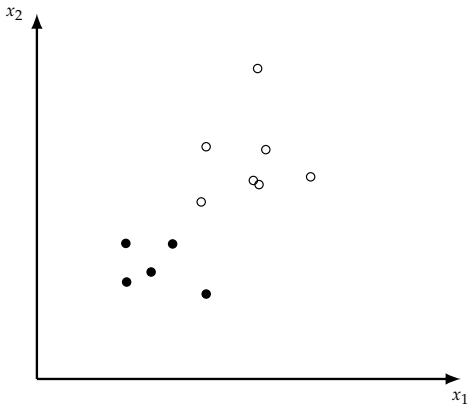
- Label the additional unlabeled samples U according to $P(c_j|x_i, \theta')$;
- Re-estimate the model parameters given the new component memberships using the maximum a-posteriori estimation $\theta' = \arg\max_\theta P(X, Y|\theta)P(\theta)$;

At first, an initial model is trained using only the labeled samples. Using this model, the unlabeled data is classified and the generative models are re-estimated using this additional information. This procedure is repeated until some stopping criterion is met, for example until the likelihood of the models does not increase any more or for a predefined number of iterations.

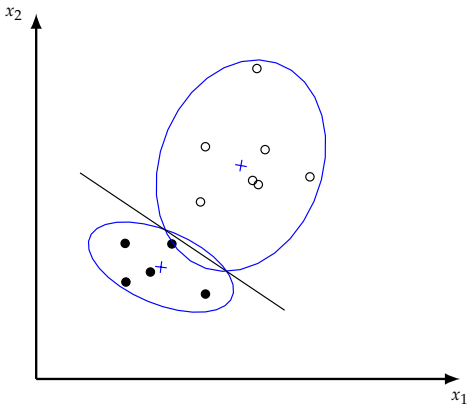
A major issue for the EM algorithm in general is that it is likely to get stuck in a local maximum. A further assumption that is made in the semi-supervised learning with generative models is that there has to be a correspondence of classes and components. In principle, any generative model such as Gaussian mixture models, naïve Bayes and hidden Markov models can be used in this context.

2.5.3 Self-Training

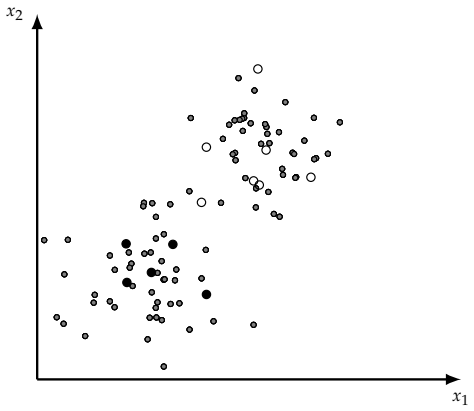
One of the most straight forward semi-supervised learning approaches is the so called self-training (Rosenberg et al., 2005). For this learning technique, a



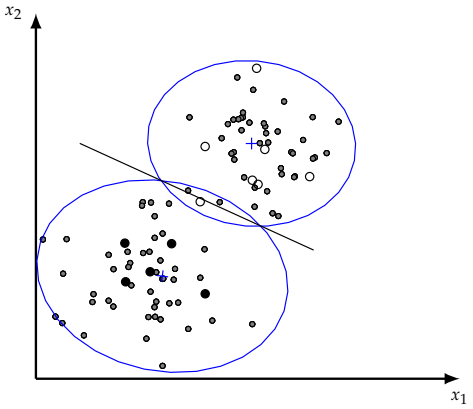
(a) Only few labeled data are available in two classes denoted as white and black points.



(b) The resulting generative model sketched by contour lines. The respective class border is also shown.



(c) Additional unlabeled data are given given as smaller, grayish filled circles.



(d) Exploiting unlabeled data renders a different, more intuitive model and respective class border.

Figure 2.17: *Estimating a generative model with two components with labeled and unlabeled samples.*

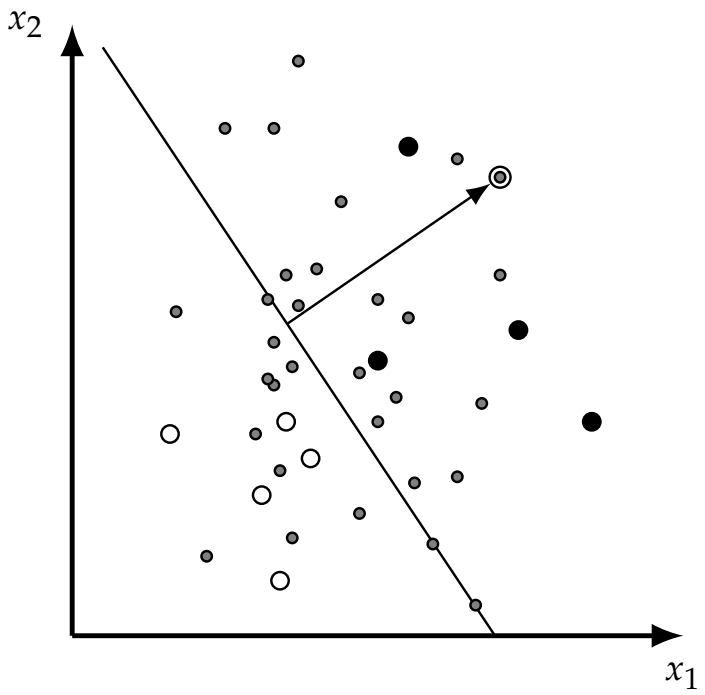


Figure 2.18: *Sample selection for self-training. The distance from the decision border is used as confidence measure.*

classifier is used to iteratively label the unlabeled data set U. In every iteration the most confident samples with respect to a qualified confidence measure are added to the labeled training set L of the classifier. Afterwards, the classifier is retrained using the new training set.

Figure 2.18 shows an example of the sample selection in a self-training setting using a linear classifier with the distance to the decision border as confidence measure. It also shows a major drawback of the approach as the selected sample is too far away from the interesting areas of the input space to be informative for the classifier.

An important aspect for the successful application of self-training is to have a correct measure of confidence in order to prevent adding noise (i.e., classification errors) to the training data. Especially, exploration towards the decision border increases the risk of adding misclassified samples to the training data. Also chaining effects, that are caused by outliers in the data, can alter the method. Thus it is possible, that an unintuitive, and probably erroneous, decision function is learned using this approach as it is illustrated in (Hady, 2011, p. 80).

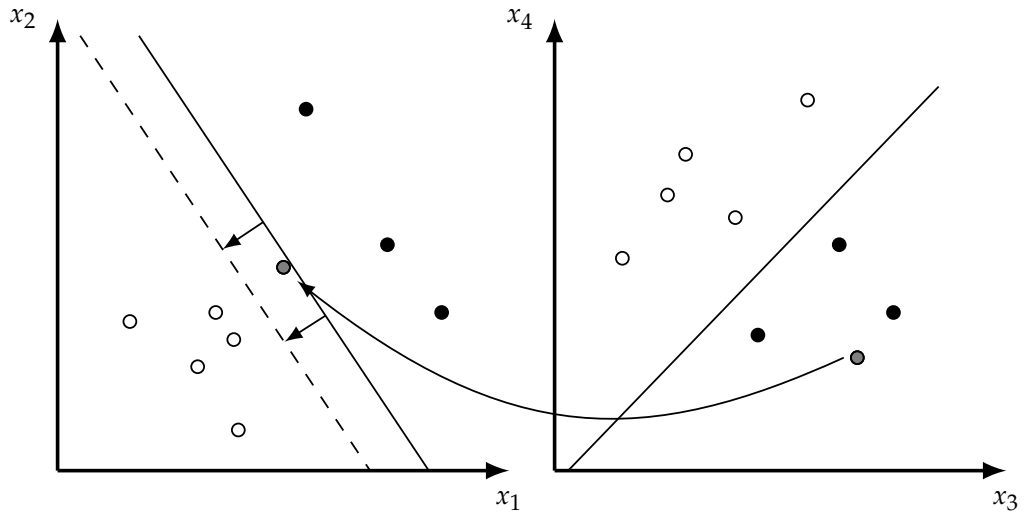


Figure 2.19: *An illustrative argument why co-training works: Two different classifiers are constructed for subsets of the feature space. The classifier on the right hand side of the figure selects the unlabeled data point that is confidently labeled as "black", by the distance to its decision border. This sample is very informative for the classifier on the left hand side and results in an adaptation of the classifier (dashed line).*

2.5.4 Co-Training

A more powerful alternative to self-training is the so called co-training (Blum and Mitchell, 1998; Ling et al., 2009). In co-training, additional data is added to the training set of a classifier by another one and vice versa. The key idea of the approach is that for the different classifiers, an unlabeled data point, that is confident for one classifier can be informative for another one. This circumstance is addressed in Figure 2.19 for two classifiers on different feature sets on a two class problem. The basic algorithm to conduct co-training is outlined in Algorithm 2.2.

To achieve such a setting, the classifiers have to be independent, given the class label. A prominent technique to achieve this is to use different feature views (i.e., representations) of the data. This is called in the literature multi-view co-training (Hady et al., 2010b). The optimal case here is of course when the different views occur naturally, for example by the usage of different feature extraction techniques. Alternatively, multiple so called views can be created by randomly sampling elements from the original feature vector (Nigam and Ghani, 2000). In the literature, there are many different techniques to generate multiple views from a single one that try to find optimal splittings based on different criteria (Feger and Koprinska, 2006).

Another crucial parameter of co-training is the accuracy of the base classifiers.

Algorithm 2.2: Co-training (Blum and Mitchell, 1998).

Input:

- a set L of labeled training examples
- a set U of unlabeled examples

Output: Learnt classifier model
Create a pool U' of examples by choosing u samples at random from U;
for *k iterations* **do**
  Use L to train a classifier h_1 that considers only the x_1 portion of x;
  Use L to train a classifier h_2 that considers only the x_2 portion of x;
  Allow h_1 to label p positive and n negative examples from U';
  Allow h_2 to label p positive and n negative examples from U';
  Add these self-labeled examples to L;
  Randomly choose $2p + 2n$ examples from U to replenish U';

If the "teacher" classifier is wrong in its labeling of the samples, that are added to the training data of the other, the performance is degrading. Hence, ensemble techniques are used to improve classification accuracy in order to reduce noise. For example the multiple classifiers can be constructed and data are added to every classifier's training set in turn by the combined ensemble of the remaining ones (Hady and Schwenker, 2010).

2.5.5 Transductive Learning

A further technique to incorporate unlabeled data in classification approaches is transductive learning (Vapnik, 1998), where the given test cases are employed as additionally available unlabeled samples. In inductive learning, a separating decision boundary is explicitly defined as a function on the whole possible space. In contrast to that, a transductive learner attempts to find an optimal assignment of categories only on the given data. Hereby the cluster assumption is exploited: It is more likely for a separating decision border to be in low density regions of the space. This means that the classes are clustered together separately.

The most common transductive learning approach is the transductive support vector machine (TSVM) (Chapelle and Zien, 2005), that aims at finding a low density region by maximization of the margin on all available data, labeled and unlabeled. The margin of the TSVM, that has to be minimized, is hence

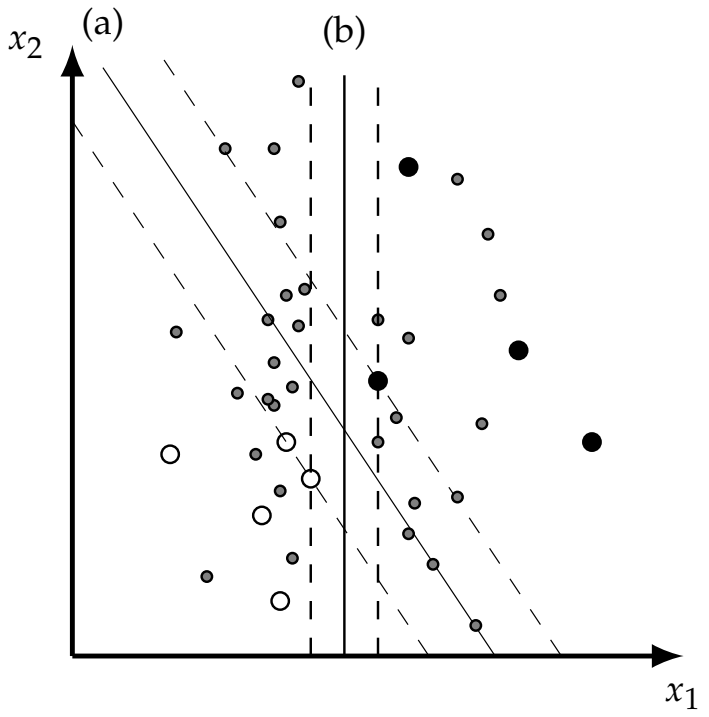


Figure 2.20: *Using only the labeled data samples, that are depicted as white and black circles leads to a different large margin classifier than using unlabeled data. Adapted from (Vapnik, 2006).*

defined as (Joachims, 2006)

$$V(y^*_{u_1}, \dots y^*_{u_k}, w, b) = \frac{1}{2} w^{\mathrm{T}} \cdot w$$

subject to

$$\forall^l_{i=1} : y_{l_i}(w^{\mathrm{T}} x_{l_i} + b) \geq 1$$
$$\forall^k_{j=1} : y^*_{u_j}(w^{\mathrm{T}} x^*_{u_j} + b) \geq 1$$
$$\forall^k_{j=1} : y^*_{u_j} \in \{-1, +1\}.$$

Here, the variables $(x_1, \dots, x_n)$ denote the n data samples together with the given labels $(y_1, \dots, y_n)$. The samples $(x^*_1, \dots, x^*_k)$ are the testing data or additional unlabeled data. The goal is now to find a labeling $(y^*_1, \dots, y^*_k)$, where the margin is maximal.

Figure 2.20 gives a graphical example, why the TSVM works: Using the additional data (gray circles) the optimization renders a different separating hyperplane than using the labeled samples only. According to the cluster assumption the decision border computed with the unlabeled data is more optimal than the one that is rendered using an inductive SVM only for the labeled training data.

2.5.6 Discussion

In this section, the common approaches to incorporate unlabeled data into the training of a statistical classifier have been described. The dominant approach is to iteratively label samples from the given data and add them to the training set. The classifier is then trained again in a supervised fashion using also the new data. Co-training extends this technique to a multiple classifier scheme, which is used to mutually label data.

Conducting a semi-supervised classifier training procedure does not guarantee by any means, that the performance is improved over a solely supervised classifier (Cozman and Cohen, 2002). Measuring classification confidences correctly is crucial in this context. These uncertainty values are further used to access the quality of a classification of a sample. If the classifier returns a high confidence for a false classification, self-training and co-training will fail. Hence, using these techniques, it is in general only feasible to improve a classifier, that is already relatively accurate (Hady, 2011).

3 Applications and Data Collections

In order to study classification architectures in human-computer interaction, the investigation of suitable corpora for classifier design and validation is mandatory. In the following sections three data collections of interactions of human subjects with a computer that are used for the numerical evaluations in this thesis are outlined. Two of these corpora are publicly available for research purposes and the third one was compiled in the context of the SFB/TRR 62 at the Ulm University. The EmoRec and the AVEC 2011 corpora comprise sessions of unconstrained human-computer interaction (compare Sections 3.1 and 3.2), while the third data base is composed of recordings of electroencephalography (EEG) of a subject conducting a visual selection task (compare Section 3.3).

There exist only few other multi-modal corpora of affective human-computer interaction as it is shown for example by Palm and Glodek (2013). The AVEC 2011 data base and the EmoRec corpus are however most fit for the experiments, that are conducted in this work. Limitations that are restricting the feasibility for the investigation of realistic interaction scenarios is for instance the fact that there are only short clips of several dozens seconds in some data collections as it is the case for example in the well-known *humaine* data base (Douglas-Cowie et al., 2007). For other corpora, the respective affective annotations are at the present still subject to debate as for example in the *last minute* corpus (Rösner et al., 2012). Another issue, that jeopardizes the feasibility of a data collection is that it may not be publicity available. An example for that is the so called *Nimitek* corpus (Gnjatović and Rösner, 2010). There are further corpora that are recorded in the context of human-computer interaction that are in principle interesting, but not recorded or labeled under an emotion theoretic paradigm. Prominent cases for such corpora are the venerable *SmartKom* data base (Wahlster, 2006) and the *Pit*-corpus (Scherer et al., 2009).

3.1 EmoRec Data Collection

The EmoRec data collections were recorded in the context of the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems"[1] in the section of Medical Psychology of the Ulm University under the supervision of Prof. Harald C. Traue. The purpose of the experiments was the elicitation of realistic emotions in a human-computer-interaction-scenario by presenting distinct stimuli to the test persons while they were conducting a given task. A major feature of the data is that a many different modalities, namely physiological channels, audio and video, were recorded for analysis.

3.1.1 Recordings

The key idea of the EmoRec recordings is to guide a subject through the PAD space of emotions. The PAD model (Russell, 1980) defines a continuous annotation space of emotions using the dimensions *pleasure*, *arousal* and *dominance*. This model defines different octants in the three-dimensional space, that are graphically displayed in Figure 3.1. Using this model an emotional state can be mapped to a distinct location in the space that is spanned by these dimensions. In this data collection, the PAD space is simplified into eight octants (for example "low pleasure, high arousal, low dominance" versus "high pleasure, low arousal, high dominance") in order to render distinguishable emotional states.

For the recordings, the test persons were seated alone in a closed room in front of a computer screen displaying a memory training task. This task was to solve multiple games of Concentration[2]. Here cards are faced downwards on a table – or a computer screen in this setting – and the player has to find pairs of cards showing the same image by successively turning over two cards. After every round the cards are turned downwards again, except if a pair of cards is uncovered. After that, a new round of the game begins. An example of the experimental setting and the computer screen as it is displayed to the test person is shown in Figure 3.2. Overall, 152 subjects passed the experiment in three successive inquiries. The test persons are between 20 and 75 years of age, and 31 % of the them are male and 69 % are female.

In order to turn a card, the participant was instructed to utter its coordinates on the board to the voice controlled dialog system: for example "A 1", "C 4" and so on. To provide an operational system to the user, the speech recognition, the selection of the card and the feedback of the system were emulated in

[1] www.sfb-trr-62.de

[2] http://en.wikipedia.org/wiki/Concentration_(game)

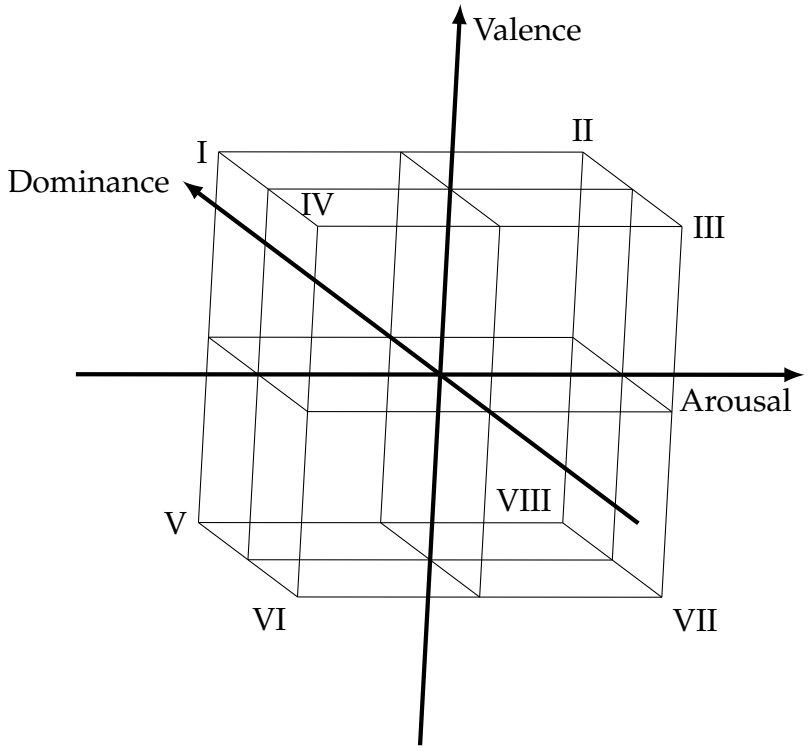


Figure 3.1: *The PAD model spans a 3-dimensional space, which can be divided into eight distinct octants. Adapted from (Walter et al., 2013).*

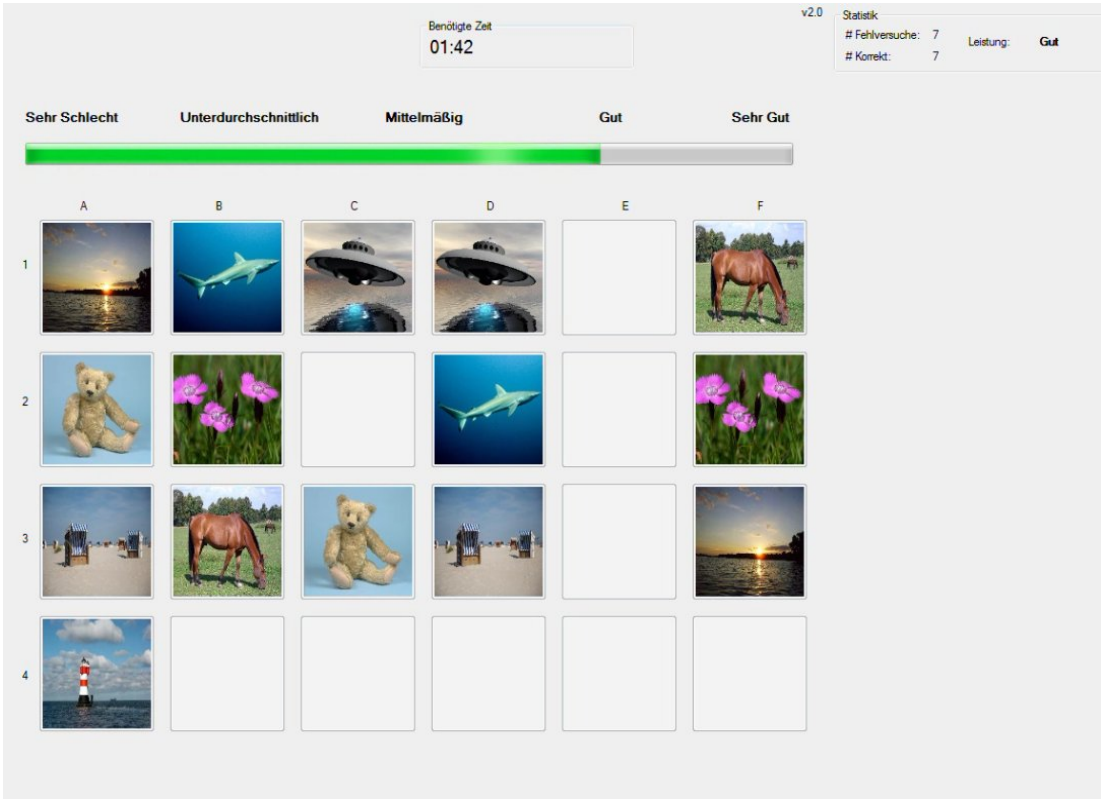


Figure 3.2: *The content of the computer screen as it is displayed to a test person: in the middle the cards that have to be turned are shown. The coordinates on the upper and left-hand side are used to reference the individual card by voice. The upper part displays the feedback given by the wizard together with the remaining time for the task.*

Figure 3.3: *A test person undergoing the EmoRec memory test. The EMG electrodes are attached on the forehead and on the cheek. On the left hand of the subject, the BVP-senor is attached to the middle finger and the SCL sensor is placed near the wrist. The subject also wears the cap that locates the EEG-electrodes on the subject's skull. Taken from (Walter et al., 2013).*

a so called *Wizard-of-Oz* (WoZ) setting (Kelley, 1983). The system, the test person is interacting with, is remotely controlled by a human experimenter using another computer in a separate room. In this way, a fully operational cognitive technical system that can adequately respond to the user is simulated. WoZ experiments are commonly used to study human-computer interaction without spending too much time for the implementation of the logic but still rendering a powerful and flexible setup. This experimental paradigm also allows to provide functionalities for a framework that are not yet fully implemented in an autonomous computer program.

Six physiological measures were recorded from the subjects together with the spoken utterances and a video portraying the subject's face frontally (compare Figure 3.3). The *heart rate variability* (HRV) is measured using an optical device that is attached to a finger. The subject's heartbeats are deduced from the translucence of the finger using a light source and a photo cell. To measure the *respiration* (Resp) of the subject, a special belt around the chest is used that expands or contracts when the subjects inhales or exhales. Furthermore the skin conductance level (SCL) is measured using electrodes on the back of a hand, where constant electrical current of 10 μA is conducted. The electric re-

PAD	+ - +	+ - +	+ + -	- - -	- + -	+ - +
Intro	ES-1	ES-2	ES-3	ES-4	ES-5	ES-6
1.5 min	approx. 3 min	approx. 3 min	approx. 5 min	approx. 3 min	approx. 4 min	approx. 5 min

time ⟶

Figure 3.4: *Experimental procedure for the EmoRec recordings: After a short introduction of the task, the subject is guided through six emotional experimental sequences (ES-1 to ES-6). The respective label in PAD space are given in the top line of the figure in terms of pluses (+) and minuses (-), where a plus signifies a high and minus a low value for the respective dimension. In short the emotional progression aims at "positive" states in the beginning succeeded by a more "negative" state and ending positively again.*

sistance of the skin is augmented by the sweat a test person oozes. Hence the transpiration of the subject is measured by the electrical conductivity between the electrodes. For two facial muscles an *electromyography* (EMG) is conducted: the *musculus corrugator supercilii* and the *musculus zygomaticus major* (Van Boxtel, 2010). The electrodes are placed over the eyebrow and the cheek of the subject. These electrodes measure the electrical potentials, that arise on muscle contractions (approx. 500 μV). The musculus corrugator supercilii shifts the eyebrow downwards on contraction. It is also used for wrinkling the forehead. The musculus zygomaticus major pulls the corner of the mouth upwards and outwards of the face. Finally, an electroencephalography (EEG) is recorded. The test persons wear a cap that fixates the electrodes on its skull during the experiment. The EEG recordings in this corpus are however not considered in this thesis.

In the literature, physiological signals are described as strong evidences for emotional processes, in particular for the arousal of a subject (Picard et al., 2001b; Stemmler et al., 2001; Christie and Friedman, 2004). For example is the frequency and the amplitudes of the SCL deflections, that are characteristic for this signal, a hint for the arousal of the subject. The same holds true for the heart rate of a person.

There has been work of other groups on classifying physiological signals, for example as described by Kim and André (2008). They use the playback of pieces of music that have a high personal value for the test subject to elicit four different category of emotions that are constituted by the quadrants of the 2-dimensional valence-arousal space. Here, the classification is conducted in a leave-one-sample-out procedure, which yields high recognition rates.

3.1.2 Annotations

Successive tasks were used to induce different emotional states to the subject in the order sketched in Figure 3.4 (Walter et al., 2011). To do so, different stimuli were presented to the subject. The system responses, that are designed to elicit negative emotions were amongst others dispraise, time pressure to the subject by setting a maximal duration for the task and wrongly or delayed execution of commands. Furthermore, the system asks the subjects to cancel the task. The respective stimuli for positive emotions were for example praising the subject or presenting an easier game with fewer cards. By these means, the subjects were passed through six experimental sequences (ES), which induce different states in the PAD space. The respective targeted affective subject states are indicated in the top row of Figure 3.4. The ES last about 3-5 minutes each depending on the difficulty of the board of the game. Before the experiment begins a short explanation of the task is given, which is depicted with the term "Intro". The subsequent ES-1 and ES-2 are designed to induce "high pleasure", "low arousal" and "high dominance". ES-1 is also inserted that the subjects get used to the task and sometimes further instructions are given by the experimenter. ES-3 is designed as a transition to the subsequent "negative" sequences. ES-4 and ES-5 are labeled as "low pleasure", "low arousal", "low dominance" and "low pleasure", "high arousal", "low dominance". ES-5 lasts a little longer than the preceding sequences as the board is bigger and it takes more time to solve the task. Finally, the "positive" sequence ES-6 , which translates into "high pleasure", "low arousal" and "high dominance" in the PAD space, is conducted to release the test persons happily from the experiment.

The classification task posed by this corpus is to discriminate the samples of ES-2 and ES-5. These two experimental sequences are designed to elicit rather complementary emotions: "high pleasure/low arousal/high dominance" versus "low pleasure/high arousal/low dominance", or short: Positive vs. negative emotion. The dimensional labels of the ES are given in the top row of the Figure 3.4. The according stimuli, presented to the subject were praises and a small board of concentration and hence an easy game for the positive sequence. In case of the negative class, the user is given a bigger board and only displeasing feedback: for example the user is criticized for his execution of the game and the subject is exposed to time pressure. Figure 3.2 displays the screen content, that is shown to the test subject. Beside the card board, a judgment of the performance of the subject and a timer is presented to the user. This procedure is designed to embed the antagonizing sequences ES-2 and ES-5 into a continuous experiment and leaving enough time between them. A detailed description of the experiment can be found in (Walter et al., 2013).

3.1.3 Physiological Channels and Features

From the SCL, the Resp, the BVP and the two EMG signals, various features were extracted on different time scales. Before the feature extraction, the data has to be carefully preprocessed. The physiological sensors show strong long term characteristics during the experiments for different reasons such as sensors may slide a little bit on the skin or the gel, that establishes the conductivity between the skin and the electrode might gradually dry out. Furthermore the sensors are also sensitive to electromagnetic noise from the surroundings. In general, a slow low- or band-pass filter is applied together with a linear piece-wise detrend (i.e., subtracting piecewise a linear least-squares-fit from the respective chunk of the data) of the time series at a 10 second basis.

The features introduced in the following were inspired by the ones described by Kim and André (2008) and were implemented in the course of Markus Kächele's Diploma Thesis (Kächele, 2011) at the Institute of Neural Information Processing of the Ulm University.

3.1.3.1 Heart Rate-related Features

Before the computation of the features a preprocessing step is conducted: after the detrending of the signal, a low pass filter at 5 Hz is applied. All heart rate related features are computed based on a 25 s time window with a 12 s offset. This duration is a trade-off between defining a window that is short enough to fit several windows in an ES and the requirement to enclose multiple heartbeats to compute meaningful statics.

The key to characterize the heart rate from the recorded blood volume pulse is to find the well known QRS complex in the signal as described for example in (Kamath and Ananthapadmanabhayuyu, 2007; Krikler, 1990). An example for a QRS complex is shown in Figure 3.5. In order to automatically compute the QRS timings an algorithm based on discrete wavelet transformation and morphological filters was used (Rudnicki and Strumiłło, 2007; Pan and Tompkins, 1985)

Based in this, the *standard deviation of the RR intervals* in a time window is defined as the first basic feature (Simson, 1981). Further, the heart rate variability (HRV) is defined as the derivation of length of RR intervals (Malik, 1996). The *standard deviation of the HRV* in a time window is used as a second feature (Sayers, 1973). The NN50 index counts the number successive RR intervals that differ more than 50 ms. When the NN50 is normed with the length of the RR interval sequence, the so called *pNN50* is defined. The pNN50 is used as a third feature for classification (Mietus et al., 2007). The *RMSSD* is defined as

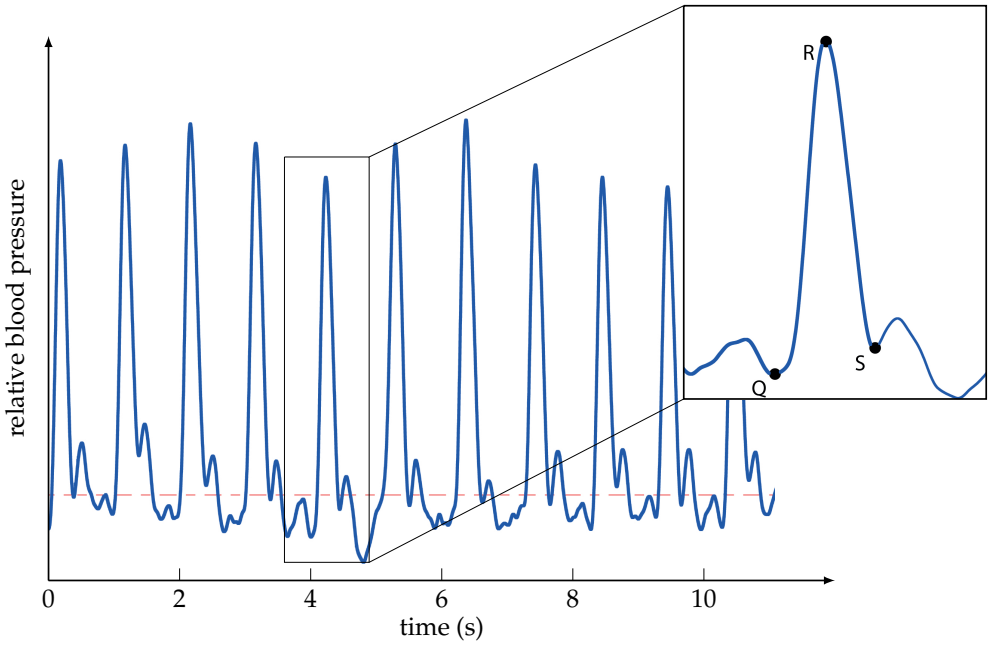


Figure 3.5: *The QRS complex in the HRV data. Adapted from (Kächele, 2011).*

the square root of the mean squared differences of successive RR intervals:

$$\text{RMSSD} = \sqrt{\frac{\sum_{i=1}^{N-1} \left(\text{RR}(i) - \text{RR}(i+1)\right)^2}{N-1}}$$

Another heart rate related feature is based on the *approximate entropy* (ApEn) that was defined by Pincus (1991) and applied to physiological data for example by Richman and Moorman (2000) and Pincus and Goldberger (1994). In order to define the ApEn suppose N measurements $u(1), u(2), \ldots, u(N)$ are given. Based on these $N - m + 1$ data windows of length m are defined as follows: $x(i) = u(i), u(i+1), \ldots, u(i+m-1)$. Furthermore a distance measure $d\left(x(i), x(j)\right)$ is defined:

$$d\left(x(i), x(j)\right) = \max_{k=1,\ldots,m} \left(|u(i+k-1) - u(j+k-1)|\right).$$

Hence, the relative frequency to find a vector with distance r is defined as

$$C_i^m(r) = \frac{(\text{number of } j \text{ such that } d\left(x(i), x(j)\right) \leq r)}{N-m+1}.$$

With

$$\Phi^m(r) = \frac{\sum_{i=1}^{N-m+1} \log C_i^m(r)}{N-m+1},$$

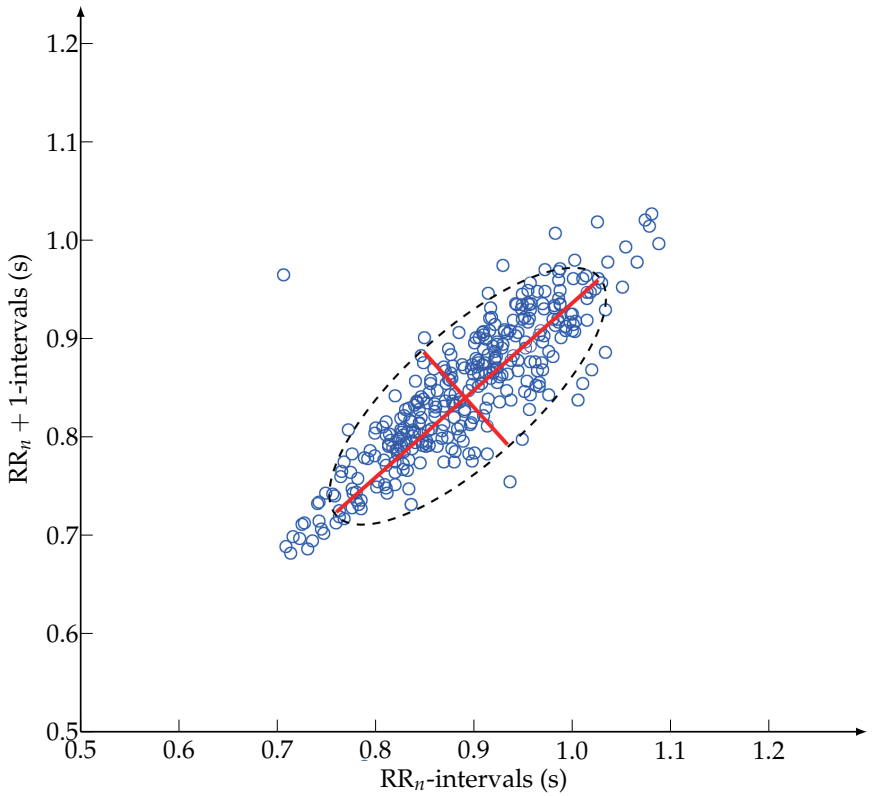


Figure 3.6: *Poincaré plot of for the RR-intervals for a real subject of the EmoRec corpus. The two principal axis of the ellipse are sketched in red. Taken from (Kächele, 2011).*

the approximate entropy is defined as:

$$\mathrm{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r).$$

In order to define meaningful features, the parameters m and r are set to 2 and 0.2 according to the literature.

The *recurrence rate* is defined using the so called recurrence plot (Eckmann et al., 1987). Suppose a time series $X = x(1), x(2), \ldots, x(n)$ and the set of all possible sub-sequences $Y_m = y_m(1), y_m(2), ...y_m(n - m + 1)$ of length $m > 0$. Then a recurrence plot is defined as the matrix that contains 1 at position (i, j), if $\|y_m(i) - y_m(j)\| < r$. Hence, the recurrence rate REC is defined as (Mewett et al., 1999):

$$\mathrm{REC}(m) = \frac{|\{(i,j)|\|y_m(i) - y_m(j)\| < r\}|}{(n - m + 1)^2}.$$

Here, the parameters are set to $m = 2\, r = 2\sigma$, where σ is the standard deviation of the samples in the time window.

Another feature of the heart rate is derived from the so called *Poincaré plot* (Karmakar et al., 2011). Here, the duration of an RR interval is plotted against the

length of the succeeding RR interval (compare Figure 3.6). The quotient of the two principal axis of the ellipse that is computed via least squares optimization to the data points is uses as feature as proposed for example by Kim and André (2008).

Further features are computed from in the frequency domain using the power spectral density (Baumert et al., 1995). Following the approach of Kim and André (2008) three separate frequency bands are defined: the "very low" (i.e., 0.003–0.04 Hz), the "low" (i.e., 0.04–0.15 Hz) and the "high" (i.e., 0.15–0.4 Hz) frequency band. The average of the high and the average of the low frequency bands were used together with the ratio of the two as features for classification.

3.1.3.2 Respiration-related Features

Features from the subject's respiration are measured using long time windows of 20 s duration with an offset of 10 s as the respiratory patterns of inhaling and exhaling are in general repeated for about 15–25 cycles per minute (Boiten et al., 1994; Bloch et al., 1991). As pre-processing steps, a linear detrending approach and a low pass filter with a cut-off frequency of 0.15 Hz (Kim and André, 2008) are applied to the raw data.

The first two features that are calculated are the *mean* and the *standard deviation* of the *first derivative* of the pre-processed signal (Kim and André, 2008). Further features are derived from the inter-peak statistics of the signal: the *mean* and the *standard deviation* of the duration of the breath intervals. Using the inter-peak timings, a Poincaré analysis analogous to the one described in Section 3.1.3.1 is conducted. Hence the *ratio of the main axes* of the resulting ellipse is used as further feature.

Another feature is the *breathing volume* of the subject, which is straightforward defined as the following integral

$$\mathrm{BV}(i) = \int_{\mathrm{valley}_i}^{\mathrm{peak}_i} |\mathrm{resp}(x)| dx.$$

It equals the area under the curve from the valley of the i^{th} breath to its peak. As the sensors return discrete values for each sample time-step the equation degenerates to the mean of the absolute values between a valley and the respective peak.

3.1.3.3 EMG-related Features

The contraction or relaxation of the muscle is perceivable in an oscillation of the EMG signal. Thus, a band-pass filter at 20–120 Hz is applied to filter out

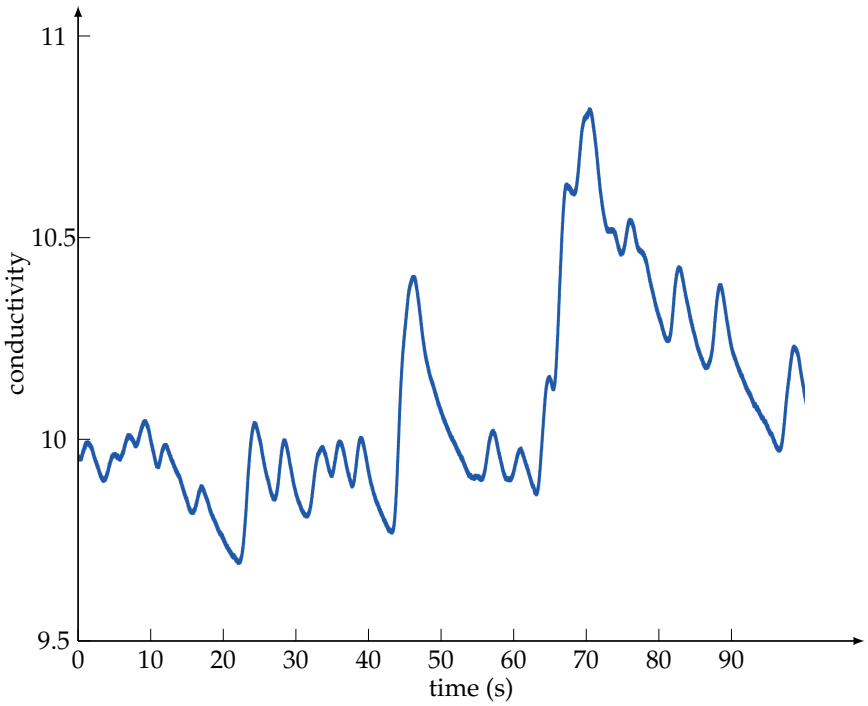


Figure 3.7: *A typical SCL curve with multiple spikes. Adapted from (Kächele, 2011).*

high and low frequency noise (Bruns and Praun, 2002). Two different kinds of features were computed on a 5 s and 20 s sliding window with offsets 2 s and 10 s. The *mean* of *first* and *second derivatives* were used as features using the shorter time windows.

Based on the 20 s time windows, a *power spectrum density* analysis (Christensen and Fuglsang-Frederiksen, 1986) was conducted. For that purpose, the spectrum from 20 Hz to 120 Hz was divided into 8 equal frequency bands with a 50 % overlap each. For each frequency band, the mean spectral power was used together with the ratio of the mean spectral power of the highest and the lowest frequency band.

3.1.3.4 SCL-related Features

The signal derived from the SCL shows spikes that are rapidly rising on a stimulus and are after climaxing comparably slowly declining (Darrow, 1937). An exemplary plot for an SCL signal is displayed in the Figure 3.7. It is also possible that further spikes occur before the latter is relieved, and hence, a superposition of the signals is observable. In order to properly capture the phenomena, high frequency noise is removed with a low pass filter at 0.2 Hz (Kim and André, 2008).

For the pre-processed signal an inter-peak statistics, i.e., the durations between two successive peaks, was computed. In a 20 s sliding window with an offset of 10 s the *number of peak occurrences* is determined. The result is further normalized with the length of the window and used as a first SCL feature. The second feature is the *average peak magnitude* in the window.

Additionally features using the derivative of the signal are used. To ensure that positive and negative gradients do not nullify each other, the negative derivatives are set to 0. Hence, the *mean* and the *standard deviation* of the *first* and *second derivative* are calculated in a 5 s window with a 2 s offset.

3.2 "AVEC 2011" Data Collection

In the following the data set provided with the Audio/Visual Emotion Challenge (AVEC) 2011 that was held in conjunction with the ACII 2011 workshop is described (Schuller et al., 2011; McKeown et al., 2010). The data collection is a subset of the "SEMAINE" data set was collected to study human interaction with a virtual agent, that was recorded in connection to the EU project having the same name[3]. For this project a so called Sensitive Artificial listener (Douglas-Cowie et al., 2008) was developed. Test persons were recorded while interacting with a 3D animated virtual character, that is designed to show and elicit certain affective behaviors (Douglas-Cowie et al., 2008):

- Prudence "who aims to make people pragmatic"
- Poppy "who aims to make people happy"
- Spike "who aims to make people angry". A sample image of this avatar is shown in Figure 3.8.
- Obidiah "who aims to make people gloomy"

The system had several dialogs prepared in order to encourage the participant to interact, for example letting the character say thing like "Go on, tell me your news" or "Have you done anything interestingly lately?" (Schröder et al., 2012). Furthermore certain phrases are designed for the respective affective color of the avatar, for example "Don't get too excited" for Obidiah and "Life's a war, you're either a winner or a loser" for Spike (Schröder et al., 2012).

The part of the SEMAINE data, that is used the AVEC 2011 is called "Solid SAL" in (Wöllmer et al., 2013), which indicates that it is recorded in a WoZ setting. Over-all three sub-challenges were proposed within the AVEC 2011

[3] `http://www.semaine-project.eu`

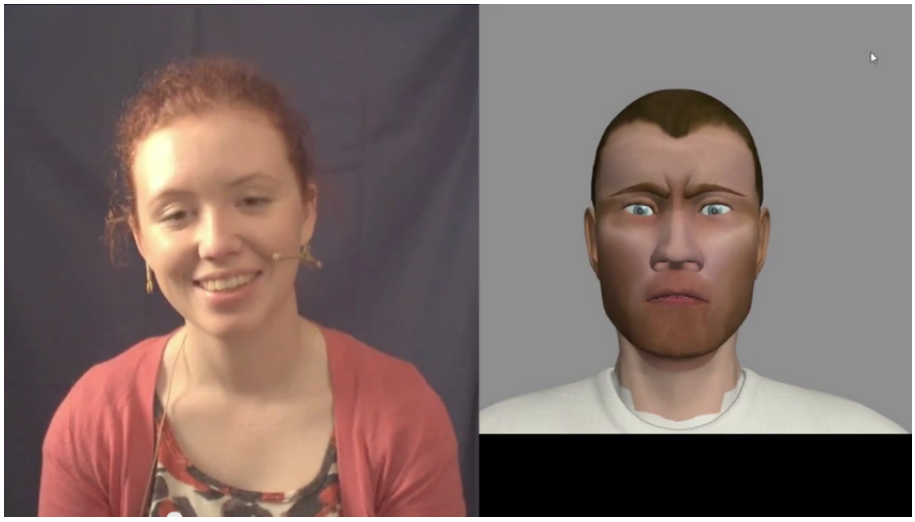


Figure 3.8: *A test person interacting with the Sensitive Artificial Listener. The virtual character "Spike" (right-hand side) is designed to aim to make people angry (McKeown, 2011).*

competition: an audio challenge on word-level, a video challenge on frame-level and an audiovisual also on video frame-level.

The recorded data was labeled in four affective dimensions: *arousal*, *expectancy*, *power* and *valence*. A brief explanation of the labels is given in Section 3.2.2. Two of the categories were annotated at a time using the so called "FEELTRACE" tool (Schröder et al., 2000). Here, a two dimensional canvas was presented to the labeler and by moving a pointer in the appropriate direction a continuous label is set. The annotations of the raters have been averaged for each dimensions resulting in a real value for each time step. Subsequently, the labels are binarized using a threshold equal to the mean of each dimension for all sessions. Every recording was annotated by at least two but no more than eight raters. Along with the sensor data and annotation, a word-by-word transcription of the spoken language was provided which partitions the dialog into conversational turns in which the participant and the machine agent are alternating. For the evaluation of the challenge only arousal was taken into account as classification of the other dimensions yielded poor results[4]. The data was partitioned into training set, development set and test set (compare Table 3.1) of 31 or 32 interactions of several minutes duration each.

3.2.1 Recordings

Audio and video material is available in over-all 63 recordings from 13 different subjects (compare Table 3.1). The data is partitioned for the official chal-

[4] `http://sspnet.eu/avec2011/`

Table 3.1: *Technical overview of the AVEC 2011 recordings. Taken from (Schuller et al., 2011).*

	Train	Development	Test	Total
Sessions	31	32	32	95
Frames	501277	449074	407772	1358123
Words	20183	16311	13856	50350
Total duration (h:m:s)	2:47:10	2:29:45	2:15:59	7:32:54
Avg. word duration (ms)	262	276	249	263

lenge into a train, development and test set. The general technical setup of the recordings is described in the following.

The video channel was recorded in the following configuration:

- approximately 50 frames per second
- 780 $\times$ 580 pixels camera resolution
- 8 bits for three color channels

The audio channel was recorded as in the following setup:

- sample rate of 48 Hz
- 24 bits per sample

An elaborate synchronization using hardware triggers for cameras, that were also recorded as a wave signal to synchronize the audio channel was realized for the corpus. Please refer to (Lichtenauer et al., 2009) for a detailed description of the set-up. This enabled a frame-wise annotation of the data. A further label, that is provided with the corpus is the timings of the uttered words and also the conversational turns of the interlocutors.

3.2.2 Annotations

The labels the used in the SEMAINE data base originate from a questionnaire study by Fontaine et al. (2007), where certain terms were assigned to affective labels. Precisely, 24 terms, that are frequently used by researchers in affective sciences and also in daily life (for example "anger", "sadness", "irritation") were presented to test persons. The task was to rate those terms according to 144 so called "emotion features", for example "pressed lips together", "felt at ease" and "caused by chance". The categorization was conducted in nine steps from "extremely unlikely" to "extremely likely", i.e., a rating how probable a distinct emotion feature is for a presented term has to be estimated. Over-all

the study has been conducted with 531 participants form Belgium, the United Kingdom and Switzerland.

The resulting 24×144 data matrices are analyzed using a principal component analysis. The top 4 principal components define 4 affective dimensions (in order of importance): "evaluation-pleasantness", "potency-control", "activation-arousal" and "unpredictably". The scores of the new dimensions for the features are found in (Fontaine et al., 2007). Further, in (Fontaine et al., 2007) the embedding of the emotional terms into the new space is displayed graphically, which provides a further insight into the semantics of the labels.

3.2.3 Audio Features

For the audio analysis we extracted a variety of standard features, that are used for spoken language understanding and transmission. They are for example prominently described in speech processing textbooks (Huang et al., 2001) or (Wendemuth, 2004). The extracted features are the fundamental frequency, LPC, MFCC coefficients and finally RASTA-PLP.

3.2.3.1 Fundamental Frequency (f_0) and Energy

From each speech segments the f_0 values are extracted, using the f_0 tracker available in the ESPS/*waves*+[5] software package. This f_0 track as well as the energy of the plain wave signal is extracted from 32 ms frames with an offset of 16 ms.

3.2.3.2 Linear Predictive Coding Coefficients

The extraction of linear predictive coding coefficients (LPC) is an auto regressive approach, where the $n-$ th sample of a time series is approximated using a function of the p preceding samples (Atal and Hanauer, 1971):

$$\hat{s}_n = \sum_{k=1}^{p} \alpha_k s_{n-k}.$$

The prediction error E is hence defined as

$$E_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} \alpha_k s_{n-k}.$$

[5] http://www.speech.kth.se/software/

Setting the partial derivatives for the coefficients α_k to zero is a straightforward way to optimize the error. As the matrix of coefficients is symmetric and positive definite, this can be computed efficiently.

LPC are still widely applied in speech processing, for example speech recognition and speech synthesis. One reason for this is that they are easily computed as, unlike other more computationally elaborate features, no Fourier analysis has to be conducted.

For the speech classification in this work, 8 LPC were computed for time windows of 32 ms length with an offset of 16 ms.

3.2.3.3 Mel Frequency Cepstral Coefficients

The mel frequency cepstral coefficients (MFCC) are computed as suggested by Fang et al. (2001):

1. The short term power spectrum is computed using the discrete Fourier transformation from windows of speech.
2. The frequency axis of the power spectrum is converted to mel-frequency:
$$M(f) = 2595 \log\left(1 + \frac{f}{100}\right).$$
3. Triangular band pass filters are applied and the log energy of every filter output is computed.
4. The de-correlated cepstral coefficients are computed using discrete cosine transform (DCT):
$$\text{MFCC}_i = \sum_{k=1}^{N} X_k \cos\left(X_i \left(k - \frac{1}{2}\right) \frac{\pi}{N}\right).$$
Here, $i = 1, 2, \ldots, M$ is the number of cepstrum coefficients, $X_k, k = 1, 2, \ldots, N$ is the logarithm of the energy of the k^{th} filter.

Using the triangular filters in mel-scale, i.e., higher densities of filters with smaller bandwidths for lower frequencies than for higher frequencies, is a feature, that is adopted from the human ear (Davis and Mermelstein, 1980). The 0-th coefficient remains often unconsidered as it can be seen as the average energies of each frequency band (Fang et al., 2001). Furthermore, $M \ll N$ is chosen in order to conduct a compression of the signal (Fang et al., 2001).

In this thesis, 20 MFCC coefficients per time window of length 32 ms with an offset of 16 ms are extracted.

3.2.3.4 Relative Spectral Perceptual Linear Predictive Coding

The perceptual linear predictive coding (PLP) is a further technique to compute characteristics for speech signals (Hermansky, 1990; Robinson and Dadson, 1956). The PLP is computed as described by Hermansky (1990):

A so-called "critical band filtering", that is similar to the triangular filtering described earlier is conducted. Sometimes trapezoidal filter banks are used instead of triangular filters. Then, an equal loudness pre-emphasis, is applied following the findings of Robinson and Dadson (1956), who model the human perception of loudness for different frequencies as follows:

$$E(w) = 1.151 \cdot \sqrt{\frac{(w^2 + 144 \cdot 10^4) \cdot w^2}{(w^2 + 16 \cdot 10^4) \cdot (w^2 + 961 \cdot 10^4)}},$$

where w is the frequency and E is the perceived loudness. Then an intensity-loudness conversion is conducted that further accounts for the nonlinear relation between the intensity and the perceived loudness by applying the cubic root for the amplitudes. Subsequently the inverse discrete Fourier transform is applied, and finally an autoregressive approach is conducted analogously to the LPC mentioned earlier.

The relative spectral (RASTA) extension for the PLP was introduced by Hermansky et al. (1992). This makes the result more robust to linear spectral distortions, for example different microphones. To do so, a logarithm is taken after the critical band analysis. After the RASTA filtering, the result is transformed back from the logarithmic domain using the exponential function.

Over-all 257 coefficients were computed every 16 ms for windows of length 32 ms.

3.2.4 Video Features

In order to process the video channel, the well-known *computer expression recognition toolbox* (CERT) by Littlewort et al. (2011) is used. A screen-shot of this toolbox is shown in Figure 3.9, where it is applied to a subject from the AVEC 2011 data collection. CERT computes human emotion related categories from subjects that look approximately frontally into the camera as it is the case for the data described in Sections 3.1 and 3.2.

CERT returns 6 high-level descriptions for an input video:

- A probability value for the six basic emotions each (i.e., anger, fear, sadness, happiness, disgust and surprise) according to Ekman (1993).

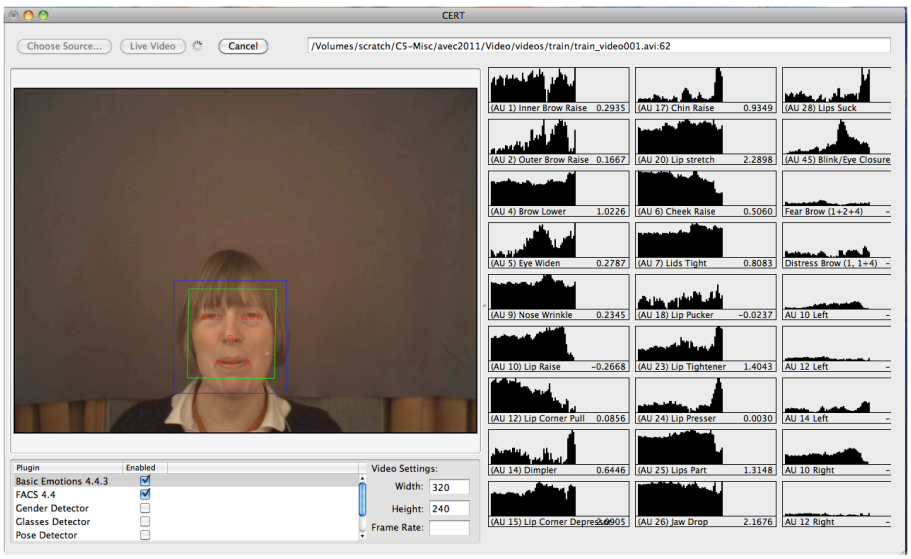


Figure 3.9: *Screen capture of the CERT software for a sequence taken from the "AVEC 2011" data collection. The respective subject is depicted on the left hand side of the figure. The values for the categories, that are extracted from the video are plotted on the right side of the figure.*

- Intensities for 19 action units, that are defined in (Ekman and Friesen, 1978). Examples for prominent action units are shown in Table 3.2 and Figure 3.10.
- A smile detector with intensity.
- A detector whether the subject wears glasses.
- An estimate of the subjects gender.
- 2D position of 10 facial reference points.
- 3D estimation of head pose (yaw, pitch and roll).

The classification is conducted in several steps (Littlewort et al., 2011), that are described in the following. First a face detector using a Viola-Jones cascade (Viola and Jones, 2001) is applied. Based on this, 10 facial regions, i.e., inner eye corners, outer eye corners and eye centers, the nose tip, left mouth corner, right mouth corner and the center of the mouth, are detected using a further classifier cascade. Additionally a post-processing step using a linear regression based on a human labeled data set is conducted to improve the segmentation. Using these characteristic 10 positions, the bounding box, that is returned by the cascade is rotated in order to estimate a better fit for the face. Subsequently, 72 Garbor filters are computed for 8 orientations and 9 spatial frequencies from

Table 3.2: *Ekman and Friesen (1978) define the Facial Action Coding System (FACS), that assigns codes to the contractions of the facial muscles. The table shows different important FACS codes and their correspondent facial movement. Taken from* `http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm`.

AU code	Peculiarity
1	Inner Brow Raiser
2	Outer Brow Raiser
4	Brow Lowerer
5	Upper Lid Raiser
9	Nose Wrinkler
10	Upper Lip Raiser
12	Lip Corner Puller
14	Dimpler
15	Lip Corner Depressor
17	Chin Raiser
20	Lip stretcher
6	Cheek Raiser
7	Lid Tightener
18	Lip Puckerer
23	Lip Tightener
24	Lip Pressor
25	Lips part
26	Jaw Drop

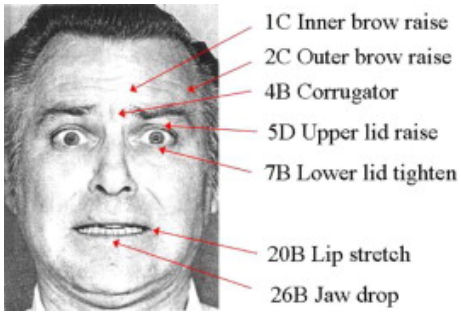


Figure 3.10: *A famous example for the assignment of action units to facial regions. The FACS codes signify different contractions of the muscles of the face (compare right hand side of the figure). A facial expression can hence be denoted by an enumeration of FACS codes. Taken from (Littlewort et al., 2009).*

every image. A linear SVM is trained for the detection of action units (AU) using mostly publicly available data sets such as the "Cohn Kanade Comprehensive Data Set for Facial Expressions" by Kanade et al. (2000). The intensities mentioned before are given as the distance to the decision hyperplane of the respective SVM, which is a popular technique to estimate the confidence of a classifier as described in Section 2.3.1. The smile detection is conducted using Haar-like features and a boosting approach. To obtain an estimate for the basic emotion, a logistic regression, that maps the estimates for the intensities for the AU to the 6 classes. This mapping is trained only on the Cohn-Kanade data collection, that incorporates an annotation of both AU and basic emotion.

In this work the outputs of the CERT software are used as features to further classify the emotional categories of the AVEC 2011 and the EmoRec data. The output of the modules "Basic Emotions 4.4.3" (i.e., Ekman's basic emotions), "FACS 4.4" and "Unilaterals" (i.e., the selection of AU shown in Table 3.2) and "Smile Detector" are considered. Thus an over-all 36-dimensional vector for every frame was obtained. CERT obviously only delivers sound values in case the face of a subject is recognized. Due to the unrestricted settings of analyzed corpora (subjects may turn away or leave the visual range of the camera) features and hence classification results may be missing for certain samples.

3.3 Pascal 2 "mind reading" data set

This section describes a corpus that comprises EEG signals, that were recorded from a subject in a visual selection task using a computer screen. Even though the corpus was compiled in a human-computer interaction scenario, it inherits a number of special features by which it is distinguished from the previous ones. Unlike the other data collection, only a single modality is captured in

this corpus even though multiple electrodes are used that deliver a signal. A further property is that due to the elicitation method that is used, the two different classes occur very imbalanced in the corpus. Finally, the corpus inherits a very strict timing of the classes by the succession of the presentations of the images on the screen.

3.3.1 Recordings

In the past few years brain computer interfaces became part of the most prominent applications in neuroscience (Chumerin et al., 2009). In the present study the goal is to investigate the possibility to automatically determine whether a human subject has just seen a target on a presented image by solely analyzing the event related potentials (ERP), that are recorded using electroencephalography (EEG). ERP typically reflect cognitive processes in the brain that follow a more or less strict timely pattern that can be visualized by filtering and averaging ERP signal recordings (Gray et al., 2004). A prototypical EEG progression can be seen in Figure 3.11. Well established states that are passed during this information process resembled in the signal are the P100, N200 and the well known P300, all named after their voltage and approximate delay of the stimulus-response (Gray et al., 2004; Dujardin et al., 1993). Furthermore, actual amplitudes and latencies in the typical ERP such as the N200 or the P300 are dependent on factors such as the subjects attention, age, the stimulus modality (for example audio or visual), and the frequency of the stimulus (Dujardin et al., 1993). Another source of ERP are motor signals which correspond to a task related physical action of a subject, for example pressing a button. Such a potential may be even further delayed after the stimulus and the aforementioned patterns. These bio-electrical phenomena are normally overlaid with heavy noise, that is caused inevitably even by subtle movements of a test person (for example heartbeats). To make the actual ERP visible a denoising technique called ensemble averaging (Sörnmo and Laguna, 2005) is applied in physiology: for all sequences of a category, the subsequent samples after a stimulus are averaged. In the present investigation visual stimuli were presented by following a typical oddball paradigm: the non target (background) type stimuli were presented very frequently, whereas the targets were displayed very rarely. According to Dujardin et al. (1993) this type of experimental setup should lead to a prominent P300 representation in the EEG.

Originating from these findings, it is compelling to design a machine classifier capable of detecting the subject's recognition of a target stimulus by monitoring the bio-electrical EEG stream. This particular setup imposes several challenges: the oddball recording technique of the data requires a special treatment due to the skewed distribution of classes. Heavily imbalanced datasets require

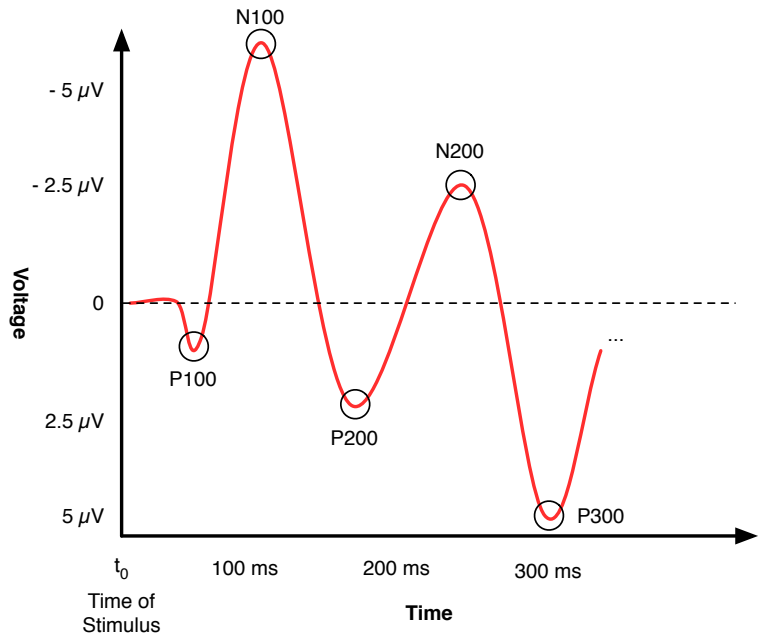


Figure 3.11: *Schematic presentation of the succession of bio-electrical artifacts after a stimulus. The letter denotes the sign of the voltage whereas the following number indicates the approximate delay after the trigger (plot adapted from Birbaumer and Schmidt, 2006).*

special treatment in order to mitigate the over-representation of a class. Popular techniques are under- and over-sampling of the training set with respect to the categories or the usage of error functions that account for skew distributions of classes (Japkowicz, 2000; Zhou and Liu, 2006). Also, the noisy nature of the employed sensors can impair the recognition performance. Methods designed to improve robustness in low signal to noise ratio conditions include low pass filtering but also information or data fusion (Kuncheva, 2002a). In this particular domain of information fusion various possibilities to ensure robustness can be applied.

In this study, the dataset provided by the "Machine Learning for Signal Processing 2010 Competition: Mind Reading[6]" is utilized. The goal of the competition is the classification of stimuli by analyzing EEG recordings. For these recordings, satellite images were presented in a fast sequence to a test person, who was instructed to push a button when a surface to air missile (SAM) site was shown. The images were shown to the subject in a resolution of 500×500 pixels for 100 ms. The data is presented in 75 blocks of 37 images leading to a total number of 2775 images. Each block is separated by a pause ended by the subject independently. However, only a marginal fraction of the images are actual triggers (i.e., a SAM site is shown) such that the task of identification

[6] `http://www.bme.ogi.edu/~hildk/mlsp2010Competition.html`

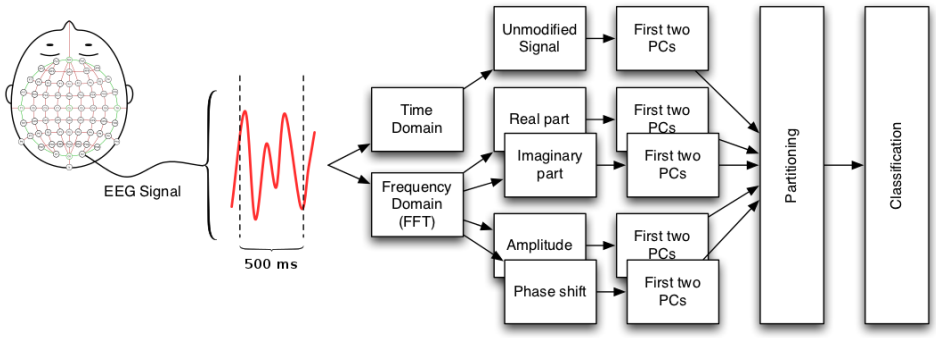


Figure 3.12: *Feature extraction procedure: the subsequent half of a second is processed in time and frequency domain. Overall,* $16 \times 5 \times 2 = 160$ *different features per partition (see Figure 3.13) are passed to the classification architecture.*

can be seen as a typical oddball paradigm task (Segalowitz et al., 2001). Out of the 58 triggers only 48 triggers were identified by the subject within a reasonable time window after the presentation of the satellite image containing an actual missile site. For the testing and training only the EEG data recorded after these 48 correctly identified triggers are used in order to ensure that the subject has actually found the target and therefore generating a meaningful EEG phenomenon.

The EEG data consists of 64 channels in total that are recorded with a sampling rate of 256 Hz. The sensors are arranged as it is shown in the center of Figure 3.13. Along with the data from these 64 electrodes the onset time of the pressed space bar and the type of displayed image ("trigger" or "no trigger") are provided in the dataset.

3.3.2 Features

In order to prepare the data for classification five different features were extracted locally from every EEG channel. The samples of a time window of 500 milliseconds following each image trigger event were isolated for subsequent analysis. This could also be conducted using unsupervised learning for sequential data like for example described in (Genolini and Falissard, 2010). The frame length of 500 milliseconds was chosen in order to fully capture the typical ERP (as depicted in Figure 3.11). The features for the analysis of the ERP were computed in both, the frequency and the time domain (Picton et al., 2000). To obtain a first feature, the first two principal components upon this sequence of samples were calculated using a PCA.

Four more features were generated by applying the fast Fourier transformation (FFT) on the aforementioned windows. The real and imaginary part of the re-

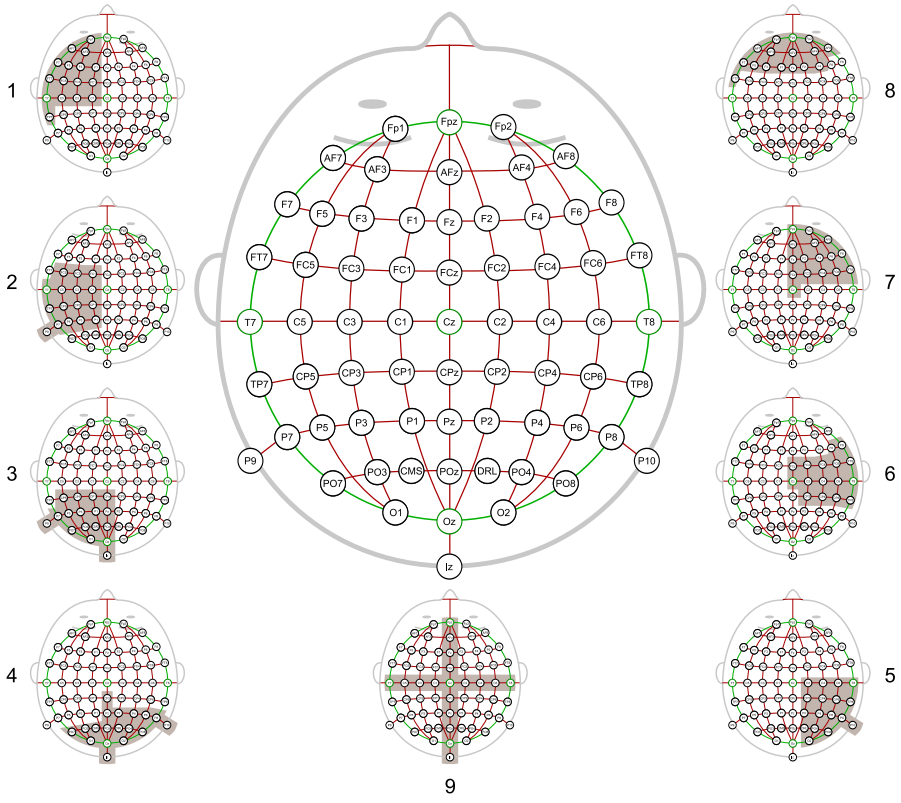


Figure 3.13: *Positions of the 64 electrodes on the scalp of the subject (Image adapted from* `http://www.biosemi.com`*). The small images (1–9) surrounding the main layout illustrate the defined partitions of the 64 EEG channels (gray regions), that will be the inputs to the classifiers of the multiple classifier system.*

sulting frequency spectrum were utilized separately to form a second and third feature. The amplitude and the phase shift of the particular frequencies were computed to form a fourth and fifth characteristic features. All these values were separately passed to a further principal component analysis, projecting the data on the two components, having the highest variance.

Figure 3.12 displays the basic steps of this feature extraction procedure. Features from both time and frequency domain are extracted from the subsequent 500 ms after a stimulus, leading to five different representations for every sensor signal. For these values the first to principal components are computed to render a compressed representation. In order to yield meaningful features for the classifier the resulting feature vectors are pooled for the individual types by defining nine overlapping partitions as shown in Figure 3.13. Thus, we gain robustness by combining multiple feature channels as well as by combining the outputs of several independent classifiers (compare Section 5.4).

A comprehensive overview of the competition can be found in (Hild et al., 2010). Many different classifier approaches have been evaluated in this context with the support vector machine being the most frequent one. Also, the concept of bagging has been widely used. Thus performances of up to 0.82 area under the ROC curve (AUC) have been reached. For a description of the top-scoring approaches please refer to (Leiva and Martens, 2010; Iscan, 2010; Labbé et al., 2010).

4 Methodological Advancements

In this chapter, the methodological developments of this work are described. Section 4.1 comprises approaches for the multi-modal and temporal fusion in the context of human-computer interaction. An approach for the partially supervised learning for uncertain teacher signals is presented in Section 4.2. Further, an adaptation to imbalanced classes of the fuzzy-input fuzzy-output SVM is described in Section 4.3. Finally, the developments are summarized and discussed in Section 4.4.

4.1 Multi-modal Decision Fusion

Many pattern recognition applications are solved in a "one sample" classification approach like for example the recognition of hand written digits or on a basis of small hand designed snippets (Kanade et al., 2000). The application of continuous emotion recognition in human-computer interaction, that is targeted in this work, requires a classification architecture, that reflects also the temporal structure of the labels. A further property of the application is that it is inherently multi-modal since the human emotions are conveyed in a manifold of different channels, such as facial expressions, verbal utterances and gestures. This can improve the over-all classification performance, but it also requires a proper combination approach.

4.1.1 Related Work

Wöllmer et al. (2013) have approached the problem in using the so called Long Short-Term Memory network introduced by Hochreiter and Schmidhuber (1997), which is a recurrent network that is controlled by "gates". These gates control the input and the output of the recurrent memory units. Further a "forget gate" is used to clear these units. The multi-modal fusion is conducted

there by either using the different channels as inputs to a single network or by creating a network per modality and then conduct a late fusion by averaging the outcomes.

Scherer et al. (2012b) study multi-modal fusion in a multi party dialog application using both hidden Markov models (Rabiner, 1989) and echo state networks (Jaeger and Haas, 2004). An echo state network is a recurrent network, that is basically trained by randomly defining a network structure. The training data is then "driven" through this network. Using the outputs of this process, a pseudo inverse is trained to create a mapping to the given labels. For both classification approaches, the multi-modal fusion is conducted after the separate classification of the different modalities. This is done by either summing the log likelihoods of the individual models or by summing up the output of the recurrent network together with a temporal smoothing procedure.

A further multi-modal fusion approach, incorporating the temporal properties of a problem is the Markov fusion network (Glodek et al., 2014, 2012a). It is based on Markov chains and it is defined by different potentials that set constraints for the output such as the outputs of the network shall not differ too strongly over time. The multi-modal combination is conduced based on a previous classification of the individual channels with a confidence value.

4.1.2 Multi-modal Fusion Architectures in Audio Visual Applications

In this section, the two main fusion schemes are proposed to investigate the multi-modal fusion together with temporal integration of intermediate results. The particular challenges that arise from the application of human-computer interaction lie in the inherent properties of the respective modalities.

The video channel provides a new sample at a relatively high rate compared, for example to the rate of the utterances of the subjects or the physiological features. For each of these frames a new feature vector is computed, for which a classification result is derived. However, it is relatively likely that this sensor fails to return a meaningful result for example because the face tracker is not able to work properly. Reasons for this can be occlusions or the subject is turning away from the camera.

The audio channel provides in principle its data in an even higher sample rate. However, in order to compute a meaningful and compact feature for the representation of the spoken language, a time windowing approach is used as described in Section 3.2.3. However, it has been shown that, in order to conduct a successful speech classification, the time granularity should at least be comparable to utterance or word length. A further property of the audio channel

is that feature vectors are not provided regularly: If the subject is not speaking, the channel cannot be used.

For all approaches in this section, it is assumed that the individual modalities are already classified appropriately. This classification is assumed to return a class decision together with a confidence for this classification. As we will see in Section 5.2, only weak classification performances are achieved using only one of the different modalities. Hence the combination approaches show promising improvements for the accuracy in these kinds of applications.

4.1.2.1 Multi-Modal Fusion *after* Temporal Integration

The first approach is mainly to conduct the temporal integration before the combination of the modalities Dietrich et al. (2001) as it is denoted in Algorithm 4.1. The algorithm is provided with the precomputed decision for the $n = 1, \ldots, N$ different continuous modalities, for example represented as a posteriori probabilities $P_t^n = (p_t^n(1), \ldots, p_t^n(C))$ together with a predefined time window length W as input.

Algorithm 4.1: Conducting multi-modal fusion *after* temporal integration

Data:

- P_t^n for $n = 1, \ldots, N$ modalities and $t = 1, \ldots, T$ frames,
- W window size

foreach $w_t = (P^n_{t-W/2}, \ldots, P^n_t, \ldots, P^n_{t+W/2})_{n=1,\ldots,N}$ **do**
  foreach $n = 1, \ldots, N$ **do**
    temporal integration, for example $\pi_t^n = \frac{1}{W+1} \sum_{s=t-W/2}^{t+W/2} P_s^n$;
  multi-modal fusion: for example $\Pi_t = \frac{1}{N} \sum_{n=1}^{N} \pi_t^n$;

Result: Π_i Combined Decision for each time window

Based on this, every sample is embedded into the surrounding samples by defining the time window as $w_i^n = (P^n_{i-W/2}, \ldots, P^n_i, \ldots, P^n_{i+W/2})$. These intermediate classification results are then combined appropriately, for example by applying the average fusion rule. Applying that to all available modalities leads to N new temporally integrated decisions. This procedure is iteratively conducted for every data sample that is incoming for classification. The new decisions for the different modalities generally improve over the decisions, that they are based on because the labels in the underlying application are not changing quickly. The size of the time window is determined by the annotation of the data: slow changes in the label allow longer time windows whereas quickly changing labels demand for smaller windows.

The temporally integrated decisions for the individual channels are finally further combined to form the over-all result, for example using the average fusion rule. This asserts an equal weight for every channel in the final combination. It is noteworthy that this may be the case even though there are potentially differently amounts of decisions in some results than others. The general structure of this information-fusion procedure is depicted in Figure 4.1(a) as a block diagram.

4.1.2.2 Multi-Modal Fusion *before* Temporal Integration

As an alternative, the multi-modal fusion can be conducted before the temporal integration Dietrich et al. (2001) as it is formally denoted in Algorithm 4.2. Analogously to Section 4.1.2.1 the inputs to the algorithm are the intermediate decisions on a fine granularity with the confidences P_i^n for the $n = 1, \ldots, N$ different continuous modalities.

Algorithm 4.2: Conducting temporal integration *after* the combination of the modalities

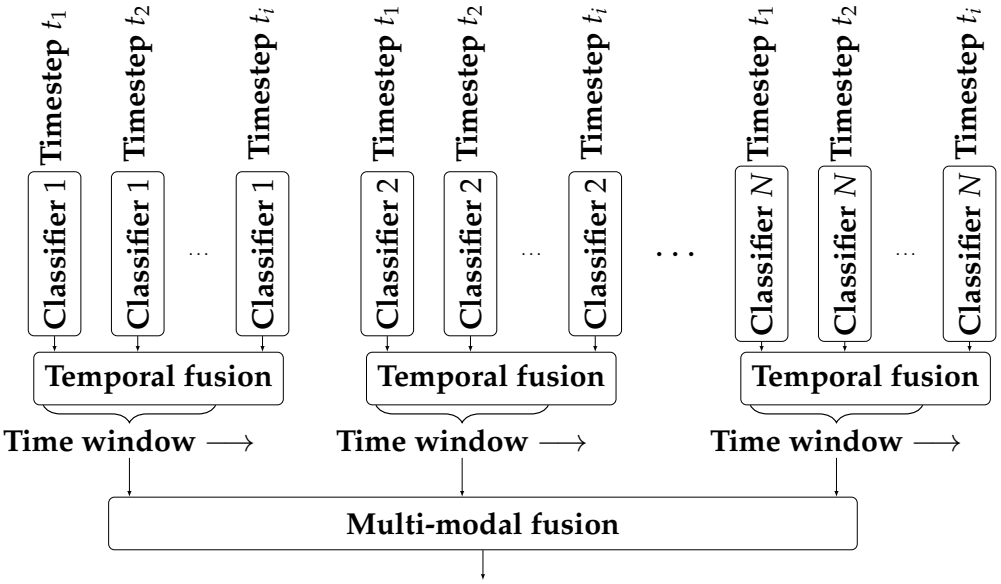
Data:

- P_t^n for $n = 1, \ldots, N$ modalities and $t = 1, \ldots, T$ frames,
- W window size

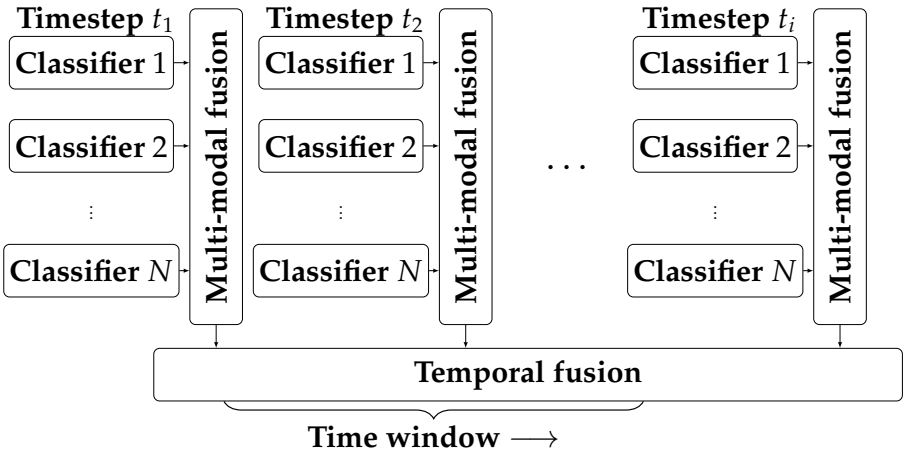
foreach $w_t = (P^n_{t-W/2}, \ldots, P^n_t, \ldots, P^n_{t+W/2})_{n=1,\ldots,N}$ **do**
  foreach $j = t - W/2, \ldots, t + W/2$ **do**
    multi-modal fusion: for example $\pi^j = \frac{1}{N} \sum_{n=1}^{N} P_j^n$;
    `/* If a modality fails, it is omitted from the fusion */`
  temporal integration: for example $\Pi^t = \frac{1}{W+1} \sum_{s=t-W/2}^{t+W/2} \pi^s$;

Result: Π_i combined decision for each time window

Again, the time window of size W is defined for the t-th decision in the n-th channel as $w_t^n = (P^n_{t-W/2}, \ldots, P^n_t, \ldots, P^n_{t+W/2})$. In this approach, the decisions for every frame in the individual modality's time window are combined with the decisions of the other modalities for the respective time step. This is for example conducted by applying the average combination rule. There may be cases where one or more or even all modalities do not provide a classification result for the reasons sketched above. Then the combined decision is drawn from using the available results only. If no decision is accessible at all, the respective time slot is omitted from the computations. This fusion step lays more weight on the modality that returns more decisions in the time window as non-available decisions are treated as mentioned before.



(a) Block-diagram of the architecture "multi-modal fusion *after* temporal integration": The N classifiers are evaluated separately for each time-step in the respective window and a temporal integration is conducted. Subsequently, the multi-modal fusion of the integrated results is conducted and the final classification result for a time window is obtained.



(b) Block-diagram of the architecture "multi-modal Fusion *before* temporal integration": For each time-step t_i each of the N classifiers is evaluated and the decisions are combined in a multi-modal fusion step. Afterwards, a temporal integration step of the combined results is conducted using a time window approach and the final result is hence obtained.

Figure 4.1: *Multi-modal classification architectures.*

After the multi-modal fusion, a temporal integration step is conducted based on these intermediate results. Thus, an over-all classification decision is rendered for a time step. It is possible, that for a particular time window no decision can be made if all channels fail to render a classification, for example if the subject turns away from the camera and it does not speak. Then the decision of the previous window is assigned to the current window, exploiting the assumption, that the labels do not change quickly. The information-fusion architecture is graphically outlined in Figure 4.1(b) as a block diagram.

4.2 Using Unsupervised Learning to Improve Supervised Classification

The annotation of corpora in the affective computing domain is particularly expensive, since the true emotional state of a subject is not entirely observable from the outside. This makes it very compelling for researchers to incorporate additional unlabeled data into the training of a statistical classifier. Recent research has been conducted using the above described techniques in the field mainly for the recognition of human emotions from speech. For example self training is studied in (Deng and Schuller, 2012) and (Esparza et al., 2012) on acted and non-acted data sets. Active learning is successfully conducted in (Zhang and Schuller, 2012) and also in (Esparza et al., 2012). A co-training approach is implemented for several corpora in (Zhang et al., 2013), confirming that adding more unlabeled data to the training does not necessarily improve classification and it can also be degraded under certain conditions.

This section introduces a learning algorithm that incorporates unlabeled data into the classification of the experimental sequences of the EmoRec corpus. With an application like the EmoRec corpus at hand, where a whole session of data with a relatively long duration is recorded but only comparably small sub sequences are explicitly annotated in the respective classes, it is appealing to use all available data, labeled and unlabeled, in the training of a statistical classifier. This is particularly desirable if the labeled data is only rarely available. This is also the case for the EmoRec corpus as the individual experimental sequences are relatively small and there are several feature approaches for the physiological domain, that require a time window of several seconds as described in Section 3.1.3.

A further distinct property of this kind of application is that the classifiers, that are naïvely constructed on single feature vectors and only the labeled data are very weak in their performance (Walter et al., 2011; Schels et al., 2012a). Hence one approach for the improvement of the classification performance, beside the information fusion techniques described in Section 4.1, is the incorporation

of unlabeled data into the training process.

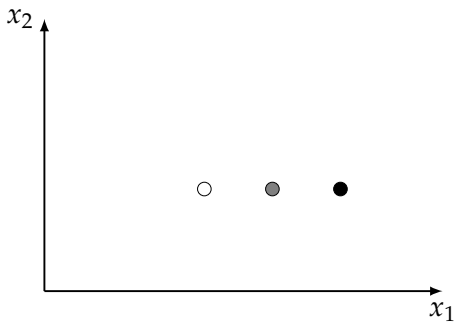
4.2.1 Proposed Partially Supervised Learning Algorithm

In this section an approach to use techniques of unsupervised learning in combination with unlabeled data to improve the supervised learning procedure is proposed. The key idea is to neglect the actual class labels of the samples and to process all available data using an unsupervised learning step (compare Algorithm 4.3). The actual classification problem is solved by a further learning step. Based on the previously attained partitioning of the data, a distance measure of the cluster centers and the data samples is evaluated. This distance could either be computed by a distance measure with respect to a cluster center if a partitioning algorithm is used, or the posterior probability of a mixture component of a fitted generative model. This results in a new representation of the data of the same dimensionality as number of cluster centers. Based on this new feature vector, a classification on the initial label is conducted using standard supervised machine learning approaches. In principle, different clustering approaches or density estimators can be used together with any compatible distance.

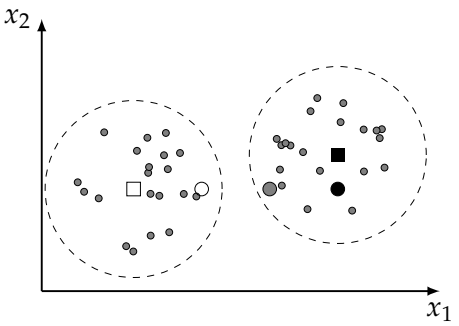
To classify an unseen data sample, it has to be transformed into the new representation. This is done analogously to the training procedure by calculating the distances to the centers: These values are computed with respect to the computed local density or the respective prototype and the obtained new representation is classified.

Another important property of the proposed approach is that the dimensionality of the features can be increased by choosing the number of cluster centers k bigger than the original dimensionality. This makes it easier for a machine learning algorithm to find a linear decision border according to Cover's theorem (Cover, 1965).

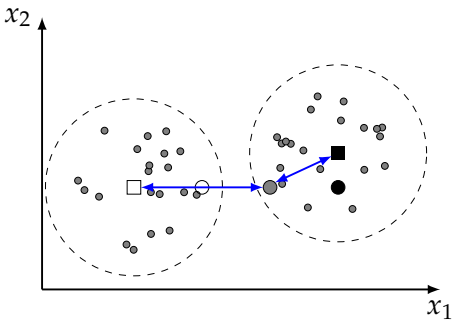
An example for the benefits of such a classification approach is given in Figure 4.3, where the partially supervised approach is evaluated for a varying number of prototypes and compared to standard pattern recognition techniques. Here the COIL 100 database was used, which comprises 7200 images of 100 different objects form different viewing angles. In order to simulate a partially labeled setting, for 92 % of the data the actual class label was removed. This means, that approximately 5 samples per class remained labeled. The features used are color histograms and orientation histograms. The ten fold cross-validation experiments show that in this configuration the pseudo inverse classifier, the linear and the RBF kernel SVM render 0.51, 0.42 and 0.39 error. The partially supervised approach using the pseudo inverse classifier together with a k-means algorithm and the Euclidean distance outperforms the



(a) A new (gray) sample has to be classified into the black or the white class.



(b) Using probability desity estimation (repsective centers denoted as rectangles) with labeled and unlabeled data, the gray sample is more likely in the "black" sample cloud.



(c) The data is re-encoded with respect the distance to the centers of the probability density function. Based on the new representation, a supervised classifier is constructed using the available labeled examples.

Figure 4.2: *The basic idea of the partially supervised approach summarized in an illustrative example.*

Algorithm 4.3: Proposed algorithm in pseudo code.

Data:

- Labeled data $\mathcal{L} = (l)_{i=1...N_L}$
- Respective labels $\mathcal{Y} = (y)_{i=1...N_L}$
- Unlabeled data $\mathcal{U} = (u)_{j=1...N_U}$
- Number of cluster centers k,

Compute k local densities or prototypes $p_1, \ldots, p_k$ using $\mathcal{L} \cup \mathcal{U}$;
foreach $l \in \mathcal{L}$ **do**

$l' = G_{p_1,\ldots,p_k}(l)$;
G is a distance or similarity measure
Examples:

1. $G_{p_1,\ldots,p_k}(l) = (\parallel p_i - l \parallel)_{i=1}^{k}$
2. $G_{p_1,\ldots,p_k}(l) = \left(\exp(-\frac{\|p_i - l\|_2}{\sigma_i})\right)_{i=1}^{k}$
3. $G_{p_1,\ldots,p_k}(l) = \left(\min(\parallel p_i - l \parallel_2, \parallel p_j - l \parallel_2)\right)_{i<j=1}^{k}$

Train classifier F on $((l')_{i=1...N_L}, \mathcal{Y})$;
Result: Classifier F;

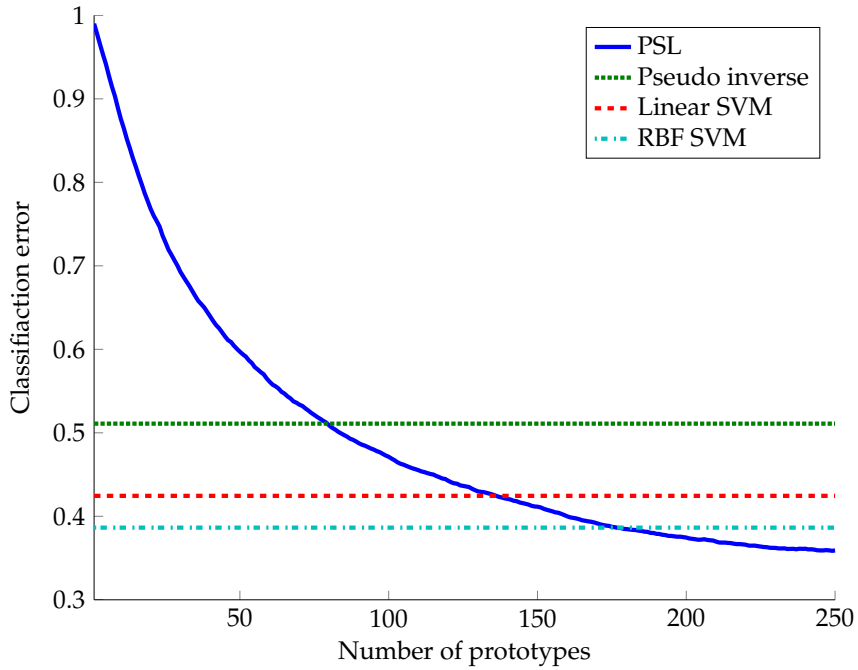


Figure 4.3: *Error rates for the Coil dataset regarding only 8 % of the available as labeled. The errors for the partially supervised approach is plotted against the number of cluster centers. The pseudo inverse classifier is used together with k-means and the Euclidean distance. Also, error rates for pseudo inverse solely and linear and RBF kernel SVM are given.*

purely supervised approach at more than 180 components on. At 250 components, the error decreases to 0.36. Hence, the partially supervised approach is able to capture the distribution of the data more precisely than it is possible without the unlabeled data. This holds especially, when the classification problem is hard, for example when the number of training samples is small.

4.2.2 Related Work

An important source for the incorporation of unlabeled data is of course the extensive literature of semi-supervised learning as for example described in Section 2.5. However, most of these techniques such as self and co-training heavily rely on a initially high accuracy as the classified samples are added to the training set of the models. Adding noise to this set degrades the performance of the model potentially dramatically. Active learning is in this context particularly expensive as the categories may be only weakly defined (Walter et al., 2011; Krell et al., 2013; Siegert et al., 2012)

The approach is related to the initialization of radial basis function networks in supervised classification (Kestler et al., 1995; Ros et al., 2007). Hereby, the

aim is to stabilize the results of the training procedure and also to speed up the process. A typical approach is to pre-train the hidden RBF-layer in an unsupervised fashion by clustering or vector quantization (Schwenker et al., 1994, 2001). Afterwards, the network is finally adapted to the labels by either solely creating a linear mapping for the output layer or back propagation for the whole network. The unsupervised step in our approach can be regarded as some sort of initialization of a "hidden layer" using all available data. Thus, the distributions of data can be estimated more reliably. After that, a second output layer is created with only the labeled data at hand.

A further analogous approach is the bag-of-visual-words approach that is to some extent popular in computer vision (Qiu, 2002). It is inspired by the common bag-of-words approach of the natural language processing. The words in this context are learned using unsupervised clustering or vector quantization, forming a vocabulary or codebook. Based on this codebook the images are represented as bags of the occurring prototypes, which are passed to a supervised classification process. A major difference to the approach that is proposed here is that the data sets for the training of the codebooks and the training of the classifier are not the same. A further possibility to use codebooks to pre-process features for the training of a classifier is the discretization of the feature space using cluster analysis as it is for example conducted in (Vogt and André, 2009). This can have positive effects for training sets incorporating only few samples per class.

The classification of human physiological signals for the detection of emotional categories has been carried out previously in different scenarios. If the classification is conducted subject dependent for one recording session on a single feature vector basis, the classification performances are generally high (Kim and André, 2008). However this setting makes it likely to run into artifacts of the sensors characteristic over time, i.e., samples that are near to each other are likely to have the same label and also a similar sensor characteristic. This does not necessarily relate to the intended phenomena.

Others use different tasks, that are separated but still in one session to train and test a classifier (Hrabal et al., 2012). This mitigates the temporal artifacts but still avoids the variances that may occur when detaching and re-attaching the senor devices to the subject's skin.

Picard et al. (2001a) consider the issue of classification of physiological measurements for one subject but for different days. This is done by introducing a so called "day matrix" to the feature vectors, which is encoding the time in which the data is recorded. It is basically a matrix that has a nonzero value only on entries, representing the respective day. Thus, an increase of the dimensionality of the data is conducted, that makes it easier to linearly classify the data as described in (Picard et al., 2001a).

4.3 Highly Imbalanced Class Distributions

In data sets, that are collected in many real world scenarios, the distribution of classes is often not balanced for example because it is more difficult to gather samples of a distinct class than of other classes. In this section, an extension of the fuzzy input fuzzy output (F^2) SVM is proposed, where the weights that are used to implement the fuzzy membership values in the slack term of the SVM are modified. The main idea for the adaptation to imbalanced classes is to punish misclassification of the over-represented categories are more severely than those of the under-represented classes.

4.3.1 Extending the F^2-SVM to Imbalanced Class Distributions

In this section, the modifications to the SVM concerning the proposed loss term for imbalanced classification tasks are outlined. The class weights for every data sample are given in two N-dimensional vectors $\mathbf{m}^+$ and $\mathbf{m}^-$ which contain the relative proportions of the two opposing classes in the training data and hence $\mathbf{m}^+ + \mathbf{m}^- = (1, \ldots, 1)$. The goal is then to maximize a soft-margin, stated as

$$\underset{\mathbf{w},b,\boldsymbol{\xi}^+,\boldsymbol{\xi}^-}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||_2 + C \sum_{n=1}^{N} (\xi_n^+ m_n^+ + \xi_n^- m_n^-) \tag{4.1}$$

where $n = 1, \ldots, N$ denotes the index of the training samples and $\boldsymbol{\xi}^+$ and $\boldsymbol{\xi}^-$ are linear slack variables penalizing a sample being misclassified. The parameter C controls the penalty of incorrect class assignments. The minimization problem of Equation 4.1 is optimized subject to the constraints

$$\begin{aligned}
\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b &\geq 1 - \xi_n^+ && (4.2)\\
\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b &\geq -(1 - \xi_n^-) && (4.3)\\
\xi_n^+ &\geq 0 && (4.4)\\
\xi_n^- &\geq 0 && (4.5)
\end{aligned}$$

where $\mathbf{w}$ and b describe the orientation and the bias of the hyperplane, and $\phi(.)$ denotes a transformation of $x_n \in \mathbb{R}^n$ into a potentially higher dimensional Hilbert space H. The corresponding Lagrangian of this optimization problem

defined by (4.1) – (4.5) is given by:

$$\begin{aligned} L(\mathbf{w}, b, \xi^+, \xi^-) &= \frac{1}{2}||\mathbf{w}||_2 + C\sum_{n=1}^{N}(\xi_n^+ m_n^+ + \xi_n^- m_n^-) \\ &\quad - \sum_{n=1}^{N}\alpha_n^+((\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b) - 1 + \xi_n^+) \\ &\quad + \sum_{n=1}^{N}\alpha_n^-((\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b) + 1 - \xi_n^-) \\ &\quad - \sum_{n=1}^{N}\beta_n^+\xi_n^+ \\ &\quad - \sum_{n=1}^{N}\beta_n^-\xi_n^-. \end{aligned}$$

Differentiating the Lagrangian with respect to the Lagrangian multipliers and $\mathbf{w}$, b, $\boldsymbol{\xi}^+$, $\boldsymbol{\xi}^-$ and subsequently eliminating the parameters of the hyperplane and the slack variables results in the dual Lagrangian:

$$\widetilde{L}(\alpha^+, \alpha^-) = \sum_{n=1}^{N}\alpha_n^+ + \sum_{n=1}^{N}\alpha_n^- - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n^+ - \alpha_m^-)(\alpha_m^+ - \alpha_n^-)k(\mathbf{x}_n, \mathbf{x}_m), \quad (4.6)$$

where the kernel function is defined by $k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^{\mathrm{T}}\phi(\mathbf{x}_m)$ and the maximization problem is now constrained to:

$$\begin{aligned} \sum_{n=1}^{N}(\alpha_n^+ &- \alpha_n^-) = 0, \text{ with} \\ 0 \leq \alpha_n^+ &\leq Cm_n^+, \text{ and} \\ 0 \leq \alpha_n^- &\leq Cm_n^-. \end{aligned} \quad (4.7)$$

In order to satisfy the Karush-Kuhn-Tucker conditions, properties

$$\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-, \boldsymbol{\beta}^+, \boldsymbol{\beta}^- \geq 0$$

,

$$\begin{aligned} \alpha_n^+((\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b) - (1-\xi_n^+)) &= 0, && (4.8) \\ \alpha_n^-((\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b) + (1-\xi_n^-)) &= 0, && (4.9) \\ \beta_n^+\xi_n^+ = (Cm_n^+ - \alpha_n^+)\xi_n^+ &= 0, && (4.10) \\ \beta_n^-\xi_n^- = (Cm_n^- - \alpha_n^-)\xi_n^- &= 0, && (4.11) \\ \forall n = 1, \ldots, N, & && (4.12) \end{aligned}$$

and properties (4.2) – (4.5) hold. A numerical solution can be computed using the *sequential minimal optimization* (SMO) approach introduced by Platt (1999a).

Once the Lagrangian multipliers $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ have been found, the parameters $\mathbf{w}$ and b of the hyperplane are determined by:

$$\begin{aligned} \mathbf{w} &= \sum_{n=1}^{N} (\alpha_n^+ - \alpha_n^-)\phi(\mathbf{x}_n), \text{ and} \\ b &= \frac{1}{2N_{\mathcal{M}^+}} \sum_{n \in \mathcal{M}^+} \left(1 - \sum_{l \in \mathcal{S}^+} (\alpha_n^+ - \alpha_n^-) k(\mathbf{x}_n, \mathbf{x}_m)\right) \\ &+ \frac{1}{2N_{\mathcal{M}^-}} \sum_{n \in \mathcal{M}^-} \left((-1) - \sum_{l \in \mathcal{S}^-} (\alpha_n^+ - \alpha_n^-) k(\mathbf{x}_n, \mathbf{x}_m)\right), \end{aligned} \tag{4.13}$$

where $\mathcal{S}^+$ ($\mathcal{S}^-$) is the set of support vectors $\alpha_n^+ > 0$ ($\alpha_n^- > 0$) and $\mathcal{M}^+$ ($\mathcal{M}^-$) is the set of unbounded support vectors with $\alpha_n^+ < Cm_n^+$ ($\alpha_n^- < Cm_n^-$). According to equations (4.10) and (4.11) ξ_n^+ (ξ_n^-) $= 0$ if the sample n is in the set $\mathcal{M}^+$ ($\mathcal{M}^-$). The bias parameter b is averaged by using Karush-Kuhn-Tucker conditions (4.8) and (4.9) to obtain a numerically stable solution.

A class decision can then be obtained by $y = \text{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b)$. To extend the SVM to a probabilistic output, the distance $d(\mathbf{x})$ of the input $\mathbf{x}$ to the hyperplane is mapped to $\widetilde{y}(\mathbf{x}) \in (0, 1)$ using a sigmoid function (Platt, 1999b) with parameter a

$$\widetilde{y}(\mathbf{x}) = (1 + \exp(-a \cdot d(\mathbf{x})))^{-1}$$

can be minimized according to the mean square error on the training data. Since $m_n^+ = 1 - m_n^-$ only m_n^+ needs to be considered and the following equation is optimized:

$$E = \frac{1}{N} \sum_{n}^{N} (\widetilde{y}(\mathbf{x}_n) - m_n^+)^2,$$

which can be accomplished by a linear regression technique.

4.3.2 Related Work

Many real world applications in pattern recognition need to deal with imbalanced training sets. Common techniques to mitigate this issue are oversampling of the underrepresented classes or under-sampling of the overrepresented classes (Japkowicz, 2000; Zhou and Liu, 2006). Another approach to

highly imbalanced data sets is utilizing a particular loss function in the chosen classifier design such as individual cost terms for each class. Such a loss function penalizes misclassification of underrepresented classes more severely than others. In the following, such a loss term is incorporated into the common formulation of the support vector machine (Schölkopf and Smola, 2001) as for example proposed in (Osuna et al., 1997). This approach has proven to be feasible under different circumstances as *fuzzy-input fuzzy-output support vector machine* technique as described by Thiel et al. (2007).

4.4 Discussion

Recent developments in computer science allow distinct applications to push the frontier forward to more subtle categories in human-computer interaction. One important factor for this is that different sensors to capture a subject and its surroundings are becoming cheaper and more accessible for everyday purposes. These sensors do not only comprise cameras and microphones but also physiological measurements from the subject's body. Further, storing measurements is also becoming cheaper, making it easier to construct learning algorithms for more complex categories like user states. This goes along with the development in computer memory capacities and computation speeds, that allows to develop and implement statistical models and classifiers on these data.

This availability of data results on the other hand in data collections that comprise not clearly defined categorizations and segmentations. This results in very hard classification tasks that provide noisy data distributions that cannot easily be separated into classes that are shipped with the data. These classes are furthermore in parts not as profoundly defined as in other technical applications, which is further discussed in Chapter 3. This originates for example from categories, that are globally defined over a distinct period of time as it is the case with the EmoRec data collection and one cannot be sure that this assumption is met in the whole segment. Another example is the AVEC 2011 data collection, which provides four different emotional categories that are not too easily accessible in practice. This holds particularly true for this application as the emotional dimensions only occur in comparably low degrees. A further issue in this context is that the different classes appear in many real world applications unevenly distributed. This originates from the fact, that the underlying experiments are commonly not arbitrarily repeatable (using dispraise for the elicitation of emotions may serve as an example for this) or the experimental setting does not allow even class distributions, for example the oddball paradigm mentioned in Section 3.3.

These challenges have to be addressed appropriately in the classification pro-

cedures. The relatively high noise level in the distributions of the data that is described earlier demands for base classifiers that have comparably few degrees of freedom. One major technique that allows to still render complexer class boundaries is the multi-modal combination of independent channels. Usually in multiple classifier system only different feature views for the respective samples are available, which makes is arguable if the independence assumptions that are often made in the classifier fusion are met in reality. By using completely different sensor devices, this assumption can be asserted stronger and the combination is generally more promising. A further distinct feature of these applications is that the data is naturally available in time series for which the different categories do not change too quickly over time. This makes the temporal integration of intermediate results a very promising approach. Both approaches have to incorporate the fact that the sample rate for the data and hence the rate of feature vectors can vary considerably for different sensor types. Finally for the problem of imbalanced class distributions, the incorporation of weights for the respective training data points is a promising approach. The weights are designed to punish misclassifications of the lesser represented class more severely than others.

The approaches, that have been introduced in this chapter will be numerically evaluated in Chapter 5. The multi-modal and temporal information fusion approaches will be tested using the EmoRec and AVEC 2011 corpora of human-computer interaction (see Section 5.2) . This comprises the whole AVEC data, that includes audio and video channels and also the audio-visual part of the EmoRec data collection in order to allow for a fair comparison. The physiological part of the EmoRec will be used for the evaluation of the partially supervised learning approach, that is described in Section 4.2 (compare Section 5.3). This part of EmoRec is well suited for the approach as the individual features are computed on a relatively long timely basis, which makes the extension of the underlying data for the statistics promising. Additional experiments for this approach on benchmark data sets are presented in the Appendix. The proposed SVM for the imbalanced class problems will be evaluated in Section 5.4. This evaluation will be conducted on the PASCAL 2 mind reading competition data collection.

5 Numerical Evaluation

This chapter resembles the numerical evaluation of the methods, that are described in Chapter 4 in the context of human-computer interaction. Before the presentation of the actual numbers, the used performance measures and statistical testing protocols are described in Section 5.1. Afterwards in Section 5.2 the multi-modal and temporal information fusion architectures are evaluated by the means of the audio-visual parts of the EmoRec and the AVEC 2011 corpora. In Section 5.3 unlabeled data is incorporated in the classification of the physiological part of the EmoRec corpus as described earlier. Finally, the SVM incorporating the class weighting mechanism is evaluated on the PASCAL 2 mind reading competition corpus. This evaluation is described in Section 5.4.

5.1 Classifier Performance Assessment

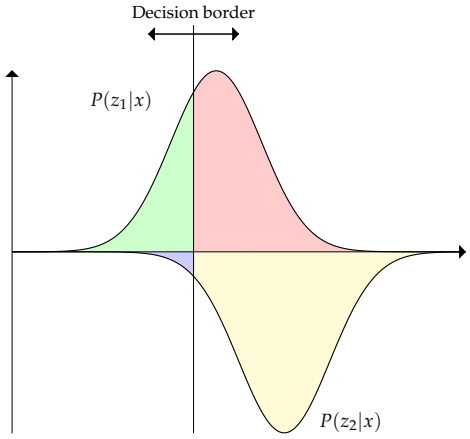
In this section the issue of numerically evaluating a statistical classifiers is discussed. It will focus on the relevant measures and the usage of data for that purpose will be discussed.

5.1.1 Error Rate and Receiver Operating Characteristics

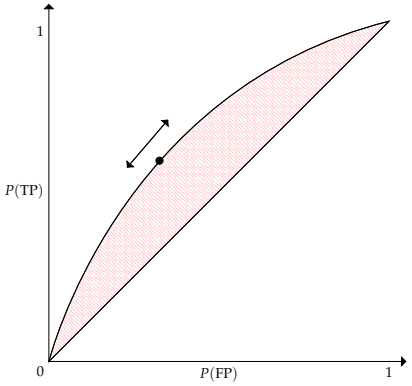
The most common and intuitive performance measure for a classification task is the error rate. It is briefly the relative amount of misclassified samples in a data set. In order to formally introduce this measure, the notation is borrowed from Webb (2002): Let $Y = \{y_i, i = 1, \ldots, N\}$ be the training data for a classifier, where each sample is defined as $y_i = (x_i, z_i)$. These samples comprise a feature vector x_i, $i = 1, \ldots, N$ and the corresponding class labels z_i, $i = 1, \ldots, N$ in a vector notation. Further is the function $\omega(z)$ the categorical class label of z and the output for a sample x_i of the classifier that is trained using Y is denoted as $\eta(x_i, Y)$. The loss function $Q(\omega(z), \eta(x, Y))$ is defined as

	True label	
Predicted label	True positive	False positive
	False negative	True negative
	1	1

(a) Confusion matrix with definition of true positive (TP) and false positive (FP) classifications.



(b) Shifting the decision border yields different TP and FP values.



(c) The plot of TP against FP for the possible decision borders returns the ROC curve.

Figure 5.1: *The logic of the ROC curve (Theodoridis and Koutroumbas, 2009).*

follows:

$$Q(\omega(z), \eta(x, Y)) = \begin{cases} 1, & \text{if } \omega(z) \neq \eta(x, Y) \\ 0, & \text{otherwise.} \end{cases}$$

Based on this the apparent error rate is defined as the average loss for the N data samples (Webb, 2002):

$$e_A = \frac{1}{N} \sum_{i=1}^{N} Q(\omega(z_i), \eta(x_i, Y)).$$

Error rate is obviously underestimating the error of the classifier, particularly for applications using models with many degrees of freedom together with only few data samples. It converges however to the true error rate if infinitely many data samples are drawn from the distribution of the data.

A further widely applied performance measure for two class problems is derived from the so called receiver operating characteristic (ROC) (Theodoridis and Koutroumbas, 2009). It is defined using the confusion matrix for the classification as it is displayed in Figure 5.1(a). There, the values for the true positive (TP) classifications and the false positive (FP) classification are taken and

test	train	train	train	fold 1
train	test	train	train	fold 2
train	train	test	train	fold 3
train	train	train	test	fold 4

Figure 5.2: *Example for 4-fold cross validation: The data are split in folds of equal size and in every run a different fold is left out of the training (white) of the classifier and used subsequently for its testing (green) (Bishop, 2006).*

varied by incorporating a classification threshold. This threshold is shown in Figure 5.1(b) for a two class problem that is displayed for Gaussian distributions. It represents the decision boundary, which is set to classify every sample left of the border as a member of the negative class and right of the border as a member of the positive class. By moving the threshold as indicated in the figure from left to right different configurations of the classifier are sampled and thus different pairs of TP and FP values are obtained. The ROC curve is thus created by plotting the relative TP values against the FP values, which typically renders a curve like the one displayed in Figure 5.1(c). The diagonal in of the table this figure indicates the border between classification results: Points above this line indicate classifiers, that are better than random guessing and points under the line indicate weak classifiers equal to or worse than random guessing. In order to access the quality of the classifier the area under the ROC curve (AuC) is a well-known measure. The larger this area, the better is the discriminative power of the underlying classifier (Bradley, 1997; Landgrebe and Duin, 2008). Especially, when the data collection comprises a highly imbalanced class label distribution the area under the ROC curve is considered an appropriate performance measure (compare e.g., Chawla, 2005; Fawcett and Flach, 2005).

5.1.2 Cross-Validation

The issue of the correct estimation of the true error rate of a classifier is commonly approached using so called holdout estimates (Webb, 2002). The most common technique in this context is k-fold cross-validation, where the available data is split into k folds of equal magnitude as exemplified in Figure 5.2 for a 4-fold cross-validation setting. In k different training and classification runs, the k-th fold is left out of the training of the classifier for which only the $k-1$ remaining folds are used. The data that is held out of the training is then used for the evaluation. This procedure is repeated for each of the k folds. The resulting error rate is then defined as the average error of the folds. This is formally defined using the previously outlined notation as follows (Webb,

2002):

$$e_{\text{CV}} = \frac{1}{N} \sum_{j=1}^{N} Q(\omega(z_j), \eta(x_j, Y_j)).$$

In this equation Y_j denotes the training set where the sample x_j is held out.

A common issue in this context is the optimization of additional classifier combination schemes or meta parameters of the classifiers for example the variance of a RBF kernel function. For these optimizations an additional validation set is generally necessary in order to avoid over-fitting on the training set but more importantly not to use the test set for such optimization.

In the context of human-computer interaction and related topics, a further application specific partitioning of the available data is the subject-*independent* evaluations and the subject-*dependent* partitions (Huang and Lee, 1993; Schuller et al., 2005). In subject-dependent settings samples of a subject are present in both, the training and the test sets, whereas for the subject independent experimental settings the samples of a subject are only either in the test or in the training set. The subject independent classification poses obviously a greater challenge for a statistical classifier.

This if further intensified in the context of physiological signals, when the study that led to the underlying corpus has a strict temporal protocol. This is the case for example when different task have to be executed in a fixed row (compare e.g., Valstar et al., 2013) or the stimuli that are presented in defined sections of the experiment (compare e.g., Walter et al., 2013). The temporal characteristics of many physiological sensors, that originate for example from a drying of the gel of the sensor pads or the pads slip slightly out-of-place over time, leads generalization issues when conducting a leave one sample out strategy within a single session for the evaluation of a statistical classifier. Hence it is mandatory to use at least different sessions, when dealing with such a scenario or follow a subject independent evaluation strategy.

5.2 Audio-Visual Classification Experiments

In this section, the evaluation of the multi-modal and temporal fusion architectures are conducted on the EmoRec and the AVEC 2011 corpora by the means of classification experiments. The implementation of the precise approaches for the classification of the features extracted from the audio and video channels is described in Section 5.2.2 for the audio channel and in Section 5.2.1 for the video channel. Based on these individual classifications of the modalities, the multi-modal information fusion architectures that were introduced in Section 4.1.2 are systematically evaluated in Section 5.2.3 with respect to differ-

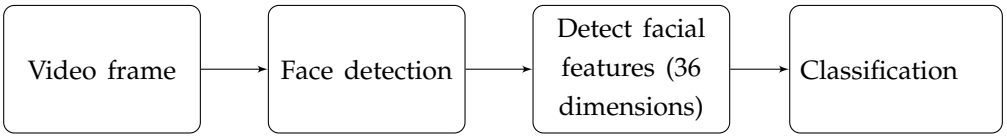


Figure 5.3: *Steps for the classification of facial expressions.*

ent time granularities. The respective findings are finally discussed in Section 5.2.4.

5.2.1 Classification of Facial Expressions

As mentioned before, the feature calculation for the video channel is conducted using the CERT tool box (compare Section 3.2.4). The subjects in the employed data are more or less looking directly into the camera and hence the tool box yields relatively stable values. Normally, one feature vector per frame is returned. But there are also cases, when no feature is returned. The most frequent reason for this is the failure of the face tracker. An obvious cause for this is that the subject is facing away from the camera. Further, glasses and excessive beards and also the facial electrodes of the EMG recording devices can sometimes disturb the video feature extraction. In these cases the respective frame location remains unset.

The output of the modules "Basic Emotions 4.4.3" (i.e., Ekman's basic emotions), "FACS 4.4" and "Unilaterals" and "Smile Detector" are considered (compare Section 3.2.4). Thus, a 36-dimensional vector for every frame is obtained.

The classification for this channel is conducted using linear least squares classifiers. Multiple of these classifiers were constructed using a bagging approach with 100 individual classifiers. Thus, complexer decision borders can be constructed despite of the simple base classifier. The class decision is computed using voting of the base classifiers. The classification uncertainty is captured by the standard deviation of the respective outputs of the individual classifiers. This base model with only few degrees of freedom turns to be very effective, as the distribution of classes for this channel is normally very noisy as seen for example in Figure 3.9. The obvious reason for this is that the expressions are very subtle in spontaneous non-acted emotional data. The general procedure of the classification of the video is depicted in Figure 5.3.

Still, the video channels renders a relatively weak classification performance in this application. As elaborated in Section 3.2.4, the models used by the CERT software are trained using acted emotional data sets. Because the occurrence of affective states in the considered data collections are much more subtle, the resulting feature vectors render a noisy class distribution, which makes discrim-

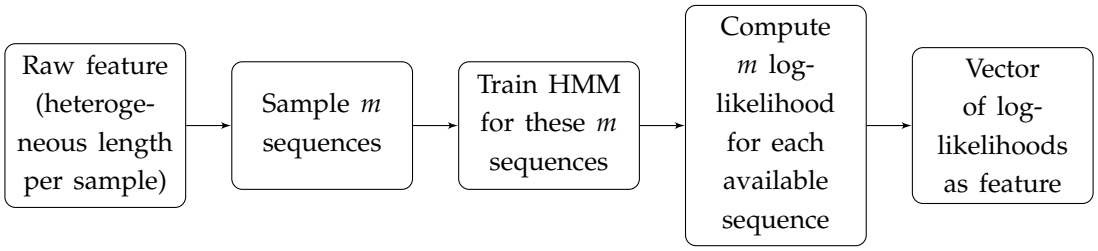


Figure 5.4: *Using HMM to transform sequences into feature vectors of uniform length.*

inative analysis difficult. Preliminary experiments showed that more complex classifiers struggle to construct meaningful class borders in this context. For example the optimization of SVM tends to assign all training samples to one class, regardless of the true class value. The false classification are thus treated as margin violations (compare Section 2.2.3). The proposed configuration of the base classifier using linear least squares pseudo inverse together with the ensemble learning scheme is hence designed to overcome this issue and still construct a meaningful classifier.

5.2.2 Classification of Spoken Utterances

A challenge for the classification of spoken utterances is that verbalized words have naturally different lengths. One way to circumvent this to conduct classification on a very low frame based level. For example in (Schels et al., 2012a) the raw MFCC frames on 250 ms time slices were used, amongst others, directly for the classification of utterances. But this straight forward approach does not seem to provide sufficient information to construct meaningful classifiers. In this work an utterance based representation using HMM and multiple features is applied. Thus the sequences of audio features is transformed into a single feature vector of uniform length, that can be processes by standard machine learning techniques.

A distance-matrix for a set of sequences $M = \{S_1, \ldots, S_n\}$ is constructed using the following algorithm (Smyth, 1997):

1. Train one HMM λ_i per sequence $S_i \in M$
2. Calculate the log likelihoods $L_{i,j} = \log P(S_i|\lambda_j)$ for every sequence with respect to every HMM. In order to mitigate the effects that are caused by the duration of a sequence, the log likelihood is normed using its length.
3. Finally, the distance between two sequences is $d(j,i) = \frac{1}{2}(\overline{L}_{i,j} + \overline{L}_{j,i})$

The result of this algorithm is a $n \times n$ symmetric distance matrix. Bicego et al. (2003) have proposed a generalization of this method by reducing the number

of HMM to "reference" sequences. This means that only $m < n$ sequences are randomly sampled from the available data and only for these HMM are constructed. Instead of computing the mean of the two log-likelihoods $L_{i,j}$ and $L_{j,i}$, the log-likelihood of the reference models is used normalized by the length of the sequence: $x_{ij} = \frac{1}{|S_i|} \log P(S_i|\lambda_j)$, where S_i is the i-th sequence and λ_j the model of the j-th samples sequence. The concatenation of these values $X = x_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$ creates a new feature vector for every sequence that is of uniform length. This is a computational alleviation especially for big databases. The algorithm is depicted as a block diagram in Figure 5.4.

An alternative for this approach is for example dynamic time warping (Wendemuth, 2004), that is widely used for audio processing for example in speech recognition (Rabiner and B.-H., 1993) and bio-acoustics (Dietrich, 2003). The HMM based method was however successfully used in affective computing using speech for example in (Schels et al., 2013b; Glodek et al., 2012b,a).

The five different audio features, that are described in Section 3.2.3 — namely MFCC, RASTA-PLP and LPC together with energy and f0 are processed separately in three "groups" of feature in the HMM preprocessing. For each group, $m = 75$ sequences were randomly selected. This value is used in order to keep it computationally feasible. The resulting 3 feature vectors per utterance are then concatenated and form the final feature.

Based on this, a random forest classifier with 50 individual trees is constructed. The number of single classifiers was not systematically optimized, but rather a sensible number in the trade-off between computational expense and degrees of freedom. The uncertainty measure, that is used for classification is the relative number of trees, that vote for a class.

The choice of a complexer classification model compared to the approach chosen for video material is made as the different audio features together with the HMM-based utterance time granularity potentially constitute a more informative modality. This means that for the video modality 25 new feature vectors are potentially available every second, whereas the audio modality provides a new feature vector in a much slower rate approximately based on a few seconds. A further characteristic of the Random Forest technique is that it provides an intuitive uncertainty measure via the agreement in the ensemble of trees. Hence this approach renders a rather robust classifiers provided enough trees as individual classification trees are defined (compare Section 2.2.4).

5.2.3 Evaluation of Multi-modal and Temporal Fusion Architectures

In the following, the numerical evaluation of the audio-visual classification architectures on the publicly available AVEC 2011 corpus and also on the EmoRec II corpus is presented. The evaluations comprise the classification of the audio and video channels solely and the two proposed information fusion architectures.

5.2.3.1 "AVEC 2011" Data Collection

All experiments are conducted in a strict *subject independent* 4-fold cross validation, i.e., the subjects that are in the training set do *not* occur in the respective test set. The different combinations of classifier decisions are evaluated for this corpus on all four available labels.

Arousal

Figure 5.5 displays the results of the audio-visual classification for the category "arousal". It shows the classification error for the four approaches with respect to the size of the employed time window in seconds.

On the single frame level (far left on the plot), the classification errors are relatively high, for both the single modalities and their combinations. The audio channel (blue line) shows a smaller error of 38 % compared to the video (red line) with 43 %. Both types of combinations are suited in the middle at around 40 % error (cyan and purple).

When applying longer time windows in the integration step, the error decreases until approximately 70 s to 80 s. the audio classifier still yields lower errors of 29 % compared to the video classifier at 36 % error. This is despite the fact that there are intrinsically more decisions for the video frames available. The combination of the modalities shows in both cases errors between the individual modalities. Combining the modalities after the temporal integration with a channel renders better classification results compared to the case, where it is conducted the other way round: 31 % versus 33 % error. A reason for this is the low performance of the video classifier, that inflicts the over-all result. Especially, when the channels are combined on a frame level, the impact of the quicker video modality is comparably high. For time windows, that are longer than 80 s, the classification error increases again.

The subject's arousal is obviously conveyed mostly over the voice of the speaker. This is in accordance to the literature, where simple thresholds for the

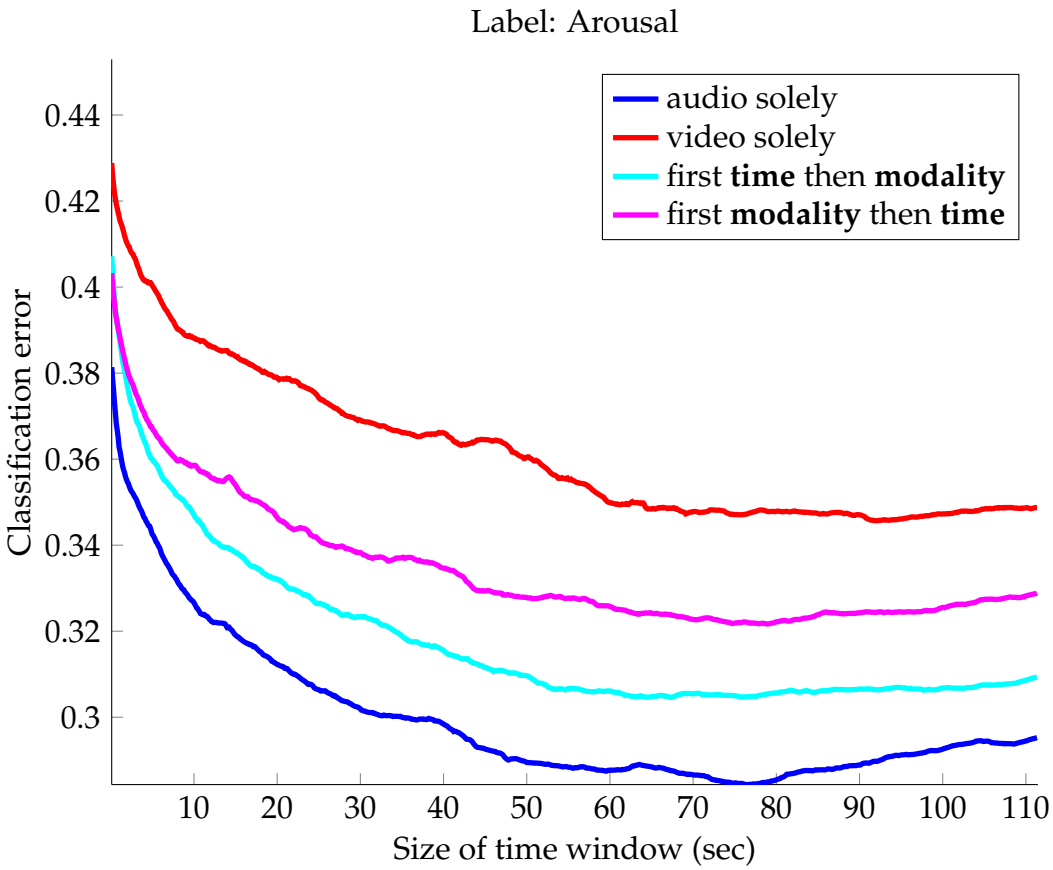


Figure 5.5: *Audio-visual classification error versus the time window size for the category "arousal" on the AVEC 2011 corpus.*

energy of the voice signal are sometimes used for the categorization of the activation of a speaker (compare e.g., Johnstone and Scherer, 1999).

Expectancy

The evaluation for the category "expectancy" is shown in Figure 5.6. Analogous to the previously described category arousal the different classifier fusion approaches are evaluated with respect to the size of the employed time window size.

For this label, the classification on a single frame is even worse than for the label arousal. For the audio channel an error rate of 45 % is computed. The video channel renders only slightly better results with an error of 44 %. The combination of the classifiers yields approximately equal error rates.

When increasing the time window, the error decreases in almost all cases except, when only the audio channel is used. The video channel improves con-

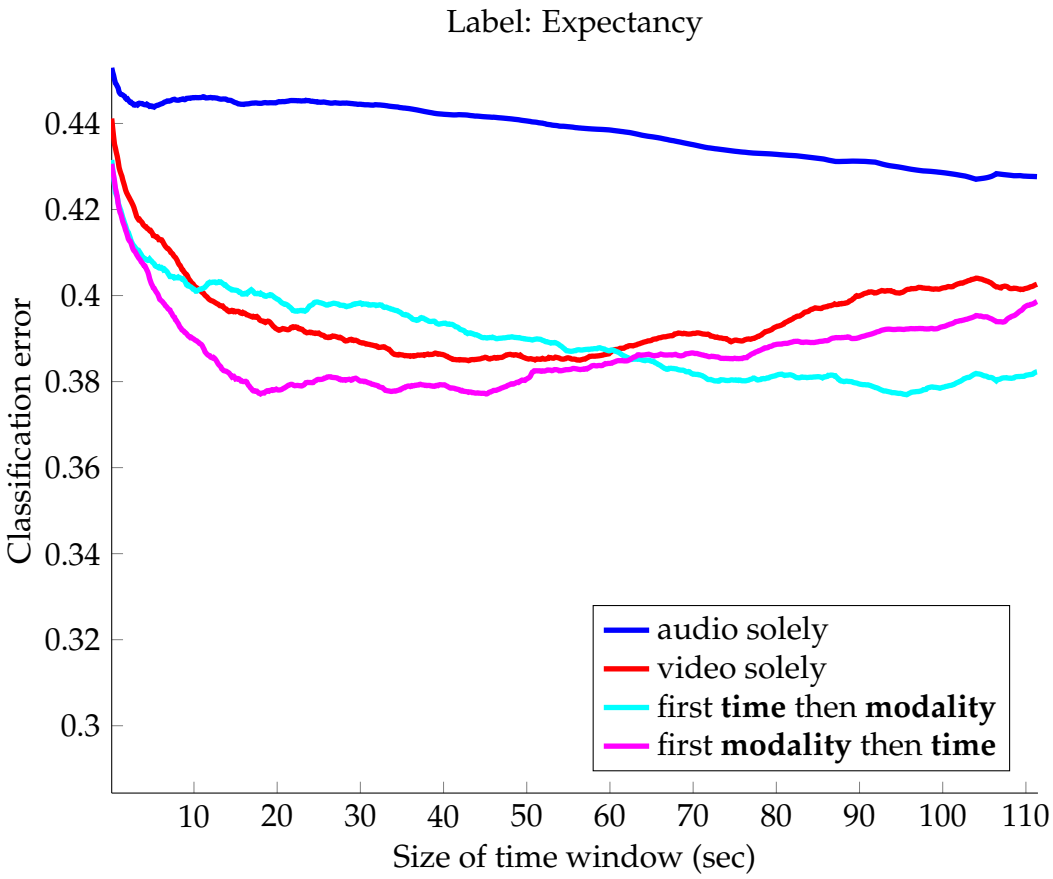


Figure 5.6: *Audio-visual classification error versus the time window size for the category "expectancy" on the AVEC 2011 corpus.*

siderably to about 39 % error at approximately 30 s time window and its performance degrades again for window sizes longer than 60 s (compare red plot in Figure 5.6).

Varying the window size has only marginal effects on the classification performance. Figure 5.6 shows evidences, that larger time windows are beneficial for the audio channely. However, the classification errors are still high for big window sizes.

The combination of the modalities for the approach, where the temporal integration is conducted after the frame-wise combination further improves over the video classifier (compare purple line in Figure 5.6). It also depends only on shorter time windows than for the single modality. The optimum region is already obtained at approximately 20 s with 37 % error. At around 45 s, the error increases again — also slightly earlier than for the video classifier solely.

The classifier fusion approach, where the temporal integration is conducted first (compare Figure 5.6, cyan plot) renders a lower classification performance

than both, the video and the other combination method. For very long time windows, this approach reaches an optimal performance for window sizes of 95 s. Here, 38 % error is rendered.

Conducting the fusion of the channels first lays more weight on the video classifier, that is more accurate for this label. Hence, the combination method outperforms the other one. It also outperforms the video classifier solely, which is here the best individual classifier.

Long time windows for the audio classifier render an improvement for this channel, such that the classifier fusion, where the temporal integration is conducted first (and the weights for the channels are equal) yields also comparably good results. However, for these time windows the underlying data are reduced noticeably and this result might not be conclusive.

Power

The results for the label "power" are similar in a sense to the ones for the label expectancy. The evaluation is summarized in Figure 5.7, where the error rate for the approaches is plotted against the size of the respective time window.

Initially, both, the audio and the video classifier share an error rate of approximately 44 % to 45 %. As before, the video classifier can benefit from longer time windows, but there is no clear optimum observable. The error is reduced quickly first and from 60 s on the improvement slows considerably down. At around 100 s, the error tilts down again, which might be an artifact of the excessive window duration. Analogous as mentioned before, the underlying data set size is also smaller for these window sizes and this might be an artifact of that circumstance.

The audio classifier is not at all benefiting from longer time windows. The error is constantly at approximately 44 % for all window sizes.

The combination of both classifiers yields the same error rates as the individual modalities at a frame basis. But both approaches can benefit from growing time windows. Conducting the combination of the channels before the temporal integration reaches an optimal window size at 30 s to 40 s with an error rate of 35.5 %. For long time windows the curve follows the one for the video channel as this modality tends to overrule the audio as described before.

The combination approach, where the temporal integration is conducted first, the error drops slower compared to the other fusion approach. The error shows no sharp optimal temporal resolution and stay at a saturation value of 36.5 % error from window sizes of approximately 50 s on.

Generally, this category is a fine example for how the combination of temporal classifiers can improve over different modalities and temporal granularity.

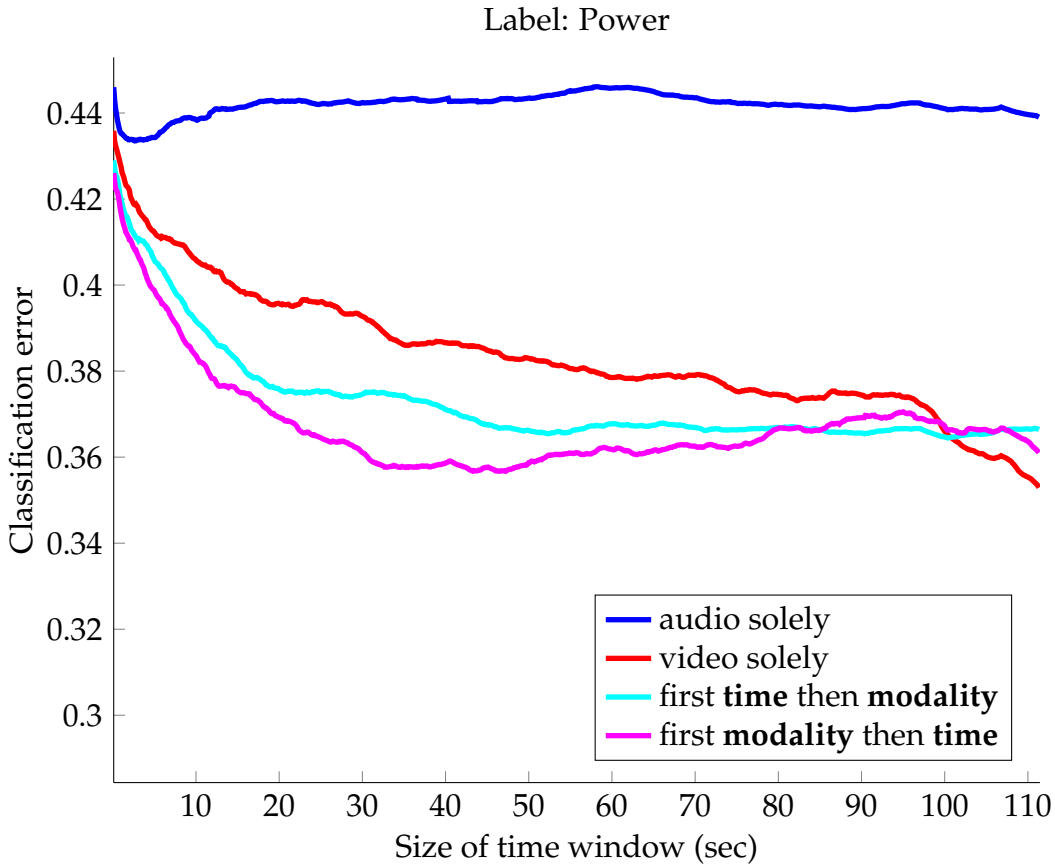


Figure 5.7: *Audio-visual classification error versus the time window size for the category "power" on the AVEC 2011 corpus.*

Even though the audio classifier renders a comparably bad classification accuracy, the combination improves over the best individual classifier for this category. A reason for that is that not only the individual accuracy is crucial for the classifier fusion but also the classifiers' diversity on the samples. Further, the uncertainty measure that is used to access the quality of a respective decision has to output sound values, for example returning wrong decisions with a high certainty.

Valence

The last category, that is provided with the AVEC 2011 corpus is the label "valence". The audio-visual classification results for this label are displayed in Figure 5.8.

For the single frame classification, the audio channel is rendering a comparably high error rate of 44 %. The video classifier and the combination of classifiers

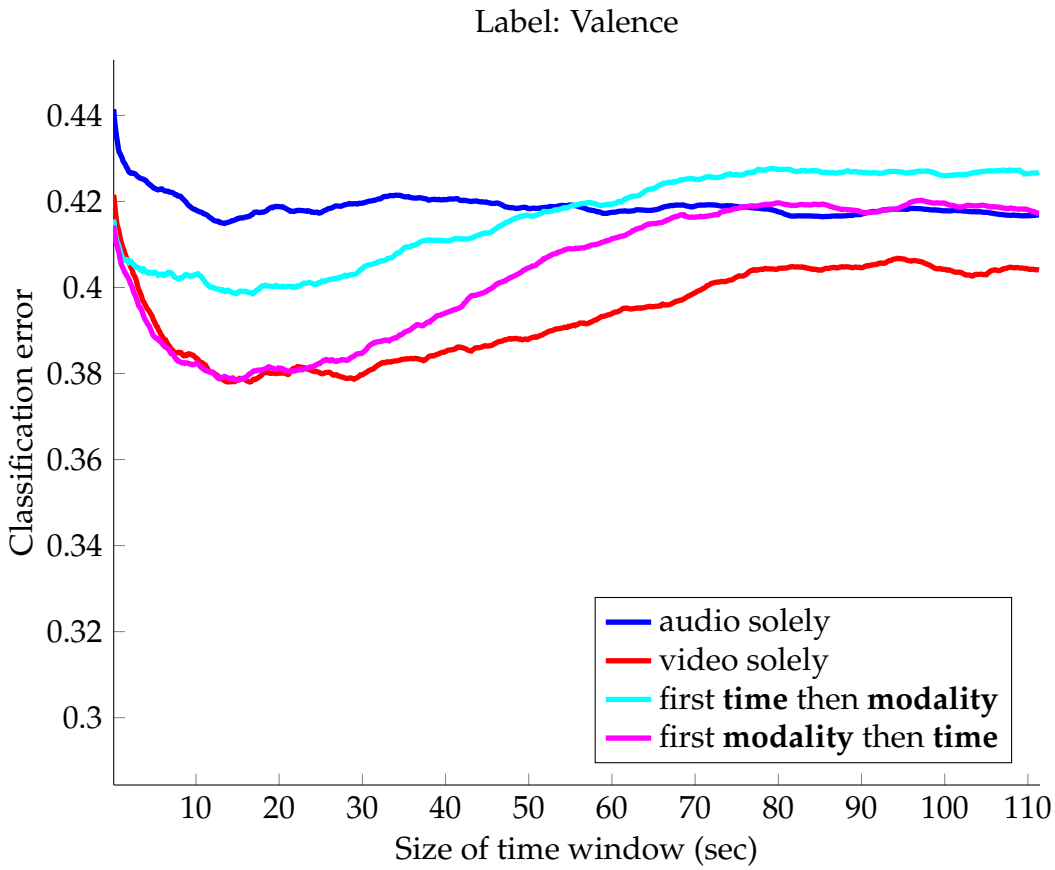


Figure 5.8: *Audio-visual classification error versus the time window size for the category "valence" on the AVEC 2011 corpus.*

yields slightly better results with approximately 42 % error.

All classifiers improve, when the time window is increasing. Again, the video classifier benefits more and renders a minimal classification error of 37 %, while the audio classifier is only slightly improved to 42.5 %. Both unimodal approaches reach the maximum performance at relatively small time windows of approximately 15 s. While the results for audio are more or less stable for further increasing window sizes, the video classifier decreases in performance, when the time granularity is more than 35 s.

The combination of the classifiers cannot improve over the best individual classifier in this case. Conducting the fusion over the channels before the temporal integration shows identical error rates for window lengths of less than 25 s. From 15 s on, a minimal error of 37 % is accomplished. When the window is grown beyond 25 s, the error increases quicker compared to video solely and an error equal to audio is measured.

Conducting the temporal integration first only improves over the audio solely classification until window sizes of less than 55 s. The optimal error for this approach is rendered at approximately 18 s to a total of 40 %. Here the optimal region is relatively sharp and the errors increase quickly again with the length of the time window to even higher rates than the audio solely.

Results on the AVEC 2011 Corpus from the Literature

Classification results for the AVEC 2011 data set are obviously generated in the course of the name giving challenge, and they are summarized by Wöllmer et al. (2013). The form of the respective result is however strongly influenced by the experimental constraints imposed by the challenge organizers. For example, the classification performance for the audio and audio-visual parts of the challenge are only calculated on a word level granularity. Further, partitions of the data that were used for the evaluation of the classifier performance were, unlike it is the case in the experiments above, not subject independent. This makes the task easier for the machine learner and the respective classification rates are higher for reasons stated earlier.

The winner of the AVEC was determined by the accuracy of the classification results that were submitted for the category arousal as the respective numbers were the larger compared to the others. The highest accuracy for this label on the audio data on word level with 60.9 % correctly classified was achieved by Ramirez et al. (2011) using latent-dynamic conditional random fields and by Glodek et al. (2011) using an ensemble approach with hidden Markov models as individual classifiers and an MLP as trainable classifier fusion scheme simultaneously. Further participants of the challenge used for example SVM for the on word level and thus rendered an accuracy of 59.8 % for the category

arousal (Sayedelahl et al., 2011). Pan et al. (2011) used an AdaBoost approach for the classification of the category arousal on a word basis and thus rendered an accuracy of 57.6 %. Using GMM based classifiers 55.3 % accuracy were achieved by Kim et al. (2011) on the same task. Finally, Cen et al. (2011) used a so-called extreme learning approach for the challenge rendering an accuracy of 52.0 %.

The organizers of the challenge also report classification results on the corpus incorporating the previous outcomes of the challenge participants (Wöllmer et al., 2013). They evaluate many variations of the recurrent long short-term memory network (Hochreiter and Schmidhuber, 1997), rendering a large variety of results on the data set. They reported for arousal on audio data on a word level the highest accuracy of 71.2 % on the test set. In the same experiment the accuracy for the other labels were 57.3 % for expectancy, 57.4 % for power and 68 % for valence.

5.2.3.2 EmoRec II

As already pointed out in Section 3.1, the EmoRec II corpus is slightly different compared to the AVEC 2011 corpus. The respective categories are not labeled manually in a session to form alternating segments with the same label, but they are embedded as whole experimental sequences into the over all sessions (compare Figure 3.4). That means for our type of experiment, that the window sizes can potentially grow beyond the border of the explicitly labeled parts. Further the sample size remains unaffected by the size of the window for classification.

ES-2 versus ES-5

The results of the experiments with the EmoRec II corpus are summarized in Figure 5.9. Analogous to the earlier described evaluations on the AVEC corpus, the error of the two unimodal classifiers (red for video and blue for audio solely) and their multi-modal combinations (cyan and purple) are shown for different window sizes for the temporal integration.

The accuracy of the individual classifiers for the EmoRec corpus are quite similar to the AVEC results for the single frame classification. The audio classifier renders slightly better results with an initial error of 40 %, while the video and the combination on this granularity perform worse at about 45 % error.

Increasing the size of the time window for the individual modalities decreases the error until 35 % for the audio channel for window sizes longer than 70 s. The errors remain at this level for larger windows of up to 90 s. Beginning from that length, the error slightly increases again.

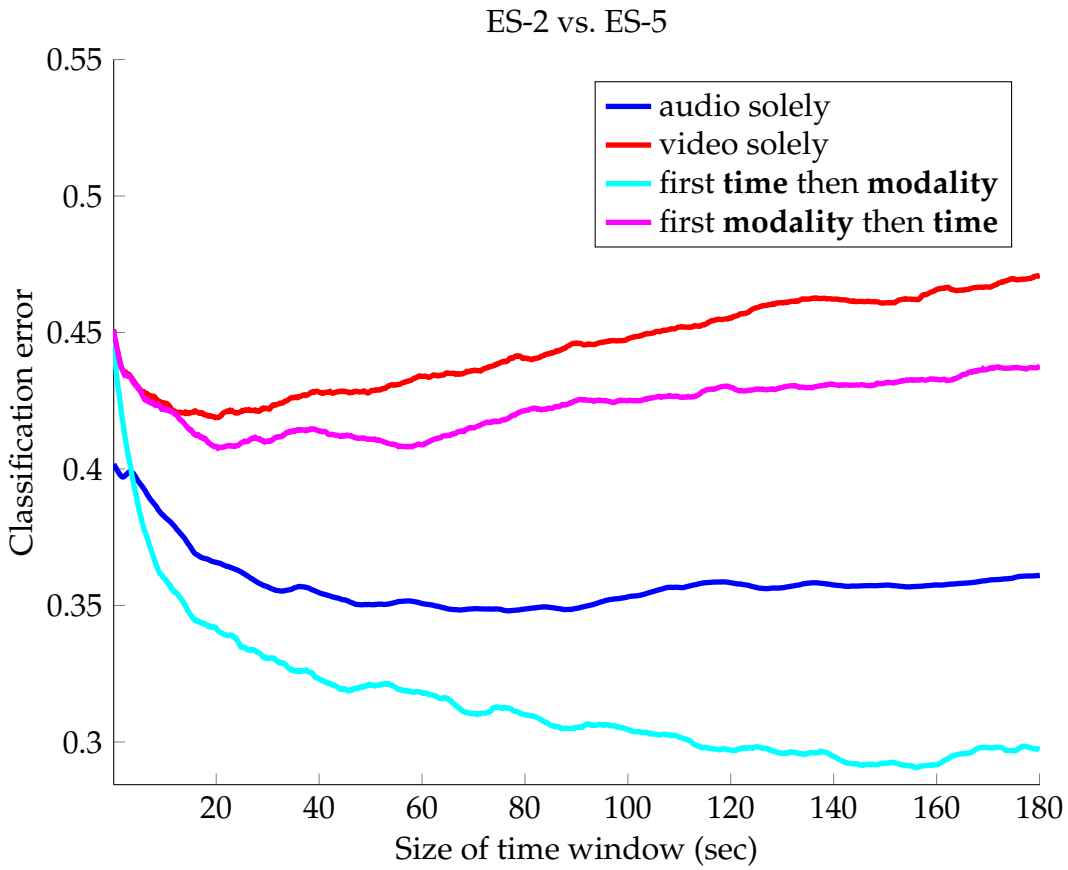


Figure 5.9: *Audio-visual classification error versus the time window size for ES-2 versus ES-5 on the EmoRec II corpus*

For the video channel, the improvement is less than for the audio. For window sizes of approximately 15 s, the error is optimal for this modality at 41 %. Larger windows increase the error — even beyond the initial value for a single frame.

Combining the individual classifiers improves over the video classifier for both approaches. The audio classifier, resembling the best individual classifier in this case, is also outperformed when the temporal integration is conducted before the fusion of the classifiers. For this approach the maximal performance is reached using relatively long window sizes of more than 140 s to 30 % error. Using window sizes of more than 170 s increases the error again slightly. The time windows are significantly longer than in the previous cases. For the underlying corpus for this experiment, the different categories do not alternate as quickly as it is the case for manually labeled data. Thus, the classification based on longer lasting processes is possible and, as the experiments show, beneficial.

Conducting the channel fusion cannot outperform the audio classifier for equal time windows. The minimum error is yielded for time windows of approximately 20 s. Here 41 % recognition error is achieved. The error remains at that level until the window size grows bigger than 60 s. Starting from there, the performance decreases, but the error remains smaller than for the video classifier for the same time window length.

Additional Results Based on EmoRec II

In principle, the experimental setting, that is used in EmoRec II offers more combinations of more or less similar two-class problems (compare Figure 3.4). In principle, these experiments can be conducted additionally for the combinations of experimental sequences ES-4 versus ES-6, as preferred by Hrabal et al. (2012), ES-2 versus ES-6, ES-4 versus ES-5 and the unification of ES-2 and ES-6 versus both, ES-4 and ES-5 (i.e., the "positive" against the "negative" experimental sequences).

The results for these settings are provided as comparison of supplemental evaluations in the appendix of this work in sections B.1 to B.4. The obtained curves are, however, less conclusive than the one for ES-2 versus ES-5, which is discussed in Section 5.2.3.2. One reason for this is that when adding the two other sessions to the data, the embedding of the sessions in the over-all recording is not optimal any more. For example the ES-6 is the last session in a recording, making symmetrical time windows impossible. Further, ES-4 is directly adjacent to ES-5 and closer to ES-2 than it is the case for ES-5.

5.2.4 Discussion

In this section, the spatio-temporal characteristics of the nature of human emotions in affective computer interaction are investigated. In order to do that, a time windowing technique was systematically evaluated together with the combination of classification results for the audio and the video channel in episodes of human-computer interaction.

The major finding of these experiments is that the classification performance is optimal for relatively long time scales. Very prototypical characteristics for the error curves have been obtained for the AVEC 2011 data set. The error is in many cases decreasing starting from high error rates on a frame level to a minimum error for time windows of 15 s to 80 s, depending on the emotional label. This shows, that the underlying human processes are passing on a relatively slow time scale, especially compared to the raw sensor sampling rates, which are commonly around 16 kHz for audio and 25 to 50 frames per second for video. Also compared to emotional phenomena in the psychological literature, these time scales are relatively long, for example for skin conductance, the respective peak in the signal occurs approximately 4 s after the eliciting event.

These findings are further confirmed by analogous experiments conducted on the EmoRec II corpus. The time windows that show minimal error rates appear at coarser time granularities, which is due to the global labeling of the experiment. This results in even longer time windows than in the AVEC 2011 corpus.

The performance of combining the audio and video channel improves in three out of five cases over the best uni-modal classifier. The evaluations show, that the fusion is most successful, when it is not the case, that one classifier is clearly outperforming the other. For example for the label “arousal”, the audio is clearly rendering lower error rates than the video channel as seen in the Figure 5.5. Hence, the fusion is not able to improve over that by combining a relatively weak classifier. On the other hand, when the performances of the classifiers are more similar, the improvement of the combinations is noticeable. Exemplary the reader is referred to the experiments for the labels “expectancy” and “power”. The results for these evaluations are shown in the figures 5.6 and 5.7. Here, the classifiers render more similar classification rates and the fusion improves over the best classifier, which is based on the video data.

The question concerning the optimal information fusion approach in human-computer interaction, i.e., is the temporal integration *before* or *after* the combination over the modalities optimal, depends heavily on the performance of the individual modality and thus on the respective label. When the classifier based on the video feature performs better, the combination method, where frame-

wise results are averaged is optimal. Thus, the over-all result is biased towards the result of the video classifier. When the slower audio classifier yields better error rates, the temporal integration is conducted before the multi-modal combination. This is the case for example for the dimension arousal in the AVEC corpus and the EmoRec II corpus, in which the weight of the features is equal and both channels are potentially improved over time separately.

Considering the channels, it is notable, that for arousal and similar classes the audio classifier is a more accurate classifier. This holds also for the EmoRec corpus as ES-2 corresponds to "low" arousal while for ES-5 the arousal is set to "high" as described in Section 3.1.1. For the rest of the classes in the AVEC corpus the facial expressions are more informative than the nonverbal communication.

5.3 Partially Supervised Evaluations on the EmoRec Corpus

The partially supervised learning algorithm described in Section 4.2 is evaluated on the physiological part of the EmoRec corpus. This corpus is especially well suited for this kind of learning approach as only few labeled data samples for the physiological channels are available. This originates from the long time scales that are required for the extraction of meaningful features from this type of signal compared for example to the audio and video channel. The evaluation of the classification of the individual physiological channels is presented in Section 5.3.1. From these individual classification results a combined multimodal classifier is constructed and its evaluation is described in Section 5.3.2. Subsequently, in Section 5.3.3 the question concerning how much additional data is needed to improve the over-all classification is addressed. Finally, a broader discussion of the findings in the context of the evaluations is presented in Section 5.3.4.

5.3.1 Classification of the Individual Physiological Channels

The features described in Section 3.1.3 are extracted not only from different modalities but also in different time scales. Hence, 6 individual base classifiers were defined, grouping the data by the type of feature and by the size of the time window: For the EMG, features that govern in time domain (derivatives of the signal and related) are grouped together (8 features) as well as features obtained from the power spectrum (3 features). Also for the skin conductance, two groups of features were defined for classification: The statistics over the derivatives are processed in a different classifier (4 features) than the statistics

of the peaks of the signals (9 features). In case of BVP and respiration such a partitioning is not necessary as the time windows of all extracted features are the same (10 and 6 features).

In order to compute the new representation, the three different settings described in Section 4.2 Algorithm 4.3 were evaluated:

- k-means clustering and Euclidean distance:

$$G_{p_1,\ldots,p_k}(l) = (\parallel p_i - l \parallel)_{i=1}^{k}$$

- Gaussian mixture models (GMM) together with EM-algorithm using the posterior per mixture component:

$$G_{p_1,\ldots,p_k}(l) = \left(\exp(-\frac{\parallel p_i - l \parallel_2}{\sigma_i})\right)_{i=1}^{k}$$

- k-means clustering and distance based on pairwise distance measure:

$$G_{p_1,\ldots,p_k}(l) = \left(\min(\parallel p_i - l \parallel_2, \parallel p_j - l \parallel_2)\right)_{i<j=1}^{k}$$

For these approaches, the number of centroids has been varied from 1 to 30.

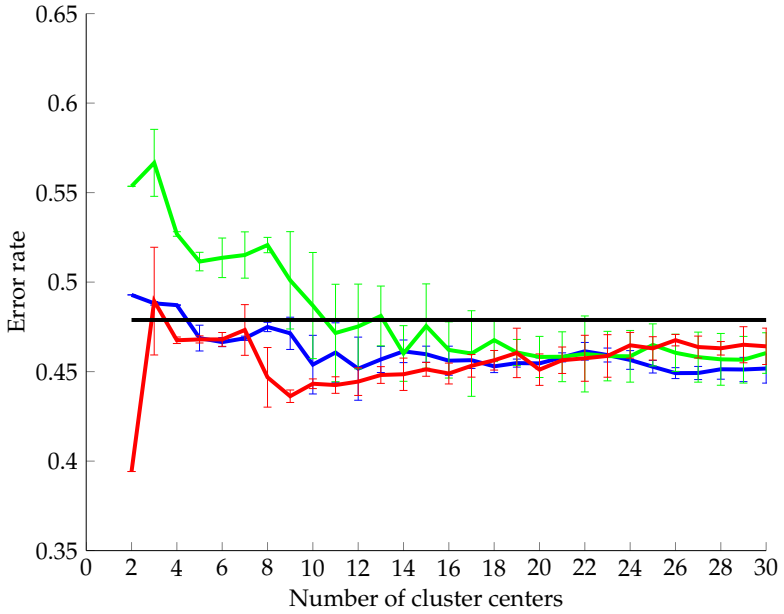
Based on this, the supervised part of the classification has been conducted using the Moore-Penrose pseudo inverse:

$$V^i = Y \lim_{\alpha \to 0+} C_i^T (C_i C_i^T + \alpha I)^{-1}.$$

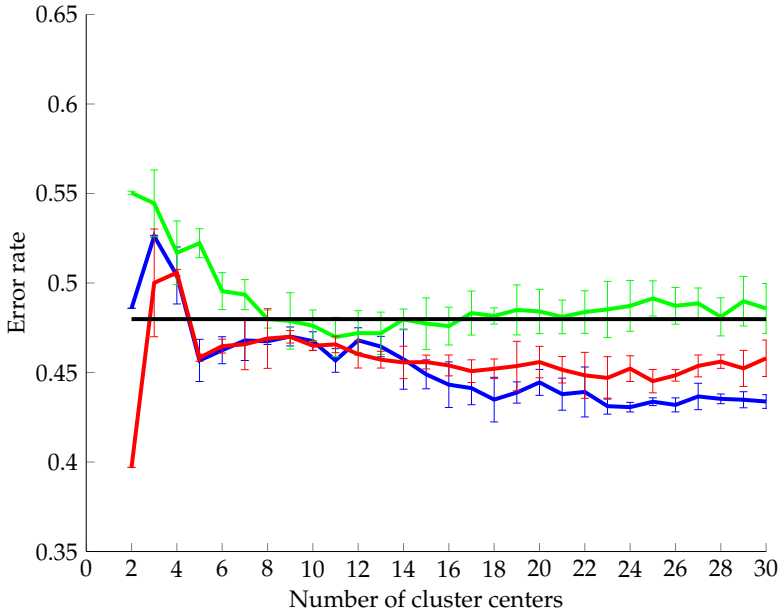
Hereby, C is the co-variance matrix of the training data $\hat{X}$, Y is the matrix of the respective labels and I is the identity matrix. A bagging approach with a bag of 100 individual classifiers is followed (Breiman, 1996). The classification experiments are conducted using a *leave one subject out* strategy. Every experiment was repeated 10 times to capture the variances that are created by the clustering procedure. Using such an elementary base classifier enables us to process noisy data as it is the case in the present application without over-fitting. However, using the bagging approach enables to still capture complexer class borders.

As reference for the experiments, a conventional supervised classification is analogously conducted. Here the bag of 100 pseudo inverses is also used is computed directly using only the available labeled data, i.e., the data of the respective experimental sessions. The temporal integration and the fusion over the channels is conducted as described above.

The performance of the model with respect to the number of centers in the pre-processing step, i.e., the complexity of the resulting network, is outlined



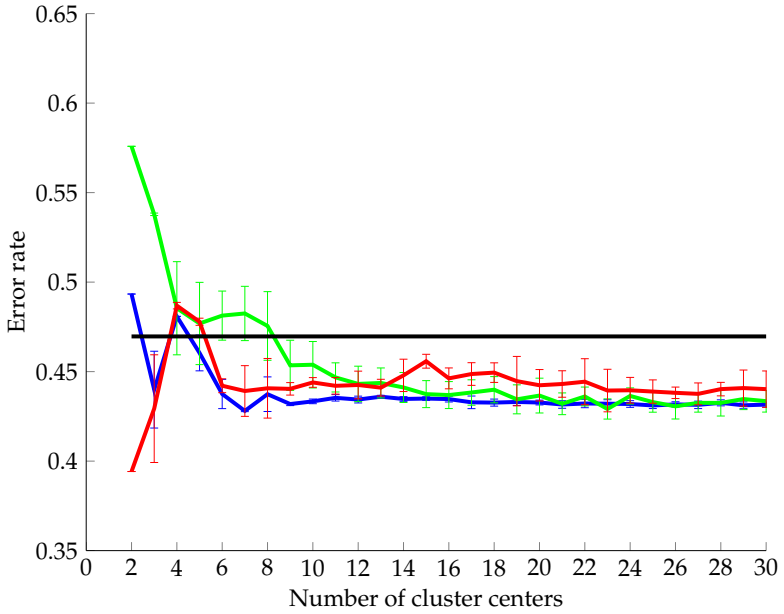
(a) Fast EMG features.



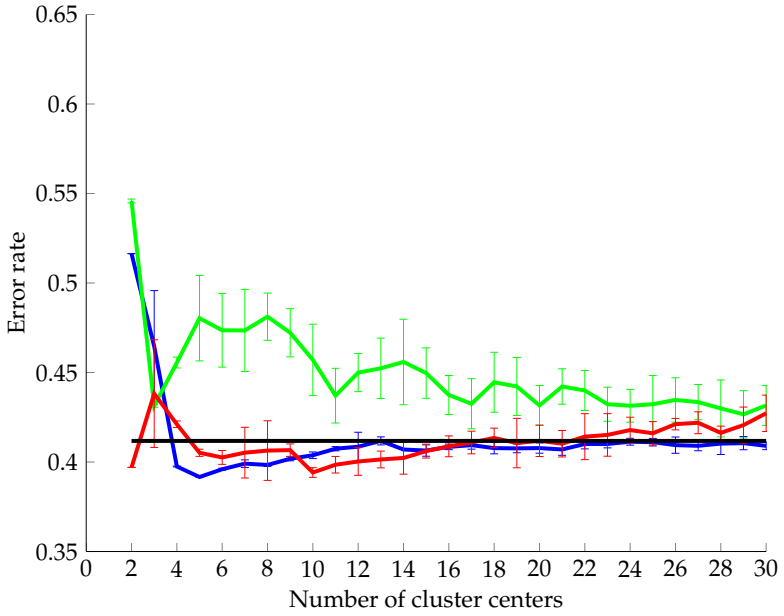
(b) Slow EMG features.

Legend:

- $G_{p_1,\ldots,p_k}(l) = (\| p_i - l \|)_{i=1}^{k}$
- $G_{p_1,\ldots,p_k}(l) = \left(\exp(-\frac{\|p_i - l\|_2}{\sigma_i})\right)_{i=1}^{k}$
- $G_{p_1,\ldots,p_k}(l) = (\min(\| p_i - l \|_2, \| p_j - l \|_2))_{i<j=1}^{k}$
- purely supervised reference



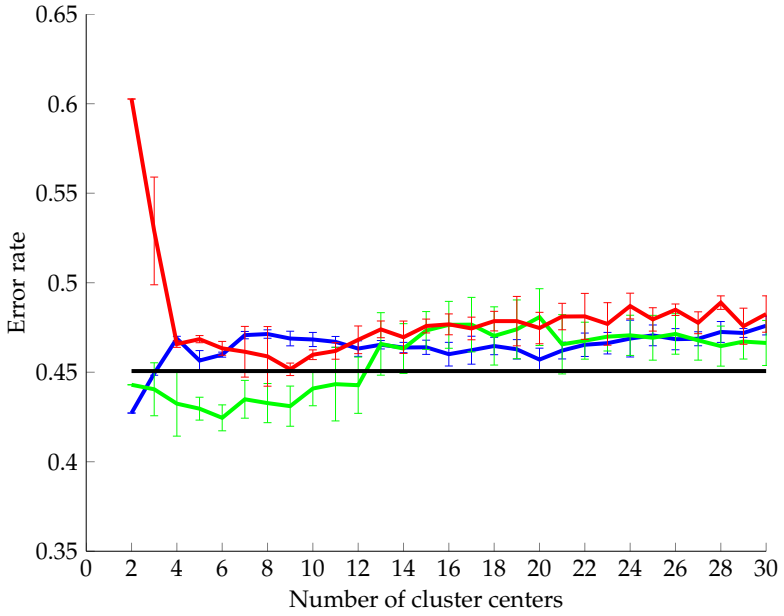
(c) Fast skin conductance features.



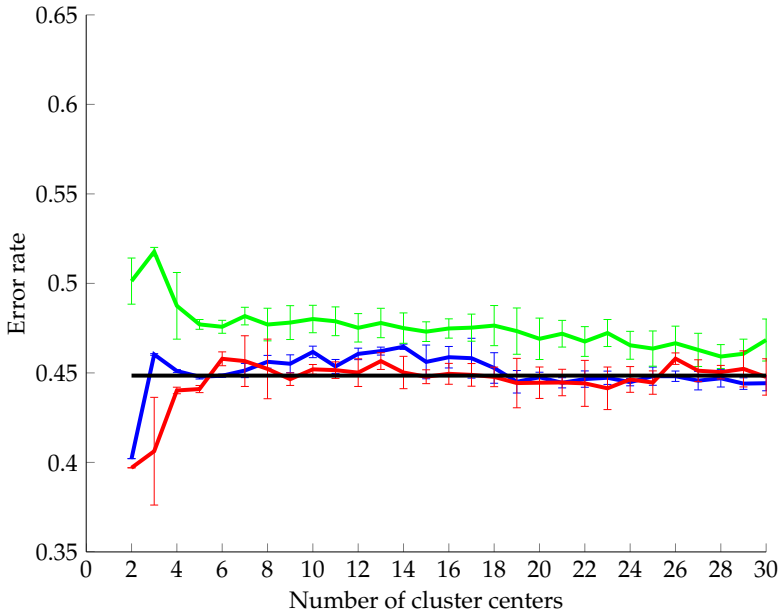
(d) Slow skin conductance features.

Legend:

- $G_{p_1,\ldots,p_k}(l) = (\| p_i - l \|)_{i=1}^{k}$
- $G_{p_1,\ldots,p_k}(l) = \left(\exp(-\frac{\|p_i - l\|_2}{\sigma_i})\right)_{i=1}^{k}$
- $G_{p_1,\ldots,p_k}(l) = (\min(\| p_i - l \|_2, \| p_j - l \|_2))_{i<j=1}^{k}$
- purely supervised reference



(e) Heart rate variability.



(f) Respiration.

Legend:

$G_{p_1,\ldots,p_k}(l) = (\| p_i - l \|)_{i=1}^{k}$

$G_{p_1,\ldots,p_k}(l) = \left(\exp(-\frac{\|p_i - l\|_2}{\sigma_i})\right)_{i=1}^{k}$

$G_{p_1,\ldots,p_k}(l) = (\min(\| p_i - l \|_2, \| p_j - l \|_2))_{i<j=1}^{k}$

purely supervised reference

Figure 5.10: *Error rates together with the standard deviation of 10 runs for the individual classifiers for different number of centers. The fully supervised reference for the respective feature is given as a black line.*

in figures 5.10 and 5.11. There, the confusion rates of the classifiers are shown together with standard deviation of the repetitions of the experiment.

For the individual channels, the recognition rates are generally low as displayed in Figure 5.10. However, the proposed method is able to outperform the purely supervised method in 3 of the 6 cases as displayed in the three figures 5.10(a), 5.10(b) and 5.10(c). The different unsupervised techniques yield approximately the same results, except for GMM where the error is often higher as seen in the figures 5.10(b), 5.10(d) and 5.10(f).

In detail, the following observations can be made: considering the purely supervised approach, the features can be subdivided into two groups. The classifiers based on the following three features perform comparably well:

- Using the "slow" SCL features the lowest over-all error of 41.2 % is achieved (compare Figure 5.10(d), black line).
- Respiration with an average error of 44.9 % (compare Figure 5.10(f), black line).
- BVP showing an average error of 45.1 % (compare Figure 5.10(e), black line).

The second group comprises the cases, where the error rates that are computed using the supervised classifier are relatively high:

- "Fast" SCL features with an average error of 47.0 %, compare Figure 5.10(c), black line.
- "Fast" EMG features with an average error of 47.9 %, compare Figure 5.10(a), black line.
- The "slow" EMG having an error rate of 48.0 %, compare Figure 5.10(b), black line.

These results show that there is a tendency for the classifiers, that are based on longer time windows yield higher classification rates. This is especially articulated considering the SCL related classifiers: the features on 20 s basis show the best classification result, while the ones based on 5 s are comparably worse. The poor results for the EMG on both time scales are relatively surprising, as the respective mount points of the sensors are very intuitively chosen for the estimation of emotion. Furthermore, the EMG is an exception for the fact, that longer time windows enhance the classification performance. The channels of BVP and respiration are somewhat in the middle. They appear to be not directly connected to the emotional state, at least in cases of small occurrences as it is the case in the application at hand.

Based on this, the effects of the proposed partially supervised method are described in the following. Here, also two main groups of cases are observed: On the one hand, there are three cases, where the supervised result is outperformed and on the other hand in three cases the result is approximately equal or for one case even a little worse.

An improvement for the classification can be observed for both EMG features. The error can be reduced to approximately 45 % as it can be seen in the figures 5.10(a) and 5.10(b). The necessary number of components varies around 10 to 20 centers. Generally the two approaches, that are based on the k-means approach yield stable results beginning at about 10 components. The one based on GMM is stable at about 20 mixtures for 5 s features but it does not converge for the features on 20 s basis. Here, no improvement can be observed. One reason for this could be that there is not enough data to estimate a proper covariance matrix, which arises from the bigger time windows.

A further case that shows an improvement is the fast SCL, which is denoted in the Figure 5.10(c). An improvement from 44 % to 45 % can be achieved for this feature. The approaches based on k-means show stable results using 6 to 7 centers. Again the GMM depends on a higher number of mixtures: Using 15 components a stable classification result is rendered.

However, there are also cases of classifiers, where no clear effect is observed. For respiration and the slow SCL, the approaches based on k-means yield similar results as the supervised case described above. This circumstance is outlined in the figures 5.10(f) and 5.10(d). Again, the GMM based approach leads to a decrease of the classification performance to error rates of 45 % and 48 %. Further, this approach shows a relatively high variance for the SCL. A reason for this could be that the longer time windows result in fewer feature vectors per time unit. Thus, it might be harder to estimate a correct covariance matrix.

In case of BVP, all variants show a comparably low performance. The error rate is around 48 % for all approaches as denoted in the Figure 5.10(e). For the approach using GMM the error rate increases around 13 mixture components.

A further observation is that weak performing classifiers can be improved using unlabeled data, whereas the ones, that show a comparably higher classification rate cannot benefit.

5.3.2 Evaluation of the Combined Classifier

To render a combined result of all available features a windowing technique is applied: From the outputs of the 6 classifiers a combined decision is computed on a one minute basis. The decisions of a classifier are collected in a time window of one minute length. This means that a final decision for a one minute

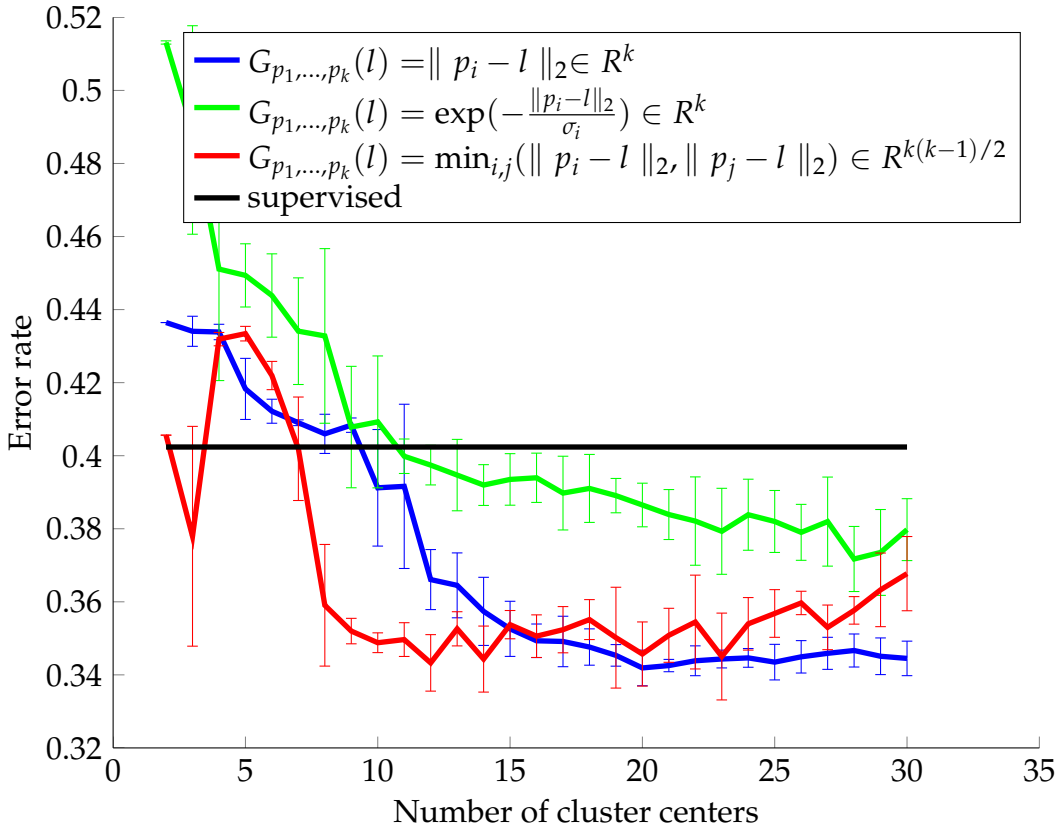


Figure 5.11: *Error rates together with the standard deviation of 10 runs for different unsupervised techniques and numbers of cluster centers for the combined classifier. The classical pseudo-inverse classifier is given as supervised reference as the black line.*

time window is formed of 5, 6 or 30 decisions, depending on the offset that is used for the extraction of the feature:

- 5 decisions based on *HVP* features (*12 s offset* per frame),
- 6 decisions based on the *respiration* (*10 s offset* per frame),
- 6 decisions for the *EMG* features of 20 s window and *10 s offset*,
- 30 decisions for the *EMG* features of 5 s window and *2 s offset*,
- 6 decisions for *SCL* features, that are computed on 20 s window and *10 s offset* and
- 30 decisions for *SCL* features, that are computed on 5 s window and *2 s offset*.

The decisions are first averaged within their respective channel and in a second step the fusion of the combined decisions for the channels is delivered. Thus, the approach accounts for the fact that the features of the different channels are computed in different temporal resolutions. Further, the temporal integration can improve individual classifiers to contribute positively to the over-all result.

A purely supervised reference approach is also evaluated. Here, the results, that are given by the supervised classifier are used to render a comparable result. The temporal integration and the fusion over the channels is conducted as described above.

The result of this combination shows promising results for both, the fusion of the individual decisions and the utilization of unlabeled data as well. The error rates of the three approaches and for different numbers of components are shown in Figure 5.11. The combined result for the conventional supervised classifier is shown in this figure as a black line.

The combination of the supervised classifiers results in a slight improvement of the best individual classifier from 41.0 % to 40.2 %.

However, the partially supervised alternatives are able to outperform this result in all cases at different numbers of components showing errors of as low as approximately 34 % (compare Figure 5.11, error plots). From about 13 to 15 components on, the supervised classifier is constantly outperformed.

The combined partially supervised classifier based on k-means and the Euclidean distance appears to be the most stable alternative. It outperforms the supervised approach at 10 centers and more and enters a saturation at an error rate of 34 % from 20 components on. The variance in the results is comparably small and using more centers does not increase the performance any more.

The second approach based on k-means using the pair-wise computation of distances is outperforming the supervised approach using fewer components at 8 centers. One reason is that the dimensionality of the new feature vector is increasing quicker than in the earlier case (i.e., $k(k-1)/2$ versus k) and hence it is more likely to find a linear separation hyperplane. However, when adding more centers the error rates increase together with the variance of the results.

Finally, the GMM based approach is showing the worst performance of the three approaches. The supervised case is still outperformed at approximately 12 to 18 centers, but only error rates of about 38 % are reached. Further, the variances are generally high in this case. The fact, that the individual classifiers show a worse performance than it is the case for the other approaches is obviously affecting the combined result.

For all partially supervised approaches, the combination of the different channels together with the temporal integration of the individual channels renders an improvement over the best individual classifier.

Generally, the variances between the classification rates for the individual subjects are very high in all cases. The standard deviations for the combined classifiers are on average 0.26, 0.21 and 0.25. For the individual classifiers the deviations range from 0.12 to 0.21. There are subjects that are relatively easy to classify and do improve with increasing number of centers and others that are by any means not classified correctly and do not improve. On the other hand, the inter subject variances are also high for the supervised approach: from 0.12 to 0.22 for the individual classifiers and 0.26 for the combined classifier.

5.3.3 The Influence of Unlabeled Data Samples

The above evaluation investigates the complexity of the over-all network with respect to its complexity (i.e., number of centers). Another question is how and how much of the unlabeled data can improve the classification. In order to provide insights into that, leave one subject out classification experiments in 50 repetitions are conducted, where the number of unlabeled samples in the preprocessing step is varied. Figure 5.12 shows the error rates for the approach using k-means clustering together with the Euclidean distance for numbers of cluster centers form 13 to 18. This distance measure was chosen as it turned out to be the most feasible one in the earlier experiments and 13 to 18 components were used here as this are the numbers of centers, where the improvement takes place. The figure also shows the results for the combined classifier. The x-axis shows the fraction of the whole available data, that is used as unlabeled. Please note that the axis is logarithmically scaled in order to show the measurements between 1 % and 10 % of usage of the data. The number of available data per feature type is shown in Table 5.1.

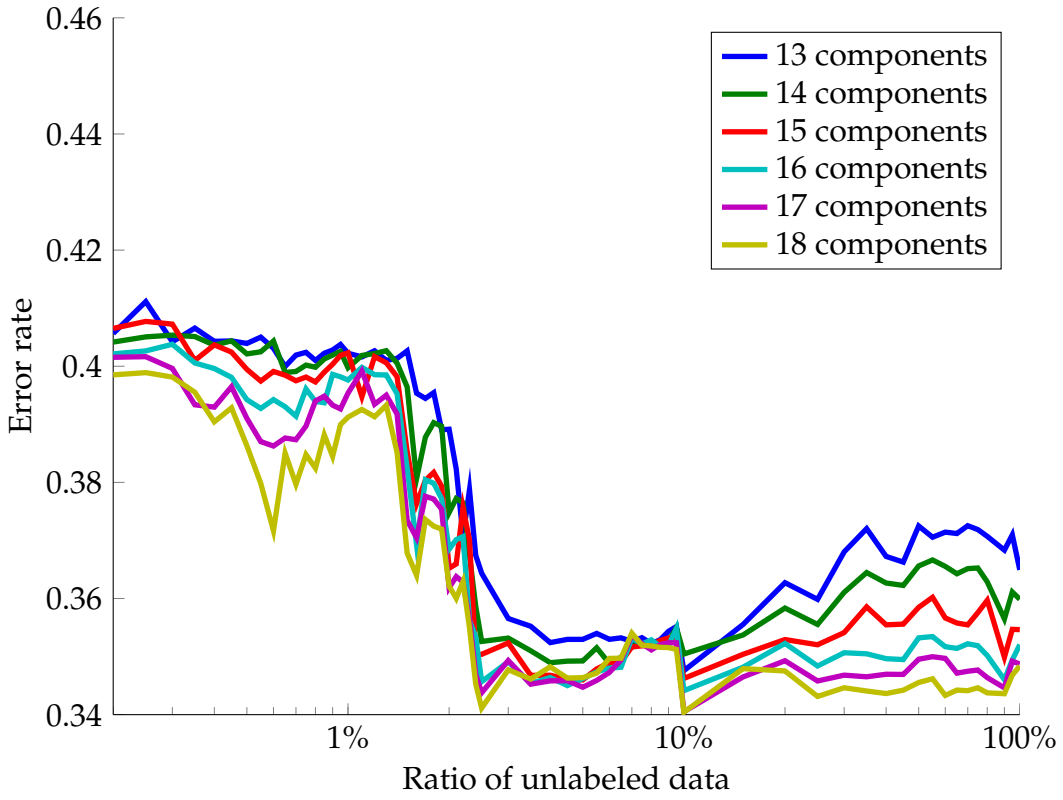


Figure 5.12: *Error rates for different ratios of all available data, that are used for the preprocessing for the combined classifier using k-means clustering together with the Euclidean distance. The classification is conduced using 13 to 18 centers. Every experiment was repeated 50 times.*

Table 5.1: *Over-all number of samples per feature type. Also the dimensionality of the respective feature vector is displayed.*

Feature type	Number of available samples	Feature dimensionality
Fast EMG	43972	4
Slow EMG	13427	18
Fast skin conductance	43972	8
Slow skin conductance	13427	5
BVP	9065	10
Respiration	13427	6

The first conclusion, that can be drawn is that the error is around 40 % for all cases, when very few data is used: when at around 1 % of the over-all unlabeled data is added in the training set of the clustering procedure. In the interval from 1 % to approximately 2.5 %, the error rapidly drops for all approaches, a little more for the cases, where a higher number of components is used. Here, an optimum is reached, especially for cases with small numbers of components. Adding more data affects the classification performance of the system, especially for 13, 14 and 15 components. The other approaches remain more or less stable on a low error level.

At first glance, the figure of 2.5 % is low. But please recall, that in this data set that equals approximately 1000 to 225 data points, depending on the number of samples available for the respective feature. The amount of samples for the different features is summarized in Table 5.1.

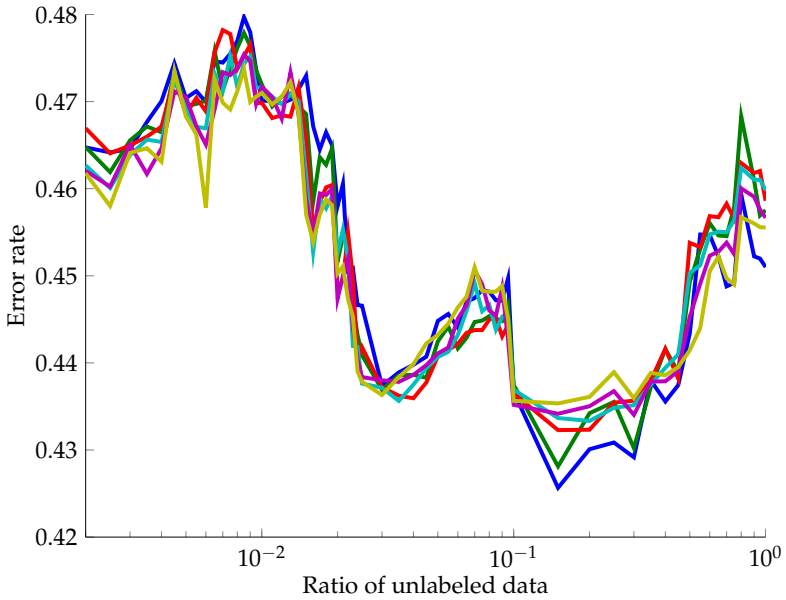
The results for the same experiment for the individual classifiers are shown in Figure 5.13. These results are unfortunately less conclusive than the ones for the combined classifier. In most of the cases, there is a local optimum for the error when between 1 % and 10 % of the available unlabeled samples are used.

5.3.4 Discussion

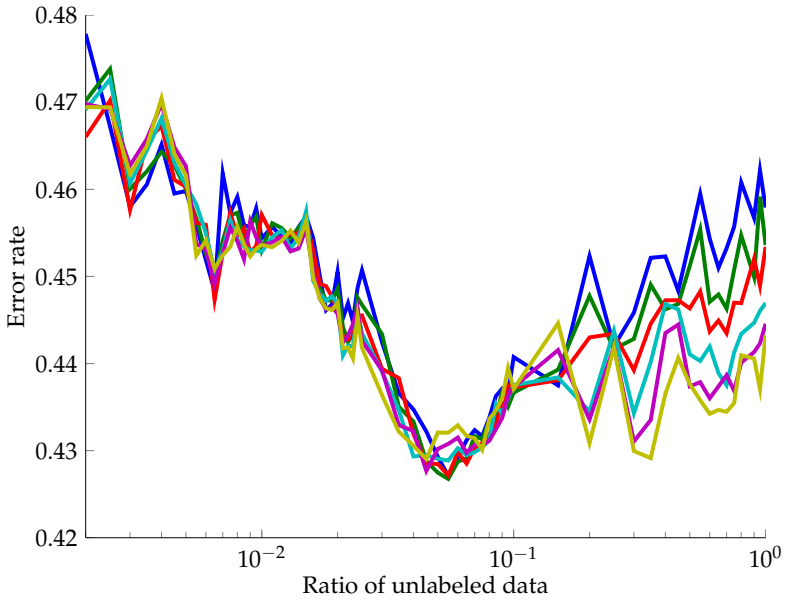
The goal of this work is to introduce a new learning approach that can incorporate unlabeled samples in applications with noisy data and class distributions. An unsupervised machine learning step as been integrated beforehand a supervised classification approach for this purpose. Using the prototypes or densities yielded from the cluster analysis, a new encoding of the data is computed, which is used then for the labeled data for classification

The method was evaluated using the classification of emotional states using physiological signals. The application provided us with labeled data for the times the elicitation process was conducted and also with data from other experimental sequences, which was regarded as unlabeled. Two different unsupervised learning strategies have been evaluated in this context: k-means as a prototypical and GMM as a generative approach. Furthermore, three distance measures were used to compute a new encoding based on the local densities.

The results for the individual classifiers show relatively low performances and the improvement over the purely supervised approach, if any, appears to be small. However, the combination of classification results in a time window and over the channels yields a noticeable improvement of the classification. From this, it can be concluded, that the decisions from the respective individual classifiers are more diverse (Kuncheva and Whitaker, 2003) than the ones of the purely supervised version. The lowest error rates are achieved at about

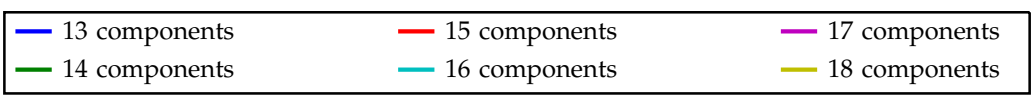


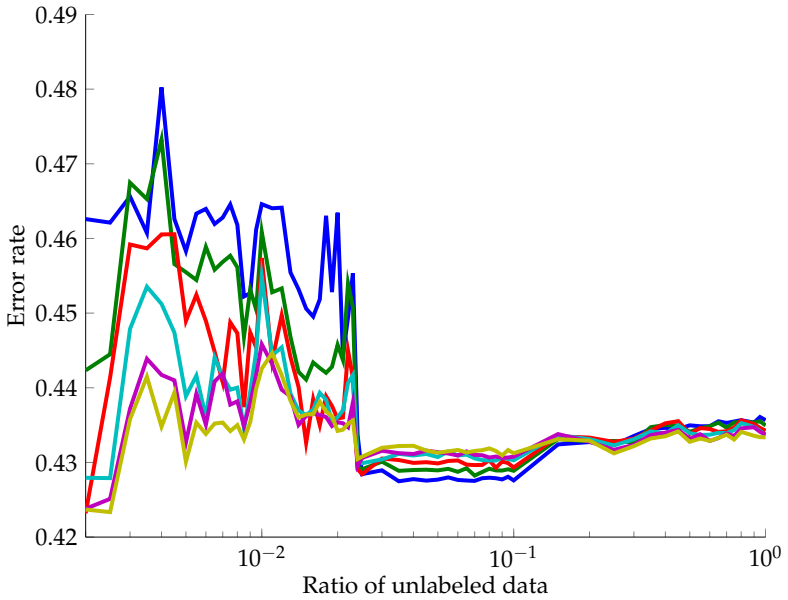
(a) Fast EMG features.



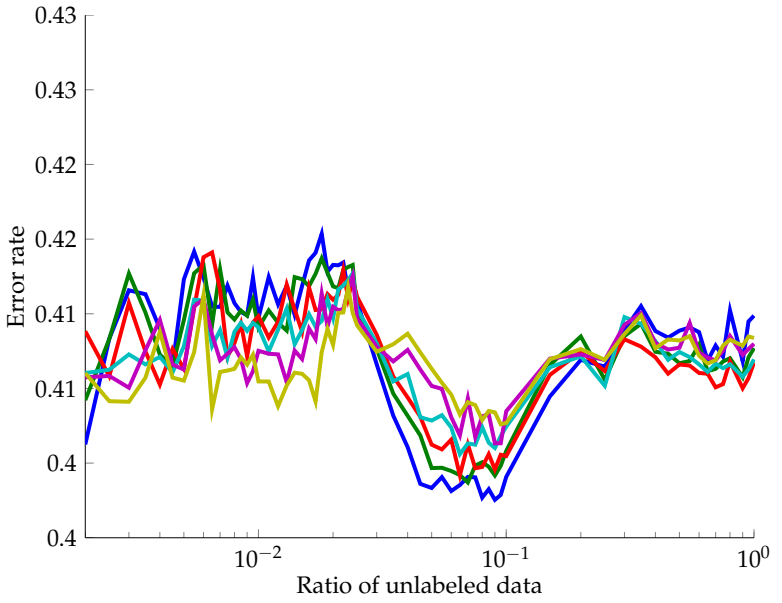
(b) Slow EMG features.

Legend:





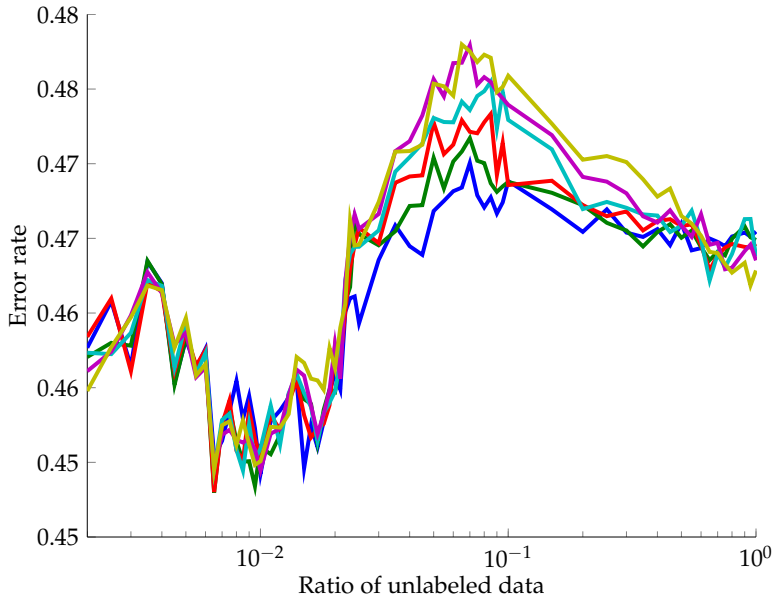
(c) Fast skin conductance features.



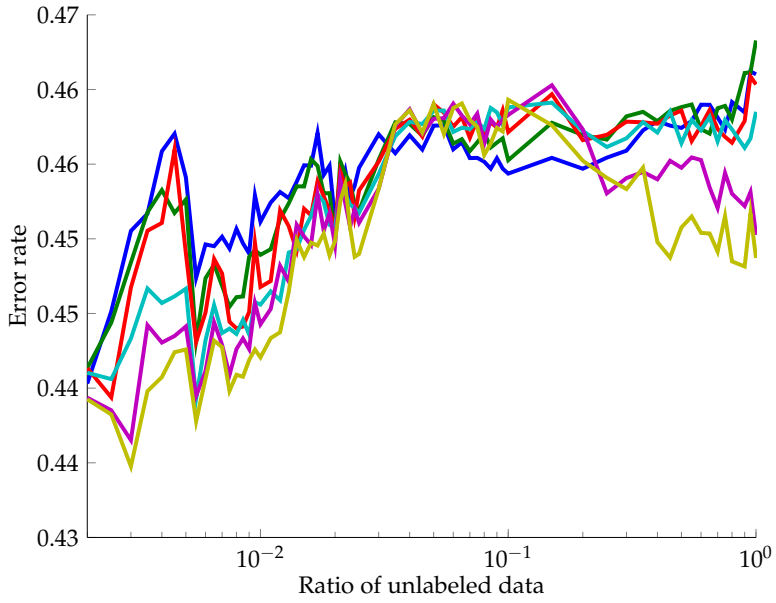
(d) Slow skin conductance features.

Legend:

13 components	15 components	17 components
14 components	16 components	18 components



(e) Heart rate variability.



(f) Respiration.

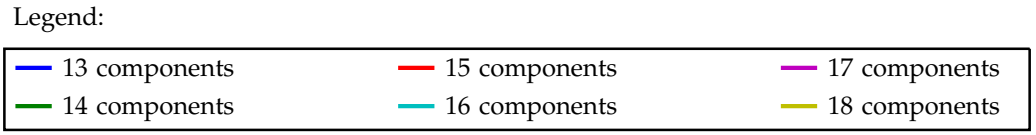


Figure 5.13: *Error rates for different ratios of all available data, that are used for the preprocessing for the individual classifiers for each feature type using k-means clustering together with the Euclidean distance. The classification is conduced using 13 to 18 centers. Every experiment was repeated 50 times.*

15 prototypes and more. This includes that the dimensionality of the data is approximately doubled.

Adding additional data the way it is conducted in the present experiments, i.e., data, that is not explicitly from the same categories, is of course only promising under certain conditions. If the samples of data resolved into clearly delimited classes, where the probability density functions for the different categories are non-overlapping and adding data from a very different partition would hardly be reasonable. But in many real world applications, this optimal setting for a classifier is not actually present: Often the data decomposes into severely overlapping distributions. There are also applications, where the particular classes are not (yet) irrevocably defined or such a definition is simply not possible due to distinct properties. Both circumstances are at hand in the application described earlier: On the one hand, the features that can be extracted from the physiological signals can be considered relatively weak for the inter individual classification. Obviously the physiology heavily depends on the individual subject. On the other hand, even though the induction of the intended emotion succeeds in the average case, it is not guaranteed by any means that every particular sample is correctly labeled.

A further step for this approach could be to label the unlabeled samples according to the computed local densities. An intuitive way could be to use fuzzy memberships according equal to the labels of the samples in the respective clusters. Hence one could further improve the supervised part of the work.

5.4 Fuzzy Output Support Vector Machines for Unbalanced Class Distributions

This particular setup imposes several challenges: the oddball recording technique of the data requires a special treatment due to the skewed distribution of classes. Heavily imbalanced datasets require special treatment in order to mitigate the over-representation of a class. Popular techniques are under- and over-sampling of the training set with respect to the categories or the usage of error functions that account for skew distributions of classes (Japkowicz, 2000; Zhou and Liu, 2006). Also, the noisy nature of the employed sensors can impair the recognition performance. Methods designed to improve robustness in low signal to noise ratio conditions include low pass filtering but also information or data fusion (Kuncheva, 2002a). In this particular domain of information fusion various possibilities to ensure robustness can be applied. In our approach, robustness is achieved by combining multiple feature channels as well as by combining the outputs of several independent classifiers.

Further, a decision fusion procedure was implemented and a comparison be-

tween the genetic algorithm based classifier selection and a standard averaging fusion approach was conducted. Generally, the combination of classifiers succeeded in improving the over-all performance. This can be interpreted as a indication for the beneficial diversity of these classifiers in combination with others.

This evaluation showed that in this application the relatively computationally expensive search procedure did not bring significant improvement. Nevertheless, there may be an advantage in the selection process regarding the complexity of the classifier. Discarding some of the available classifiers reduces the efforts that have to be spent not only on the classifier fusion step, but also these classifiers obviously do not have to be evaluated. This reduction comes with the drawback of a costly search procedure which is, however, conducted off-line prior to testing. Furthermore, the fact that there are no tremendous differences concerning the length of the bins in the histogram might as well be an argument to explain, that skipping the classifier procedure does not decrease the performance.

5.4.1 Construction of Individual Classifiers

Before the computed features are classified using the proposed machine learning approach, the EEG channels were partitioned into nine overlapping areas containing up to 18 channels at a time. Eight partitions are chosen as coherent slices and the ninth one is defined as the horizontal and vertical cutoff of the EEG device's layout (see Figure 3.13 for an overview). These partitions were defined from a machine learning point of view rather than from physiology: thus, one can provide more information for individual classifiers than using only one electrode, but it is still possible to conduct decision fusion. For every partition, the features extracted from the respective electrodes were concatenated to form a new feature. Thus, a feature level fusion approach is realized and by doing so the individual classifiers that are constructed based on this representation are supposed to get informative input.

Subsequently, the resulting 45 sets of channels – due to the combination of the five kinds of features with the nine partitions – were trained and classified separately. Preliminary cross validation experiments for k-nearest neighbors, multi layer perceptrons and various types of SVM showed, that weighted SVM – with Gaussian kernel – outperforms the others by the area under the ROC curve (AUC). The SVM that conducts the classification is trained using a Gaussian kernel. As described in Section 4.3, a loss term was integrated in this SVM approach tackling the issue of the imbalanced training and test sets as only 48 positive samples are available opposed to as many as 2700 negative samples. The performance of the classifiers is determined by the AUC. The

results of this first classification step are ranging from a classification performance of 0.478 AUC, which is close to random, to 0.836 AUC. An overview of the performances of all constructed classifiers is depicted in Table 5.2, showing the average of 15 six-fold cross-validation rounds.

It can be observed from Table 5.2 that the individual performance depends on the chosen feature extraction approach. Especially the features extracted from the real part of the FFT and from time domain results in strong performances, while features from phase shift reveal rather weak performances. On the other hand the actual partition seems to be less important considering this measure: the variability in the columns is relatively small compared to the previously mentioned findings.

5.4.2 Classifier selection and fusion using genetic algorithms

Algorithm 5.1: Multiple classifier system for the classification of ERP

Input:

- Set C of individual classifiers with $|C| = n$
- Individual classification rates on the validation set

Randomly generate a pool of classifier teams of size k as binary vectors $r \in \{1, 0\}^k$ and $r_i = 1$ if the classifier a member of the team;

for $1 \ldots N$ *iterations* **do**

1. Cross-over: split of the binary vectors and re-combination
2. Mutation: bit-flip with probability p
3. Evaluation of the performance of the new classifier teams
4. Selection: choose the k fittest classifier teams

Output: The optimal team of classifiers on the validation set

In order to further enhance the performance of the proposed classifier, a decision fusion step was implemented to combine the obtained outputs. An averaging classifier fusion was applied for the combination of the individual classifiers because of stability reasons (Kittler et al., 1998). In order to find a suitable combination of classifiers, a basic genetic search algorithm approach was implemented as described in Algorithm 5.1. The classifier selection was optimized locally in every cross validation run: a validation set – i.e., the data

Table 5.2: *Performances of the constructed individual classifiers in terms of area under the ROC curve. The numbers refer to the partitions defined in Figure 3.13. The classifiers group themselves regarding the performance by feature types: on the one hand features from the real part of the FFT and PCA in time domain perform well, on the other hand features from phase shift coefficients are only slightly better than random. The standard deviation of the conducted runs is given in parentheses.*

Partition of Electrodes	Real part of FFT	Imaginary part of FFT	Amplitude	Phase shift	PCA over time
1	0.622 (0.089)	0.733 (0.091)	0.669 (0.083)	0.480 (0.087)	0.714 (0.101)
2	0.744 (0.087)	0.774 (0.081)	0.593 (0.092)	0.586 (0.088)	**0.836** (0.065)
3	0.758 (0.093)	0.723 (0.090)	0.573 (0.082)	0.497 (0.042)	0.794 (0.082)
4	0.795 (0.078)	0.711 (0.085)	0.610 (0.076)	0.481 (0.082)	0.811 (0.078)
5	0.689 (0.095)	0.698 (0.102)	0.602 (0.086)	0.502 (0.089)	0.659 (0.111)
6	0.677 (0.102)	0.764 (0.086)	0.577 (0.084)	0.478 (0.085)	0.750 (0.090)
7	0.654 (0.096)	0.722 (0.097)	0.493 (0.089)	0.481 (0.071)	0.784 (0.070)
8	0.630 (0.075)	0.755 (0.082)	0.682 (0.086)	0.468 (0.092)	0.796 (0.084)
9	0.661 (0.101)	0.714 (0.098)	0.650 (0.083)	0.522 (0.084)	0.737 (0.100)

of one fold of the training data – is left beside in the training of the individual classifiers and the fitness of the classifier ensembles is determined and optimized on this set. For the subsequent experiments, the number of maintained individual solutions and the maximum number of epochs were set to 10.

Table 5.3: *Results of the classifier fusion and the classifier selection approaches in terms of area under ROC curve. The results including the classifier selection step are based on 76 search attempts. The results without classifier selection are computed with 15 separate 6-fold cross validations. The standard deviation of the conducted runs is given in parentheses.*

Fusion procedure	**AuROC**
classifier fusion/selection with GA	0.860 (0.074)
classifier fusion solely	0.853 (0.081)
best individual classifier	0.836 (0.065)
random forest with class weighting	0.789 (0.047)

Results for this averaging classifier fusion approach are reported in Table 5.3 with and without classifier selection procedure. Using classifier selection, combinations of classifiers, which further increased the area under the performance were found: The performance of the classifier selection process on the test set is 0.860 AUC while the actual performance when skipping the selection process is marginally smaller (0.853 AUC). Both approaches do actually outperform the optimal individual classifier showing the highest performance (see Table 5.2) . These results reveal that the classifier selection step does only yield marginal benefits in this particular application.

The number of selections of the 45 individual classifiers in 4050 classifier selection experiments is depicted in Figure 5.14. Even though the discriminant message of this figure may be subtle one could argue that the features computed from the amplitudes of the FFT (green) and maybe the phase shift (cyan) or the real part (blue) are selected more often than the others. Especially in case of Phase shift features this is surprising, because these features show relatively poor performances in Table 5.2.

In order to compare the proposed approach to state of the art classification techniques the experiment has also been conducted using Breiman's random forest (Breiman, 2001) extended as proposed by Chen et al. (2004), which takes into account the imbalanced class distribution by weighting with respect to the frequency of the classes. The random forest classifier has been trained using 500 individual trees. These results are also listed in Table 5.3. It can be observed that this state of the art approach is outperformed by the proposed SVM in this application.

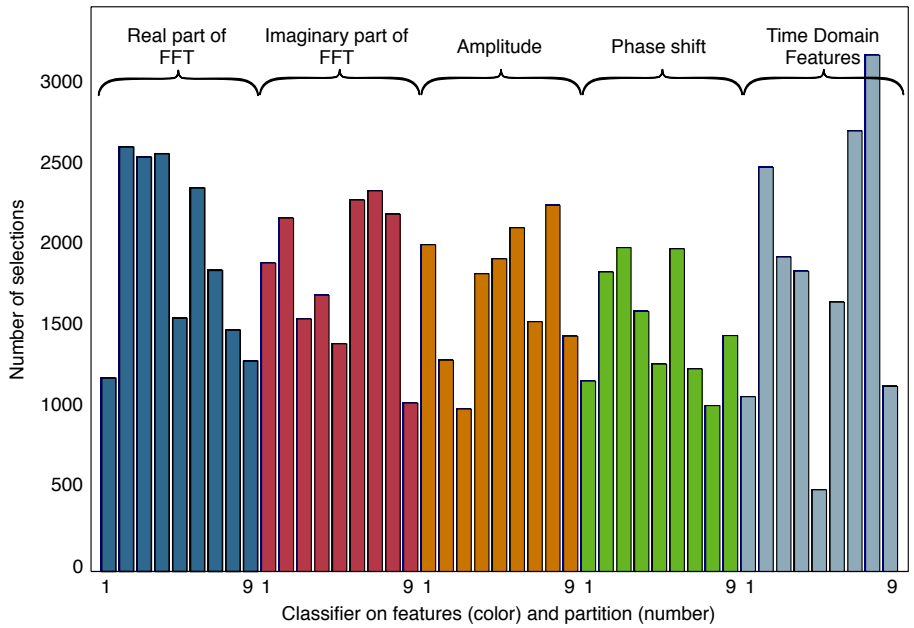


Figure 5.14: *Number of appearances of a particular classifier being selected: the x-axis depicts the different partitions of the electrodes using digits 1 to 9 and for the 5 feature types that are color encoded: real and imaginary parts of the FFT are plotted in blue and red, amplitude and phase shift are plotted in green and cyan and finally the feature in time domain is shown in black. The y-axis shows the number of selections.*

5.4.3 Discussion

An information fusion approach to discover ERP in EEG data recorded using an oddball paradigm is described and numerically evaluated in this section. Firstly 45 individual classifiers were trained using various feature extraction strategies and a low level information fusion technique defining partitions of EEG electrodes. Utilizing this first fusion step, the individual classifiers could be constructed to reveal a good performance even though the underlying data is noisy and the distribution of the isolated features of a channel are heavily overlapping.

6 General Discussion

The goal of this thesis was to contribute to the field of human-computer interaction by making the communicational process more intuitive and more accessible for a human user of a system. The dialog with a computer is adapted to more humanly ways of interaction, that are very common in inter human communication but are up to the present times not taken into account in the interaction with technical systems. Besides the increasing computing and storage capabilities, that are available as the technical development advances, the availability of cheap sensor devices makes it feasible to reach that goal. These sensors are very prominently audio and video devices, but also recording devices for physiological signals such as skin conductance or electromyography.

Different novel approaches for the classification of user states in human-computer interaction have been proposed in this work and they were extensively evaluated based on various data collections. The focus of the thesis is on the estimation of emotional states in real world affective corpora during the interaction with a technical system. One issue in this context is the investigation of information fusion architectures to combine classifications of multi-modal and continuous affective computer interactions into a unified categorization. Further, the incorporation of unlabeled data into the classification process is investigated to improve unreliable classifiers, that are generally obtained in natural human-computer interaction. Thus, the classification of physiological signals can to some extent be extended from subject-dependent to inter-individual affect recognition.

Another contribution of the work lies in the learning of statistical classifiers in the context of imbalanced class distributions in training sets as it regularly is the case in human-computer interaction. It is implemented by the means of a class weighting mechanism into the fuzzy-input fuzzy-output SVM. The algorithm has been evaluated on a corpus of EEG signals, where the recording paradigm results automatically in very imbalanced class distributions. Also, a decision fusion scheme and a classifier selection approach were realized and evaluated on the same data.

6.1 Information Fusion Architectures in HCI

The human-computer interaction is approached in this work by the means of statistical pattern recognition and the combination of individual classifiers in multiple classifier systems. This architecture relies specifically on independent individual classifiers, that are most likely obtained if the underlying features are per se independent. This is particularly the case when the application provides different sensors as it is the case here as outlined earlier. A unique challenge in this context is posed by the varying inherent technical properties of the different data sources such as their sample rates and their respective scales. This demands for more complex fusion architectures compared to traditional classification problems as they can be found in many benchmark data sets.

Beside the multi-modal fusion over different sensory outputs, the temporal integration of intermediate classification results of one channel helps to increase the recognition performance. It has been shown in this thesis, that the different emotional categories show a relatively distinct optimal time granularity in which the classification error is minimal. This can also be used to gain a deeper understanding of the underlying emotional concepts, revealing in which frequency the different categories may alternate in a free human-computer interaction.

This differs obviously between the concrete applications, however it was shown that the optimal classification rates are rendered in a relatively slow frequency compared for example to the sample rates of the different sensors. A further insight into the emotional label system in a realistic human-computer interaction scenario is that the different categories are conveyed in a different intensity by the individual modalities. For example the label arousal is detected most successfully using only the audio signal whereas considering the other labels the facial expression analysis is generally more successful.

6.2 Annotation of Data in the Context of HCI

The evaluation of the approaches is conducted mainly using corpora of affective human-computer interaction, namely the EmoRec and the AVEC 2011 corpora. These kinds of data collections lead the way towards a new understanding of human-computer interaction by incorporating the emotional dimensions of human communication into the over-all process. This marks a starting point for a broader and deeper integration of artificial technical systems into the everyday interaction with computers in the future and the underlying concepts are still under debate. The two main approaches to the design of an expressive corpus under relatively unrestricted conditions are also reflected

in the employed data sets: One possibility to record an affective corpus, which will be referred as "global" labeling, is the usage of predefined stimuli, that are presented to a test subject.

Such stimuli are commonly implemented as feedback from the system, for example praise from the technical interlocutor or delayed reactions, which the subject is interacting with. Hence, stimuli that are thought to elicit similar reactions of the user are presented in blocks to produce continuous sequences of the same category. This global approach inherits several unique features, that makes it favorable for some researchers: It does not require any post-hoc labeling procedure, which is often expensive and vulnerable to errors and it asserts that the experimenter retains the control over the recording. However, the approach comprises also particular drawbacks, that have to be taken into account: It is not feasible to assert the success of the individual stimulus to elicit the desired emotional state. This makes the resulting label uncertain, even in the ground truth of the data, which makes the training of a classifier a challenging task. In principle, it can be said, that it is very easy to frustrate a user by emulating a broken user interface or by overburdening the user with unsolvable tasks. Eliciting other reactions is unfortunately by far more difficult.

A further property that is often encountered in these kinds of corpora is that the experimental setup follows a distinct temporal order, resulting in an order of segments, that are labeled in the same class. The complementary approach, which might be analogously called local, is to manually label the material after it is collected. This makes it easier to provide a recording scenario, that allows a free interaction of the user with the provided artificial interlocutor. Also, no inherent temporal structure of the data is given by the experimenter, which makes it easier for an automatic classification system to avoid to fall for temporal artifacts. However, this labeling procedure is unfortunately expensive and prone to errors as the true emotional state of a subject is not easily accessible for an external person, if at all.

This demands for the usage of multiple labelers for the same recording, hoping that the labeling errors can be averaged out by doing so. This further increases the costs of labeling and corpora that are labeled by more than three subjects are relatively rare. Another issue that arises in the labeling procedure is in the timing of the succeeding labels. As the recordings are naturally provided in time sequences, the segmentation is a non-trivial problem, which can result in skewed label traces. In summary, it can be stated that the resulting annotations, and hence the teacher signals for a training process, are weakly defined compared to other pattern recognition applications.

6.3 Feasibility of Unlabeled Data in Classification

The general difficulties in the annotations of real-world affective corpora and the fact that there is commonly only few data for the different subjects available in the training sets makes it appealing to use additional unlabeled data in the training in the classifier. Unfortunately classical approaches of semi-supervised learning rely on base classifiers, that show initially a high recognition rate in order to avoid to add false classifications to the training set of the classifiers. In this thesis a more robust approach to incorporate unlabeled data into the training process by conducting a clustering approach or density estimation on all available data in order to compile a new representation of the data by computing the distances to the different centers.

It has been shown experimentally that the approach improves over a purely supervised approach for noisy class distributions and only few labeled data samples at hand. Together with the combination of the classifier outputs from multiple modalities and their temporal integrations, the partially supervised approaches are distinctly suited for the application of human-computer interaction.

6.4 Integration into a Greater System

The recognition rates, that can be achieved in unconstrained human-computer interaction scenarios are unfortunately still by far lower than the ones that are rendered for acted data in comparable classification studies. This originates mainly from the fact, that the acted data collections reflect a wide variety of expressions, that comprise mainly full blown occurrences of emotions. However, these full blown expressions are not likely to occur in realistic everyday interactions between humans interlocutors and are even more unlikely in interactions with computers.

The term “emotion” is a broadly defined term, that is obviously not the only factor, that is and will be important for the interaction with technical systems and the definition and recognition of more complex user states as an input to an interface. Further relevant information could be long-term goals of the user, specific tasks that have to be conducted by the means of a technical system, the interaction history of a user with different devices and of course traditional direct inputs to an application.

The automatic recognition of the human emotional categories should hence be viewed in the context of a greater framework, which integrates the over-all interaction and assistance for a task, that is relevant for the user. This implies for the over-all system that it may not be necessary to query a new classification

for the emotional user state at every technically possible instant of time.

A very intuitive way to conduct this is to access the certainty of the individual decisions of a decision by maintaining fuzzy values further after the respective information fusion steps, that are passed to the next level of processing. The classification could thus be rejected by the application if the certainty of the decision is smaller than a threshold, that could also be dynamically adjusted. Another promising approach is to incorporate feedback from the application into the classification. Such a feedback could be, for example, a newly labeled sample, that is passed to a re-training mechanism in order to improve or personalize a distinct classifier.

Another alternative is to exploit high-level knowledge about the application in order to improve the classification performance. This technique is also well-suited to avoid false alarms, that are likely to disturb a natural interaction with the system. A very intuitive example for that is that the design of an interaction process favors different expressions compared to others, for example when a difficult step of a task has to be conducted inherently for the application. This could be incorporated into the classifier by manipulating the prior probabilities of the different categories, which is also a thinkable approach to conduct an adaption of the classification system to a specific user.

6.5 Conclusion

The incorporation of the results of the affective computing community into next generation user interfaces is still an open issue for research. It is particularly non-trivial to find obvious actions for a traditional computer interface, which is designed in a conservative imperative paradigm, which neglects every kind of subtle form of non-verbal communication. However, future interfaces for technical systems will embrace the user much more, than it is possible in the present technologies. The recent developments in the computer sciences and related fields of research aim at the incorporation of such new categories into the interaction of users with technical systems. This is a highly interdisciplinary objective, which is challenging for both, the pattern recognition and the psychological parts of this undertake. This thesis is intended to be a part of this development to bring the human and the technological counterparts of new user interfaces closer together.

7 Summary of the Contributions

Novel user-interfaces for human-computer interactions will bring the communication with a technical system beyond the present question and answer paradigm. This will be enabled by the means of new input modalities and additional sensory channels such as microphones, cameras and various physiological sensors. The additional information is processed according to psychological shindigs from the emotion theory and also the well-known patters in physiological signals, such as the ERP in EEG signals.

In the following, the three main contributions of this thesis in this context are summarized. In Section 7.1, the contributions in the context of temporal and multi-modal fusion architectures are outlined together with the results of the numerical evaluations on two affective corpora. Further in Section 7.2, the contributions in the context of the incorporation of unlabeled data for unreliable classifiers is summarized. Also, the findings of the numerical evaluations for the proposed approach in the context of subject-independent classification of physiological signals are outlined. Finally, the contributions for the training of statistical classifiers with training sets having an imbalanced class distribution are described in Section 7.3. This comprises the extension of the fuzzy SVM to imbalanced classes and its evaluation in the context of the detection of ERP in EEG data.

7.1 Multi-modal and Temporal Fusion

In order to detect user states in human-computer interaction multi-modal decision fusion architectures were proposed and evaluated using the AVEC 2011 and the EmoRec corpora (Schels et al., 2014a, 2013a,b, 2012a, 2009; Schels and Schwenker, 2010; Glodek et al., 2013a, 2012b, 2011; Meudt et al., 2013; Scherer et al., 2012a, 2011; Schmidt et al., 2010; Walter et al., 2011; Schwenker et al., 2010). These fusion architectures comprise a combination of the decisions of

different sensory channels but also the temporal integration of preliminary decisions in time series. The evaluations were conducted to investigate the size of the time windows that are optimal for the detection of the user states and also the optimal order for the combination of the modalities.

The temporal integration of the intermediate classification results turned out to be successful in all experiments, drastically reducing the classification error for the individual modalities for longer time scales. However there are differences in the results for the different corpora, that were investigated. There is a clearly optimal time window for the recognition of the different categories in the AVEC 2011 corpus. These vary from approximately 10 s to up to 70 s. For the EmoRec corpus, even longer time windows render optimal results for the discrimination of the experimental sequences. In this case the optimum is not so clearly observable but after 100 s – 120 s the performance of the classifier does not increase further. An argument for these differences are the differences in the recording paradigms of the corpora: The EmoRec labels are set globally over a distinct period of time which leads to longer labels compared to the manually set labels in the AVEC corpus.

The temporal integration experiment clearly shows that the classes do not change very quickly over time, which allows longer time granularities. Also, the intermediate frame-wise classifications are comparably weak, such that many individual decisions are needed to improve the final result. A rather notable result is that the optimal recognition error is rendered on comparably longer time scales than it is reported for more direct physiological reactions to external stimuli. For example compared to the approximately four seconds time window that is needed for the skin conductance to react after an external stimulus in the literature. Improving the individual classifiers using this approach also makes multi-modal fusion more promising.

The results of these multi-modal fusion experiments are in their details dependent on the different applications and the respective categories. The category arousal is clearly better conveyed by the voice of the subject than using the facial expression and hence the combination of the audio classification with the video does not decrease the classification error. For the labels expectancy and power the multi-modal fusion improves over the best individual classifier: In both cases, combining first the audio and video classification results and conducting the temporal integration afterwards on the combined decision is the more eligible approach. This results from the fact the video classification is more reliable than the classification based on the audio solely and this approach gives more weight to modality, that provides its samples more frequently. The category valence is clearly better classified using the video signal compared to the audio classification. The combination of the two modalities does unfortunately not improve the over all classification. For experiments on

the EmoRec data set, the audio classifier performs better than the one based on the video, which can be explained by the arousal, that is in the definition of the experimental sequences. However the facial expressions also render better results compared to only arousal in AVEC corpus. One reason for that is that the valence parameter also is varied in the experimental sequences. The combination of the two modalities further improves the classification by giving equal weights to the different channels. This reflects the comparably good performance of the classifiers.

In summary, it can be stated that the different affective categories are conveyed in different ways: Some are better recognized using the audio channel and others are better classified on the video channel. In all cases, the integration of multiple succeeding decisions improves the classification of the affective state. The success of the multi-modal fusion depends notably on the respective label. One important factor is the performance of the individual channels, which should not differ too much on average. But not only the individual performance is important in order to improve over the best individual classifier but also a certain diversity of the classifiers and together with correct confidence estimates are important for the combination.

7.2 Partially Supervised Learning in Human-Computer Interaction

A further contribution is in the field of partially supervised learning in the context of human-computer interaction (Schels et al., 2013a, 2014b, 2012a,b, 2011; Hady et al., 2010a). The estimation of probability densities or cluster prototypes under the usage of additional unlabeled data was used as pre-processing for a further supervised training step. This can be viewed as a classification network with one hidden layer that is trained in an unsupervised fashion based on additional unlabeled data. The approach allows the usage of different clustering techniques and a variety of distance measures for a re-encoding of the data. The complexity of the resulting network can be regulated by the number of cluster centers or density components in the unsupervised learning approach. A further important parameter for the approach is the amount of additional unlabeled data, that is provided to the training process.

The approach was evaluated using the physiological part of the EmoRec corpus. This corpus provides, besides the physiological data, audio and video recordings. However, it is particularly compelling to use additional data for the subject independent classification of the physiological part as the features are computed on a longer time scale, which results in fewer data. The other channels are sampled in the order of milliseconds, which leads to a denser dis-

tribution of data. Also, a decision fusion approach, where the classification results for different channels are temporally integrated and afterwards combined over the different physiological channels is implemented. Thus, the approach incorporating the unlabeled data is outperforming the purely supervised reference approach.

The approach has been investigated with respect to the complexity of the resulting network and how the unlabeled data improves the classification. It has been shown, that the error of the classification decreases with the number of centers. The re-encoding approaches, that render a higher dimensionality for the representation of the data have lower error for fewer centers. A further notable result is that the variant based on k-means clustering is the most stable investigated approach over the number of centers. It converges for approximately 20 center at an average error of 34 %. Using a GMM as the unsupervised step together with the a posteriori probability for the individual components, the classification error decreases lesser. A reason for that could be in the estimation of the covariance matrices, for which usually lots of data is required.

Secondly, the amount of additional unlabeled data samples that are necessary to improve the classification is investigated. In principle up to 9,000–50,000 unlabeled data points from the additional experimental sequences and the different subjects are available. However it is shown that after comparably few samples, the error does not decrease further. This is concretely the case after incorporating approximately 2 – 3 % of the available data, which corresponds approximately 180–800 samples, depending on the channel and the feature type. For this sample size, the underlying distribution of the data seems to be adequately estimated for the subsequent classification.

The results for these individual classifiers that are defined for mainly the different physiological channels, are also outlined in the experimental section. However, the results for these classifiers are not as conclusive as the ones for the combined classifier. This shows again, that not only the accuracy of the individual classifier is important for a multiple classifier system, but also the diversity of the classifier team. Thus, the combination of the classifiers bears the opportunity to correct false decisions for the individual channels.

7.3 Imbalanced Classes

Many data collections, that represent real world problems have to deal with the problem that the distribution of classes is not balanced for all classes (Schels et al., 2013c, 2010). Hence a class based weighting algorithm is developed, that is derived from the fuzzy input fuzzy output SVM in order to mitigate the

imbalanced class distributions in the training process of a classifier. The fuzzy SVM uses weights for the class membership of a sample in the loss function to find an optimal hyperplane. In this work the weighting mechanism was adapted to define a higher loss value for the samples in the underrepresented classes.

The algorithm was evaluated using the Pascal 2 mind reading competition data collection, which comprises a recording of EEG. For this data, the P300 ERP is elicited using the oddball paradigm, which means that the target stimuli is presented only sparsely compared to the background patterns. This paradigm naturally results in imbalanced class distributions. Additional to that, a multiple classifier system is constructed, where different feature approaches and different overlapping partitions of EEG electrodes are used to obtain diverse classifiers. This renders relatively accurate individual classifiers and gives the opportunity to improve over the best individual classifier. The combined classifier outperforms the best individual classifier and also competing approaches. This is further supported by the fuzzy output of the SVM where confidence values are used for the estimation of the class memberships. A subsequent classifier selection procedure was conducted on a validation set, which led to no further increase in performance. One reason for this is that a further validation set for the selection has to be split from the initial training set, which leads to a decrease of accuracy for the individual classifiers.

Appendices

A Partially Supervised Results for Standard Data Sets

In this chapter of the appendix, the results of a further evaluation for Algorithm 4.3 are shown for different standard benchmark data collections, which are mostly retrieved from the well-known UCI Machine Learning Repository (Bache and Lichman, 2013). The used data sets from this source are namely:

- "Ionosphere"[1]: 351 samples; 2 classes
- "Iris"[2]: 150 samples; 3 classes
- "COIL-100"[3] (Nene et al., 1996): 7200 samples; 100 classes
- "Fruits" (Fay, 2007): 840 samples in 7 classes.

For the evaluation of the partially supervised approach a predefined portion of the data is regarded as unlabeled by neglecting the given label for the respective samples. This portion is varied as seen in the different sub-figures of figures A.2, A.4, A.5 and A.6. The number of cluster centers is also varied in order to examine the architecture for partially supervised approach, which is plotted over the x-axis of the sub-figures. The supervised stake in the approach is implemented using a Moore-Penrose pseudo inverse linear classifier. For comparison, classification experiments are also conducted only using the labeled portion of the data sets. Three different classification approaches are evaluated in this context: a further purely supervised Moore-Penrose pseudo inverse, a support vector machine using a linear kernel and an additional SVM using a RBF kernel. Generally the supervised pseudo inverse classifier renders lower classification rates than the linear SVM, which renders itself lower classification rates than the nonlinear RBF SVM. The partially supervised approach

[1] `http://archive.ics.uci.edu/ml/datasets/Ionosphere`
[2] `http://archive.ics.uci.edu/ml/datasets/Iris`
[3] `http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

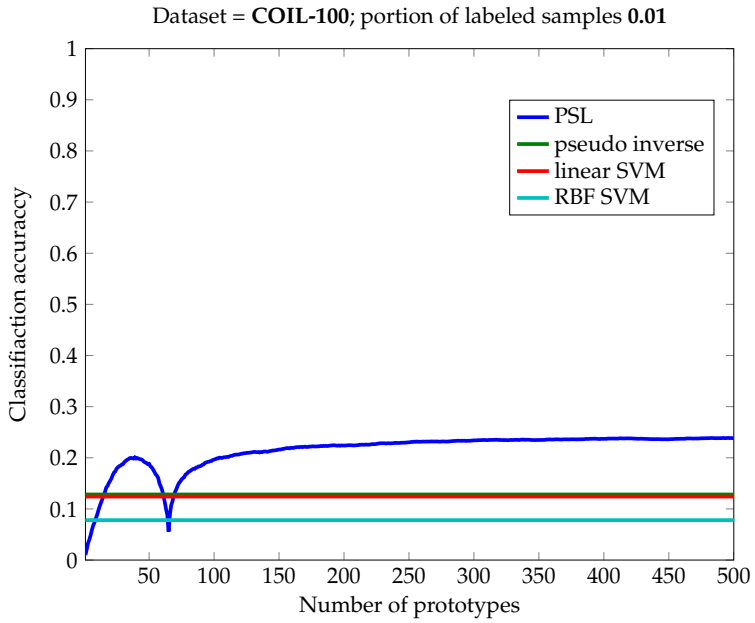
yields generally speaking equal or sometimes higher recognition rates if the number of cluster centers is large enough. However, when the number of cluster centers approaches or is equal to the number of labeled data samples, the drops dramatically, which is a phenomenon, that is well established in the literature, compare for example (Hoyle, 2011; Schäfer and Strimmer, 2005). The mathematical reason for this lies in the fact that for this setting, the number of non-zero eigenvalues that are small is increased.

A.1 COIL-100 Data Set

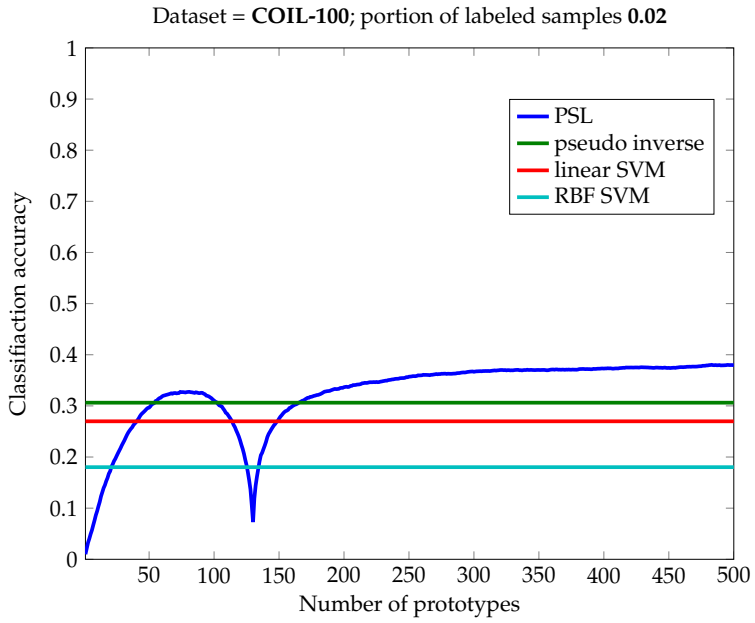
The Columbia University Image Library (COIL-100) is used as one data set for the evaluation of the partially supervised algorithm. It comprises visual images of 100 different objects taken from 72 defined point of views. The resolution of the images is 128×128 pixels. Samples taken from the COIL-100 library are shown in Figure A.1. Following (Fay, 2007), color histograms of size 24 were extracted from each of the color channels of the images. Concatenating the three resulting histograms yields a feature vector of size 72.

Figure A.1: *Sample images from the COIL-100 data collection (Nene et al., 1996).*

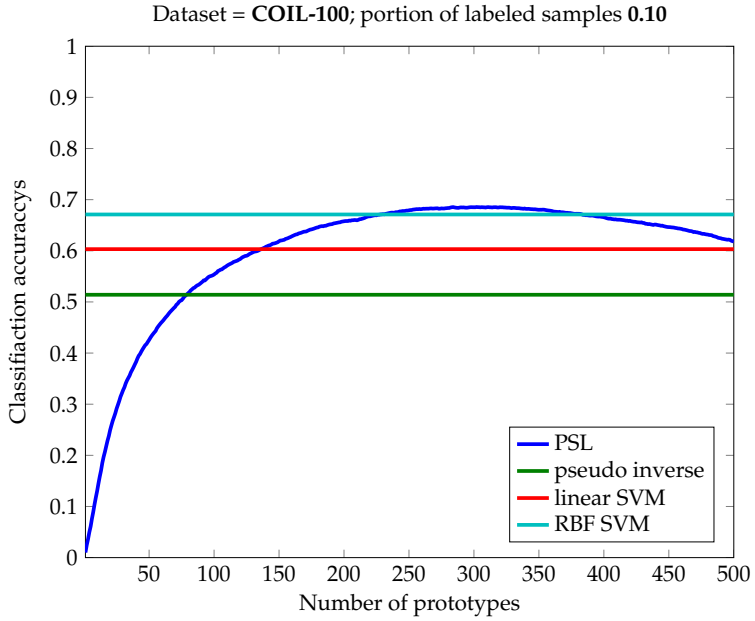
The numerical evaluations of on the COIL-100 library are presented in Figure A.2 for different portions of data for which the labels are removed. The accuracy of the classification increases generally when additional labeled data is added to the training set. However, the partially supervised approach suffers lesser from the absence of the labeled data than the purely supervised approached, that are also evaluated. Only, when the number of labeled samples is approximately equal to the number of prototypes in the setup, it can be clearly observed, that the pseudo-inverse is failing (compare Figures A.2(a) for 70 prototypes and Figure A.2(b) for 140 prototypes). Adding more labeled samples to the training set significantly improves the accuracy as seen in figures A.2(c) and A.2(d). In these cases, the accuracy of the partially supervised approach outperforms the linear classifiers and shows approximately the same performance as the RBF SVM.



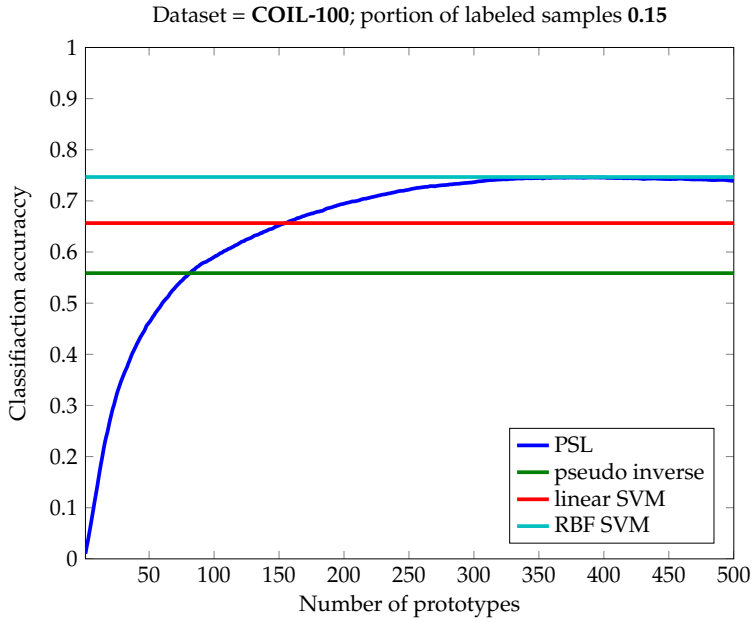
(a) Classification results for the COIL-100 data set for different number of prototypes when 1 % are labeled.



(b) Classification results for the COIL-100 data set for different number of prototypes when 2 % are labeled.



(c) Classification results for the COIL-100 data set for different number of prototypes when 10 % are labeled.



(d) Classification results for the COIL-100 data set for different number of prototypes when 15 % are labeled.

Figure A.2: *Numerical evaluations of the four classification approaches on the COIL-100 data set for different portions of the training data provided with labels.*

A.2 "Obst" Data Set

The "Obst" or Fruits data collection was assembled in the context of the EU project "MirrorBot" (Wermter et al., 2005). Its goal was to construct a robot, that is capable of identifying and grasping fruits on a table. Hence the data collection comprises images of seven different fruits, i.e., green apples, lemons, oranges, red apples, red plums, tangerines and yellow plums. The fruits were photographed under different angles, varying light conditions and different positions in the picture in a resolution of 346×288 pixels and every object was portrayed 120 times. Analogous to the COIL data set, color histograms with 24 bins for each of the three RGB color channels were created Fay (2007). The concatenation of the resulting histograms renders a feature of dimensionality 72.

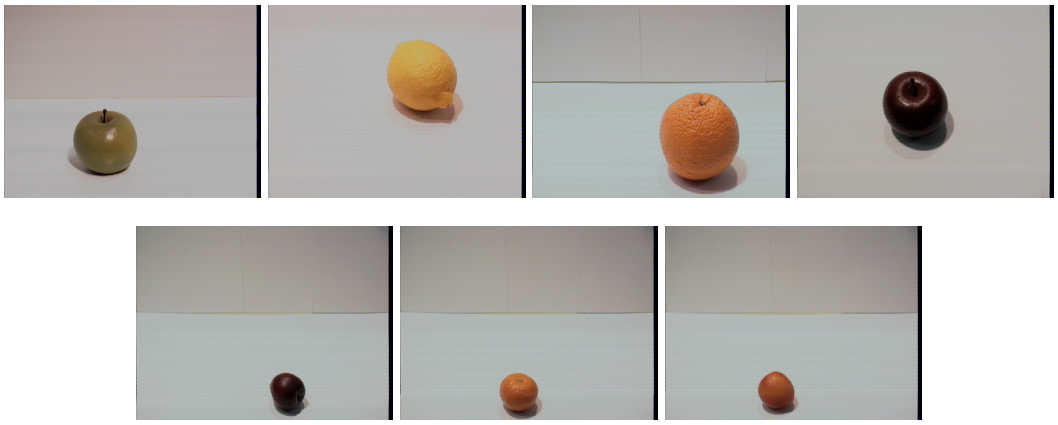
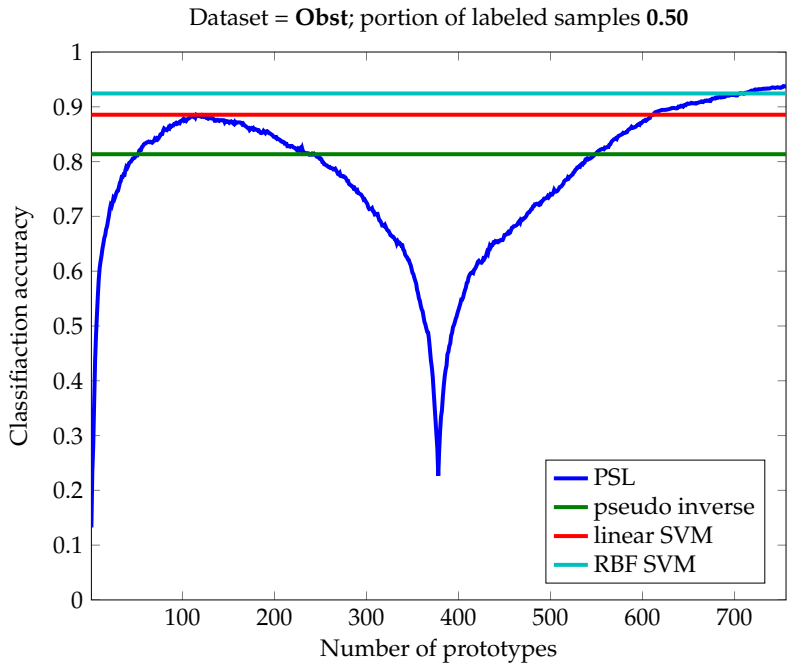
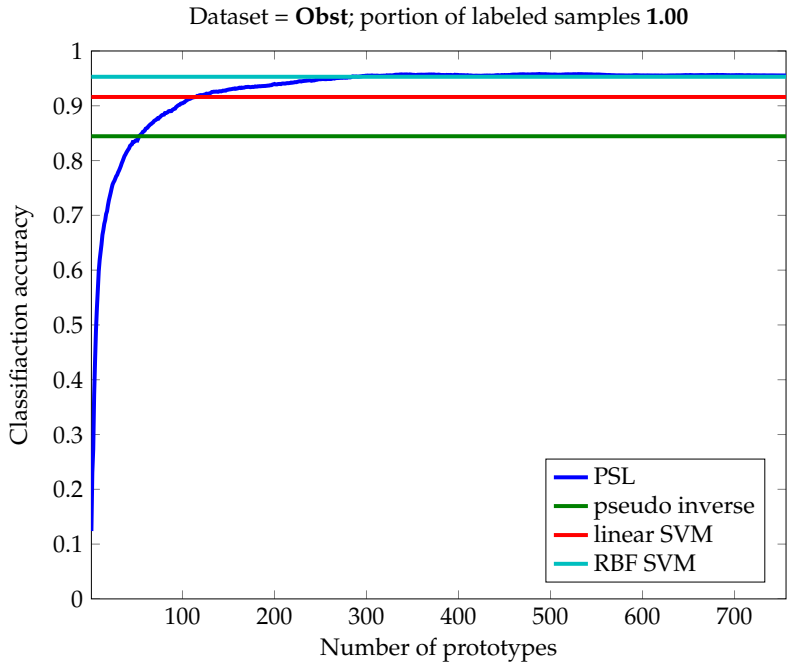


Figure A.3: *Sample images from the "Obst" data collection: green apple, lemon, orange, red apple, red plum, tangerine and yellow plum (from top left to bottom right) (Fay, 2007).*

The results of the numerical evaluation for a different number of prototypes are shown in figures A.4(a) and A.4(b) for portions of labeled data of 0.5 and 1 in the training set. It can be seen in the figure A.4, that the classification task is much easier than the COIL-100 data collection. When only half of the training data is labeled, the partially supervised algorithm performs equal to the linear SVM for smaller numbers of prototypes and equal to the RBF SVM for higher number of components. At around 380 components, the pseudo-inverse classifier, that is situated on the unsupervised procedure fails as mentioned earlier. When a label is provided for all the available training (Figure A.4(b)) the accuracy increases with the number of components and converges around the accuracy of the RBF SVM at 95 %. The pseudo-inverse classifier is outperformed from approximately 100 prototypes and the linear SVM from 150 prototypes on.



(a) Classification results for the "Obst" data set for different number of prototypes when 50 % are labeled.



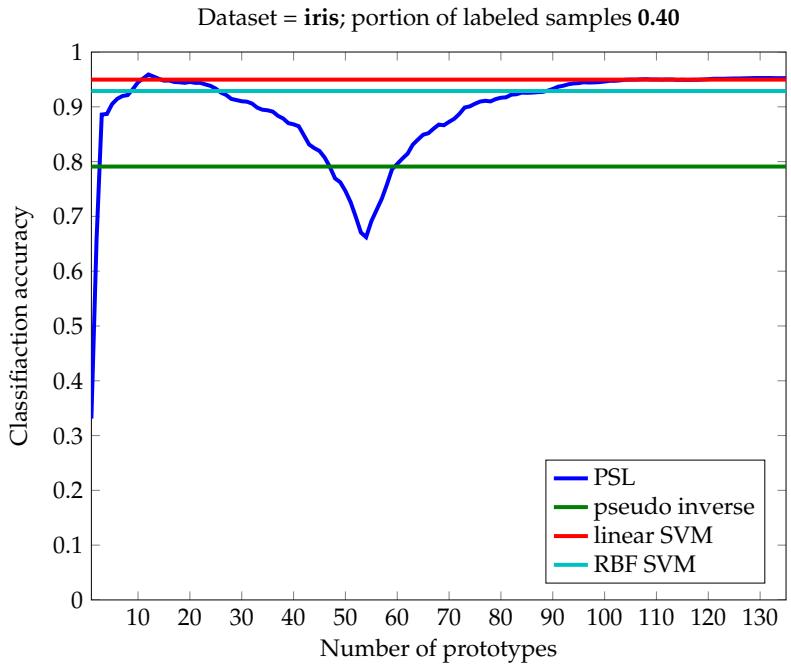
(b) Classification results for the "Obst" data set for different number of prototypes when 100 % are labeled.

Figure A.4: *Numerical evaluations of the four classification approaches on the "Obst" data set for different portions of the training data provided with labels.*

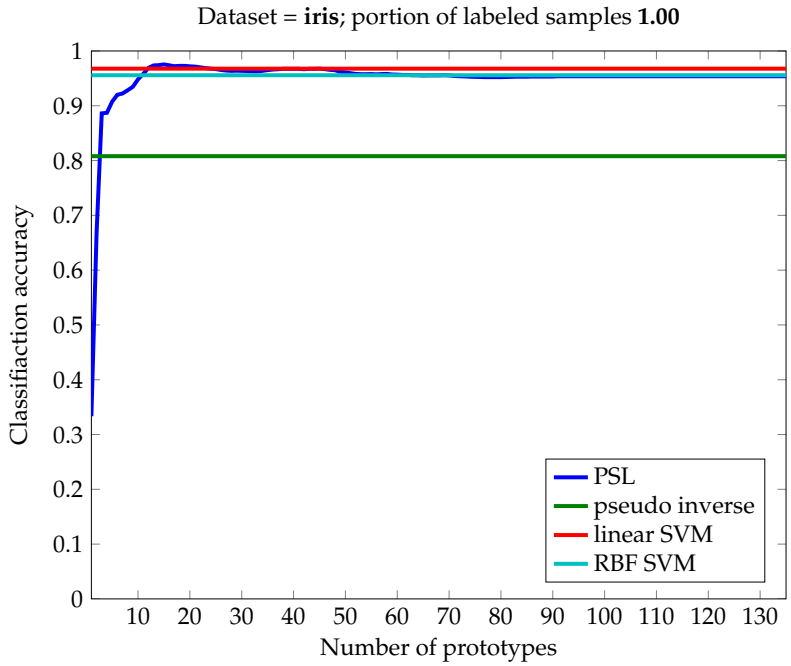
A.3 Iris Data Set

The Iris data collection is a very prominent data set hosted in the UCI repository, describing three different types of iris plants, namely Iris Setosa, Iris Versicolour and Iris Virginica. The features that are available for this data set are the petal and sepal widths and lengths, which provides a four-dimensional feature vector for each sample.

The numerical evaluations for this data set are shown in Figure A.5 for portions of 0.4 and 1 of the available training set labeled. It can be observed that the classification task is comparable easy to solve, rendering high accuracies. In Figure A.5(a) the number of prototypes in the partially supervised algorithm is varied and its accuracy is compared to the other classification approaches for 40 % of the available data labeled. It can be observed, that the accuracy increases with the number of prototypes until the instabilities of the pseudo-inverse occur as mentioned earlier. After that the accuracy increases again and converges at the error rate of the linear SVM. The RBF SVM is outperformed at approximately 90 prototypes. A similar situation is observable when the whole training set is labeled as it can be seen in Figure A.5(b). Again, the linear SVM performs best in this setting but the RBF SVM is very close to it in terms of accuracy. The purely supervised pseudo-inverse approach is only slightly improved using more labeled data. Considering the partially supervised algorithm, the accuracy increases until 15 prototypes are used and converges around the performance of the two SVM approaches.



(a) Classification results for the Iris data set for different number of prototypes when 50 % are labeled.



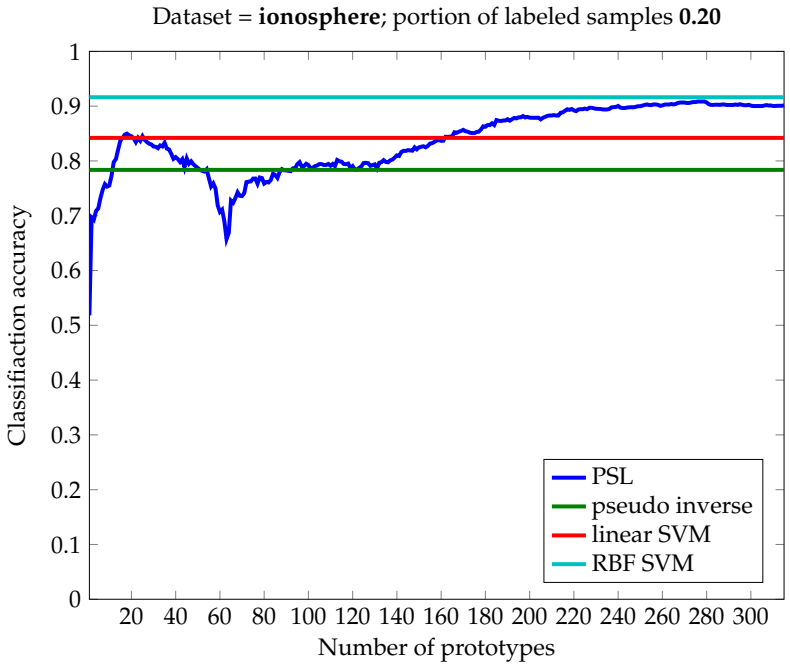
(b) Classification results for the Iris data set for different number of prototypes when 100 % are labeled.

Figure A.5: *Numerical evaluations of the four classification approaches on the Iris data set for different portions of the training data provided with labels.*

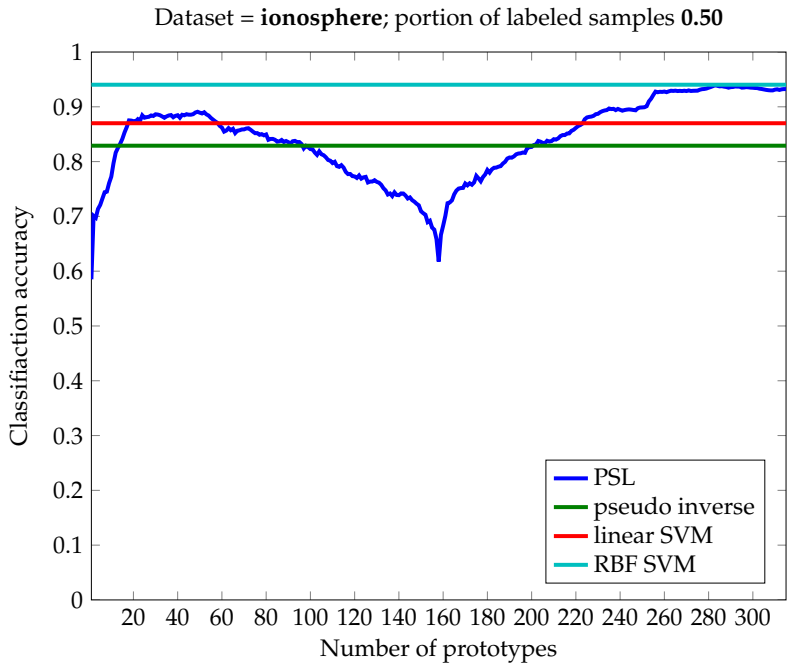
A.4 Ionosphere Data Set

The Ionosphere data set is a popular benchmark data set, that is also located at the UCI machine learning repository. It comprises radar measurements of 16 radar antennas if different frequencies. This data set constructs a two class problem, where the class label signifies whether the radar is reflected by the ionosphere or not. This is labeled as "bad" and "good" in the data collection. The data set provides a 34 dimensional feature vector, that is characterizing the radar pulses.

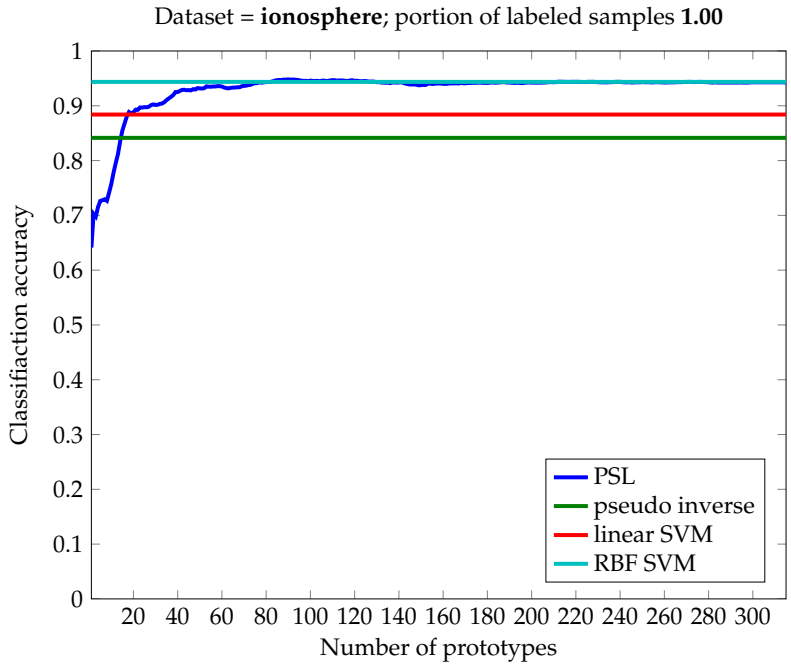
The results of the numerical evaluation for this data set is shown in Figure A.6. Figure A.6(a) shows the results when 20 % of the training set are annotated. It is shown that the partially supervised approach shows an approximately equal accuracy compared to the RBF SVM for a high number of prototypes. Figure A.6(b) shows slightly higher over-all accuracies when the labeled portion of the training data is increased to 50 %. The curve for the partially supervised approach is slightly smoother and the increase of the error is shifted to the right as the number of samples for the pseudo-inverse is increased. In Figure A.6(c), the whole training set of the Ionosphere data set is provided with labels. It can be seen, that the accuracy increases quickly to the level of accuracy of the RBF SVM for this setting. The linear SVM and the supervised pseudo-inverse show still an inferior classification accuracy.



(a) Classification results for the "Obst" data set for different number of prototypes when 20 % are labeled.



(b) Classification results for the Ionosphere data set for different number of prototypes when 50 % are labeled.0.5



(c) Classification results for the Ionosphere data set for different number of prototypes when 100 % are labeled.

Figure A.6: *Numerical evaluations of the four classification approaches on the Ionosphere data set for different portions of the training data provided with labels.*

B Supplemental Results for the Temporal Integration

The numerical evaluation of the information fusion techniques in Section 5.2 focused only on the discrimination of the two experimental sequences ES-2 and ES-5. However, there are further experimental sequences in the over-all sessions, that have been recorded in the experiments as displayed in Figure B.1.

In this chapter of the appendix, the analogous classification experiments, that were previously conducted with respect to ES-2 and ES-5, are executed with the inclusion of the ES-4 and ES-6 for the sake of completeness. These experimental sequences are labeled as "low pleasure, high arousal" and "high pleasure low arousal". The following figures display the classification experiments analogously to the Figure 5.9 in Section 5.2 as plots of the error rate against the size of the time window for the temporal integration. The error rates for the audio and video channels solely and two different fusion approaches are shown. The results for ES-6 versus ES-4, as it is for example conducted by Hrabal et al. (2012), are displayed in Figure B.2, the analogous experiment for the ES-2 versus ES-4 are shown in Figure B.3, and the experiments for ES-6 versus

PAD	+ - +	+ - +	+ + -	- - -	- + -	+ - +
Intro	ES-1	ES-2	ES-3	ES-4	ES-5	ES-6
	approx.	approx.	approx.	approx.	approx.	approx.
1.5 min	3 min	3 min	5 min	3 min	4 min	5 min

time ⟶

Figure B.1: *Experimental procedure for the EmoRec recordings: After a short introduction of the task, the subject is guided through six emotional experimental sequences (ES-1 to ES-6). The respective label in PAD space are given in the top line of the figure in terms of pluses (+) and minuses (-), where a plus signifies a high and minus a low value for the respective dimension. In short the emotional progression aims at "positive" states in the beginning succeeded by a more "negative" state and ending positively again. Reproduction of Figure 3.4*

ES-5 are seen in Figure B.4. Finally, fusion experiments for the unification of the "similar" experimental settings ES-2 and ES-6 and also ES-5 and ES-6 are shown in Figure B.5.

The results of these experiments are less conclusive than the original one for ES-2 versus ES-5 as the error decreases only slowly for larger time window in Figures B.2 and B.4 or not at all as seen in Figures B.3 and B.5. A reason for this result could be that the new classes are not as uniformly embedded in the experimental recording as the originals ES-2 and ES-5. This has effects on the technical side of the experiment as the reliable classification of the experimental sequences relies on relatively large time windows. For example, ES-6 is directly at the end of a recording, which makes difficult to accumulate over larger time windows. Another example is given for the discrimination of ES-5 and ES-6, that are in a direct neighborhood to each other and larger time windows will obviously affect the classification.

A merely psychological problem of experimental sequences, that are temporally near to each other originates from the fact, that the affective states of the subjects do not change very quickly in this experiment. Hence it is possible, that the effect of the presented stimuli did not (yet) have an effect on the state of the subject or the targeted states of the previous experimental sequences may continue to have an effect until the present time step. Unfortunately, these speculations can not be verified as the true internal state of the subject is not inferable.

B.1 Classification of ES-6 versus ES-4

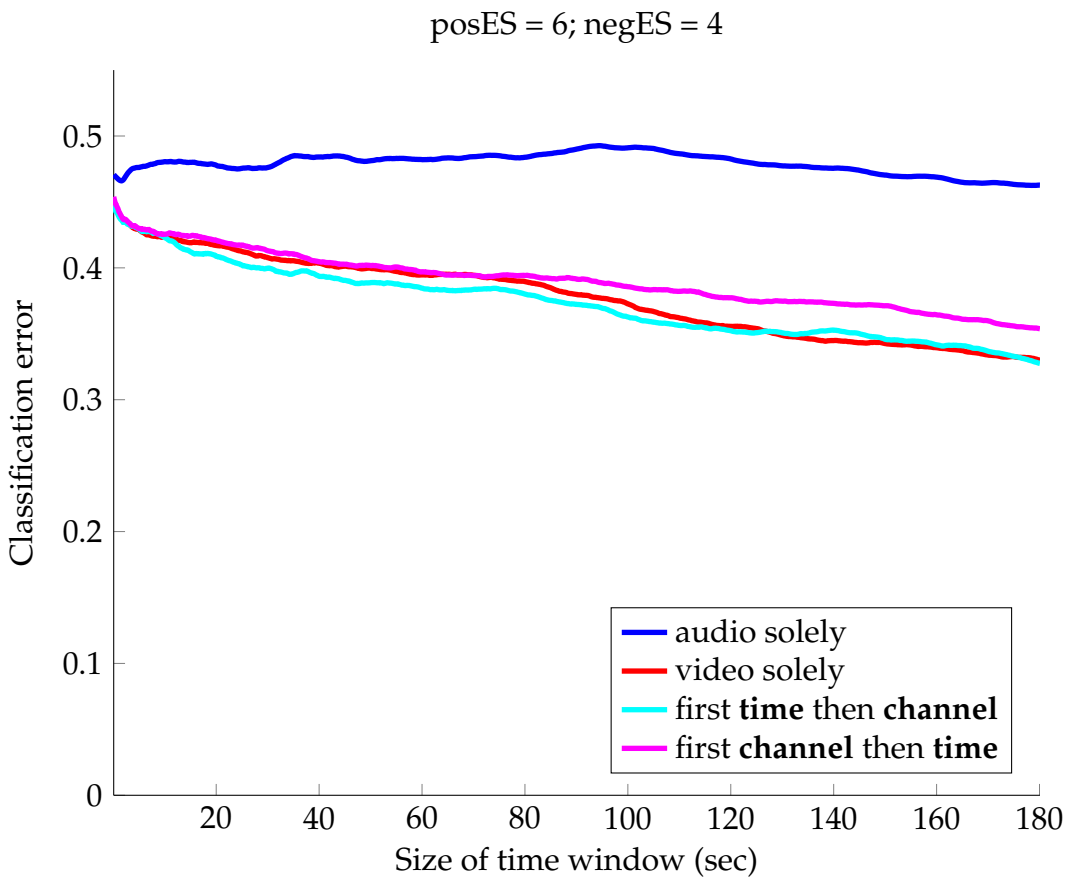


Figure B.2: *Audio-visual classification error versus the time window size for ES-6 versus ES-4 on the EmoRec corpus. The audio data does not provide enough information to discriminate the two experimental sequence and this does not change for larger time windows. The classification of the video material renders slightly better classification results and the combination of the intermediate results in a time window improves the classification with the size of the window. The multi-modal fusion approaches do not further improve the accuracy and follow more or less the curve for the video classification, which is not surprising as the performance for the audio data is poor.*

B.2 Classification of ES-2 versus ES-4

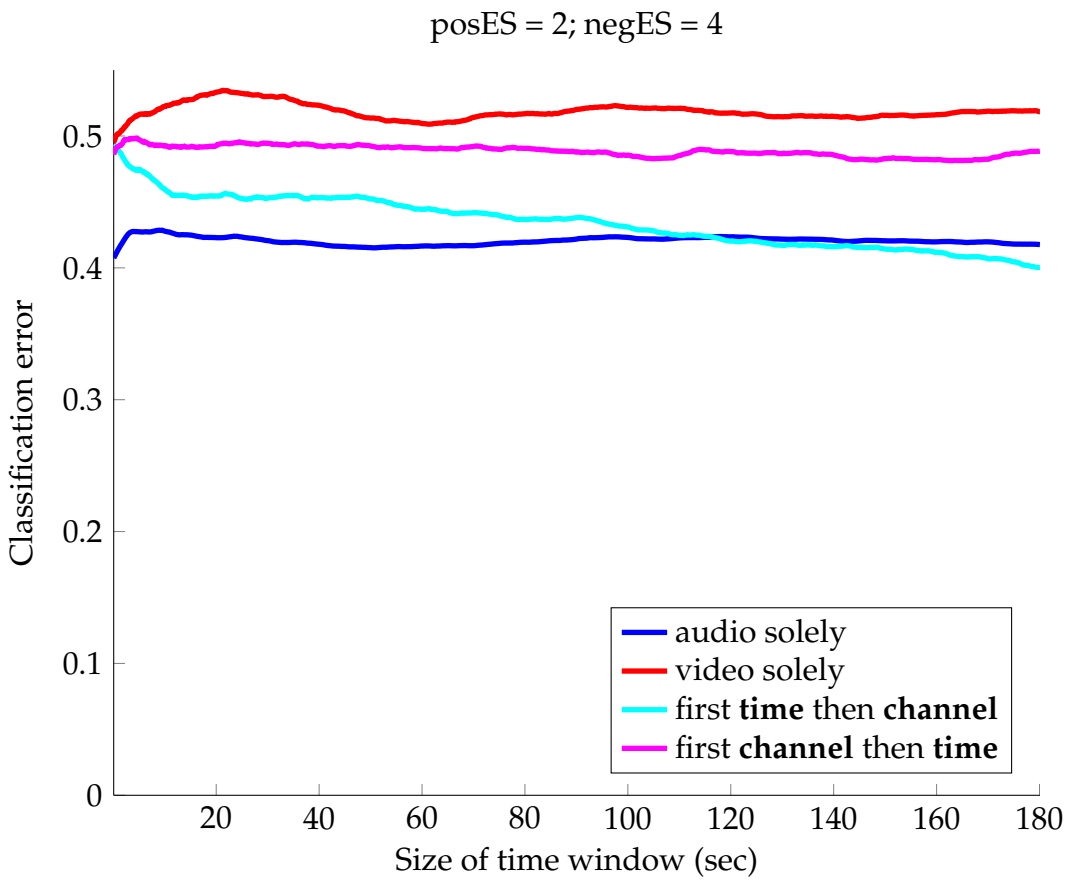


Figure B.3: *Audio-visual classification error versus the time window size for ES-2 versus ES-4 on the EmoRec corpus. The classification accuracy for the video data is rather low and it does unfortunately not improve for larger time windows. The classification of the audio data renders slightly lower errors but it also does not improve for larger time windows. Only one of the competing fusion architectures improves the accuracy of the classification: combining the modalities after the temporal integration (cyan) renders an increasing accuracy for larger time windows, however this does not improve significantly over the audio classification alone. The second fusion approach shows no significant effect for the error of the classification.*

B.3 Classification of ES-6 versus ES-5

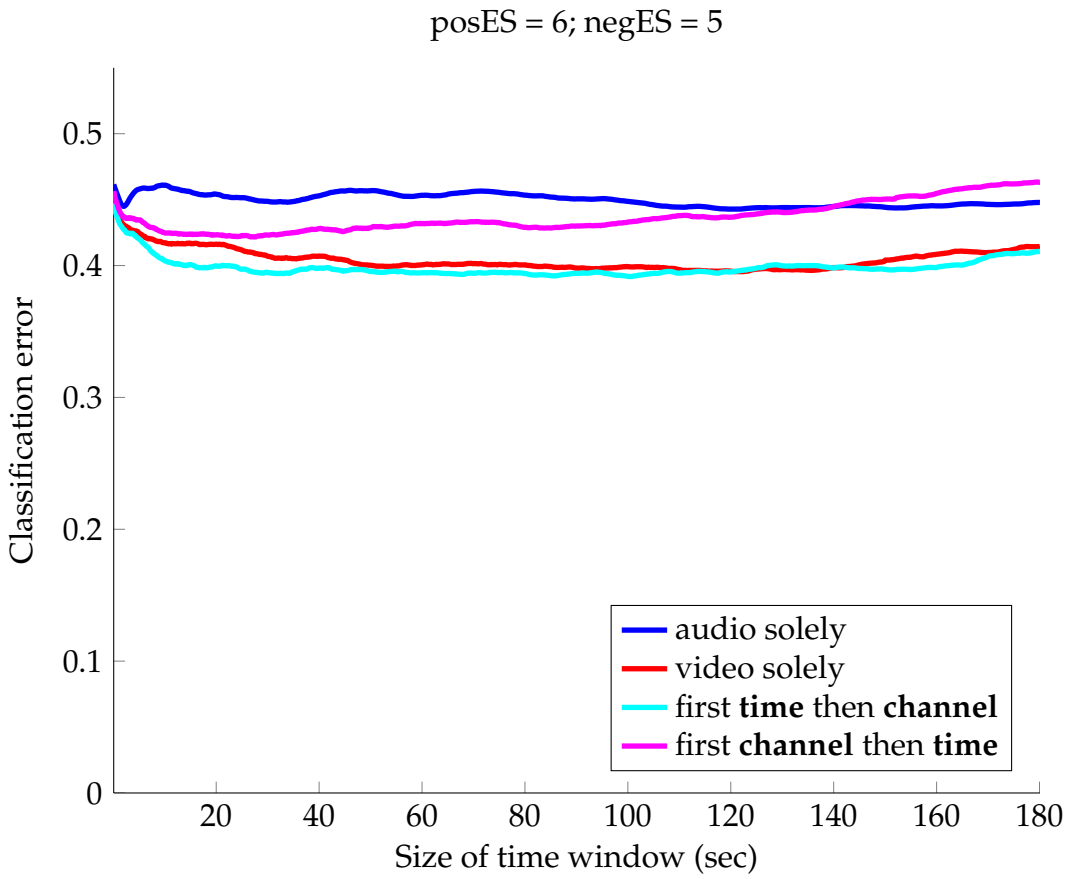


Figure B.4: *Audio-visual classification error versus the time window size for ES-6 versus ES-5 on the EmoRec corpus. The audio classification renders only high errors for the discrimination of these experimental sequences. This does unfortunately not improve for larger time windows. The classification based on the video data of the corpus renders slightly lower error rates, especially for larger time windows. There seems to be a very flat optimal region for time windows as the error increases for very long windows. The fusion approaches do have no or only little effect for the error rates. The variant, where the temporal integration is conducted first follows more or less the curve for the video data. The other fusion architecture, i.e., first the combination of the modalities and the temporal integration afterwards, renders even higher error rates.*

B.4 Classification of ES-2 and ES-6 versus ES-5 and ES-4

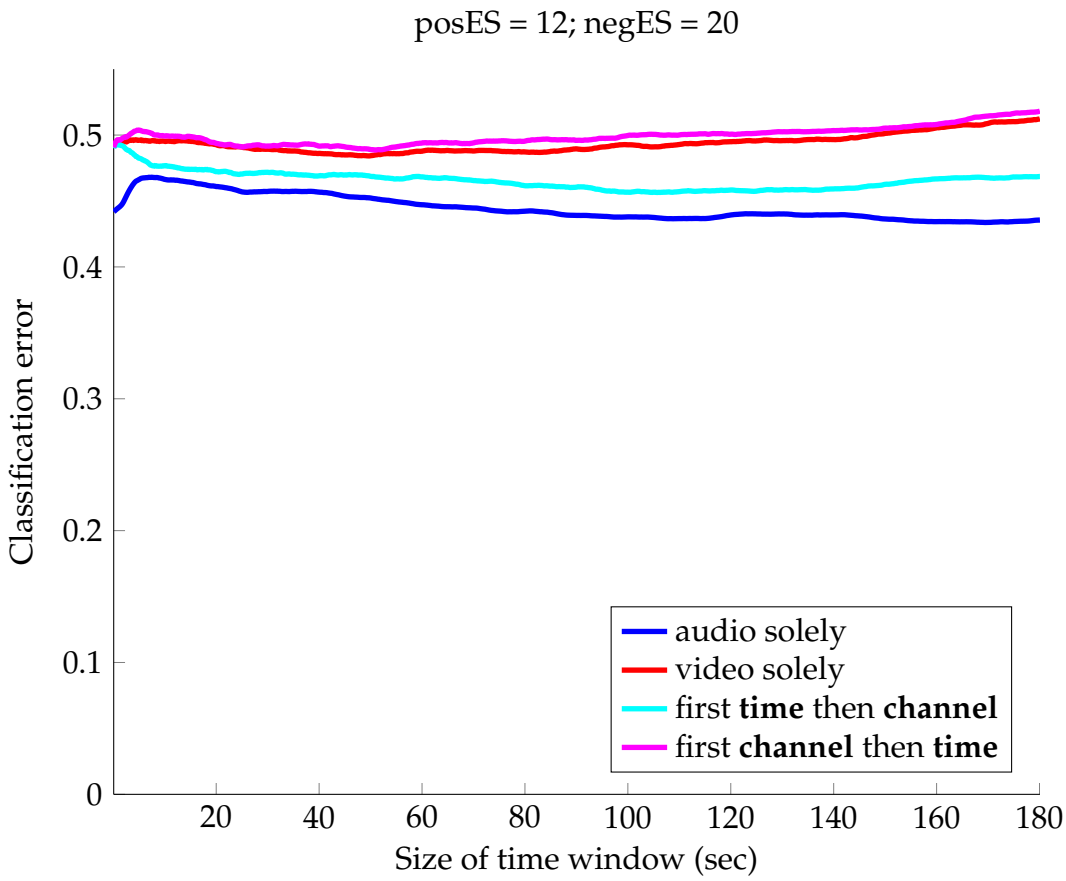


Figure B.5: *Audio-visual classification error versus the time window size for the* unifications *of ES-2 and ES-6 versus ES-5 and ES-4 on the EmoRec corpus. The classification based on the video data renders no meaningful classification results with an error of* 0.5*. This does not improve for larger time windows. The classification of the audio data shows slightly lower error rates, which are merely decreasing a little bit for when increasing the window size. The fusion approaches do hence not improve over the best individual classifier. They follow either the curve for the video data solely or lie in the middle of the curves for the two individual modalities.*

List of Figures

List of Tables

List of Algorithms

Bibliography

Abe, N. and Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 1–9. Morgan Kaufmann Publishers Inc. (Cited on page 53.)

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655. (Cited on page 76.)

Bache, K. and Lichman, M. (2013). UCI machine learning repository. (Cited on pages 35 and 157.)

Baluja, S. (1999). Probabilistic modeling for face orientation discrimination: learning from labeled and unlabeled data. In *Proceedings of the 1998 conference on Advances in Neural Information Processing Systems II (NIPS)*, pages 854–860. (Cited on page 53.)

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1-2):105–139. (Cited on page 50.)

Baumert, J. H., Frey, A. W., and Adt, M. (1995). Analysis of heart rate variability. background, method, and possible use in anesthesia. *Der Anaesthesist*, 44(10):677–686. (Cited on page 71.)

Bennett, K. P. and Campbell, C. (2000). Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.*, 2:1–13. (Cited on page 29.)

Berkhin, P. (2002). Survey Of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA. (Cited on page 14.)

Bicego, M., Murino, V., and Figueiredo, M. (2003). Similarity-based clustering of sequences using hidden markov models. In Perner, P. and Rosenfeld, A.,

editors, *Machine Learning and Data Mining in Pattern Recognition*, volume 2734 of *Lecture Notes in Computer Science*, pages 86–95. Springer Berlin Heidelberg. (Cited on page 108.)

Birbaumer, N. and Schmidt, R. F. (2006). *Biologische Psychologie*. Springer, Heidelberg, german edition. (Cited on pages 83 and 178.)

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. (Cited on pages 13, 18, 21, 22, 23, 24, 25, 26, 28, 30, 31, 33, 34, 35, 40, 105, 176, and 179.)

Bloch, S., Lemeignan, M., and Aguilera-T, N. (1991). Specific respiratory patterns distinguish among human basic emotions. *International Journal of Psychophysiology*, 11(2):141–154. (Cited on page 71.)

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings Conference on Computational Learning Theory*, pages 92–100. (Cited on pages 39, 57, 58, and 187.)

Bocklet, T., Stemmer, G., Zeissler, V., and Nöth, E. (2010). Age and gender recognition based on multiple systems - early vs. late fusion. In *INTERSPEECH*, pages 2830–2833. ISCA. (Cited on page 41.)

Boiten, F. A., Frijda, N. H., and Wientjes, C. J. (1994). Emotions and respiratory patterns: review and critical analysis. *International Journal of Psychophysiology*, 17(2):103–128. (Cited on page 71.)

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159. (Cited on page 105.)

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140. (Cited on pages 50 and 122.)

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. (Cited on pages 34, 50, and 140.)

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA. (Cited on pages 33 and 34.)

Brown, G. and Kuncheva, L. (2010). "Good" and "bad" diversity in majority vote ensembles. In Gayar, N., Kittler, J., and Roli, F., editors, *Multiple Classifier Systems*, volume 5997 of *LNCS*, pages 124–133. Springer. (Cited on page 49.)

Bruns, T. and Praun, N. (2002). *Biofeedback: Ein Handbuch für die therapeutische Praxis*. Vandenhoeck & Ruprecht, Göttingen. (Cited on page 72.)

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. In *Proceedings of Interspeech 2005*, pages 1517–1520. (Cited on page 5.)

Campbell, W., Sturim, D., and Reynolds, D. (2006). Support vector machines using gmm supervectors for speaker verification. *Signal Processing Letters*, 13(5):308–311. (Cited on page 41.)

Cen, L., Yu, Z. L., and Dong, M. H. (2011). Speech emotion recognition system based on l1 regularized linear regression and decision fusion. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 332–340. Springer-Verlag. (Cited on page 117.)

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. In *Tenth International Workshop on Artificial Intelligence and Statisitcs*, pages 57–64. (Cited on page 58.)

Chawla, N. (2005). Data mining for imbalanced datasets: An overview. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer. (Cited on page 105.)

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357. (Cited on page 35.)

Chen, C., Liaw, A., and Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. Technical report, Department of Statistics, University of Berkeley. (Cited on page 140.)

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46. (Cited on page 39.)

Christensen, H. and Fuglsang-Frederiksen, A. (1986). Power spectrum and turns analysis of emg at different voluntary efforts in normal subjects. *Electroencephalography and Clinical Neurophysiology*, 64(6):528–535. (Cited on page 72.)

Christie, I. C. and Friedman, B. H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *International Journal of Psychophysiology*, 51(2):143–153. (Cited on page 66.)

Chumerin, N., Manyakov, N. V., Combaz, A., Suykens, J. A. K., Yazicioglu, R. F., Torfs, T., Merken, P., Neves, H. P., Van Hoof, C., and Van Hulle, M. M. (2009). P300 detection based on feature extraction in on-line brain-computer interface. In *KI'09: Proceedings of the 32nd annual German conference on Advances in artificial intelligence*, pages 339–346. Springer-Verlag. (Cited on page 82.)

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal Artificial Intelligence Research*, 4:129–145. (Cited on page 52.)

Cooper, D. B. and Freeman, J. H. (1970). On the asymptotic improvement in the out-come of supervised learning provided by additional nonsupervised learning. *IEEE Transactions on Computers*, 19(11):1055–1063. (Cited on page 54.)

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. (Cited on page 29.)

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions On Electronic Computers*, 14(3):326–334. (Cited on pages 31 and 93.)

Cozman, F. G. and Cohen, I. (2002). Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, pages 327–331. AAAI Press. (Cited on page 60.)

Cybenko, G. (1992). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 5(4):455–455. (Cited on page 28.)

Darrow, C. W. (1937). The equation of the Galvanic skin reflex curve: I. the dynamics of reaction in relation to excitation-background. *Journal of General Psychology*, 16(2):285–309. (Cited on page 72.)

Darwin, C. (1978). *The expression of emotion in man and animals*. HarperCollins, London, 3rd edition. edited by Paul Ekman. (Cited on page 2.)

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366. (Cited on page 77.)

De Stefano, C., Sansone, C., and Vento, M. (2000). To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 30(1):84–94. (Cited on page 39.)

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366. (Cited on page 16.)

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247. (Cited on page 37.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39. (Cited on page 18.)

Deng, J. and Schuller, B. (2012). Confidence measures in speech emotion recognition based on semi-supervised learning. In *INTERSPEECH*. ISCA. (Cited on page 92.)

Dietrich, C., Schwenker, F., and Palm, G. (2001). Classification of time series utilizing temporal and decision fusion. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 378–387. Springer Berlin Heidelberg. (Cited on pages 89 and 90.)

Dietrich, C. R. (2003). *Temporal Sensorfusion for the Classification of Bioacoustic Time Series*. PhD thesis, Universität Ulm. (Cited on page 109.)

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer. (Cited on pages 42, 43, and 45.)

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. (2008). The sensitive artificial listener an induction technique for generating emotionally coloured conversation. In Devillers, L., Martin, J.-C., Cowie, R., Douglas-Cowie, E., and Batliner, A., editors, *LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4. ELRA. (Cited on page 73.)

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., and Karpouzis, K. (2007). The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, pages 488–500, Berlin, Heidelberg. Springer-Verlag. (Cited on page 61.)

Duin, R. P. W. (2002). The combining classifier: to train or not to train? In *16th International Conference on Pattern Recognition*, volume 2, pages 765–770. IEEE. (Cited on pages 46 and 48.)

Dujardin, K., Derambure, P., Bourriez, J. L., Jacquesson, J. M., and Guieu, J. D. (1993). P300 component of the event-related potentials (ERP) during an attention task: effects of age, stimulus modality and event probability. *International Journal of Psychophysiology*, 14(3):255–267. (Cited on page 82.)

Eckmann, J.-P., Kamphorst, S. O., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973. (Cited on page 70.)

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48:384–392. (Cited on page 78.)

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto. (Cited on pages 2, 79, 80, and 185.)

Esparza, J., Scherer, S., and Schwenker, F. (2012). Studying self- and active-training methods for multi-feature set emotion recognition. In Schwenker, F. and Trentin, E., editors, *Proceedings of the First IAPR TC3 Workshop on Partially Supervised Learning (PSL'11)*, LNCS, pages 19–31. Springer. (Cited on page 92.)

Fang, Z., Guoliang, Z., and Zhanjiang, S. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16(6):582–589. (Cited on page 77.)

Fawcett, T. and Flach, P. A. (2005). A response to Webb and Ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58(1):33–38. (Cited on page 105.)

Fay, R. (2007). *Feature Selection and Information Fusion in Hierarchical Neural Networks for Iterative 3D-Object Recognition*. PhD thesis, Universität Ulm. (Cited on pages 157, 159, 162, and 181.)

Feger, F. and Koprinska, I. (2006). Co-training using RBF nets and different feature splits. In *International Joint Conference on Neural Networks*, pages 1878–1885. IEEE. (Cited on page 57.)

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York. (Cited on page 49.)

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. (2007). The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057. (Cited on pages 3, 75, and 76.)

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *THE COMPUTER JOURNAL*, 41(8):578–588. (Cited on pages 14 and 20.)

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139. (Cited on page 50.)

Ganesalingam, S. (1989). Classification and mixture approaches to clustering via maximum likelihood. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(3):455–466. (Cited on page 54.)

Genolini, C. and Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics*, 25:317–328. (Cited on page 84.)

Giacinto, G. and Roli, F. (2001). Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9—10):699–707. (Cited on page 49.)

Glodek, M., Reuter, S., Schels, M., Dietmayer, K., and Schwenker, F. (2013a). Kalman filter based classifier fusion for affective state recognition. In *Proceedings of the International Workshop on Multiple Classifier Systems (MCS)*, volume 7872 of *Lecture Notes in Computer Science (LNCS)*, pages 85–94. Springer. (Cited on pages 8 and 149.)

Glodek, M., Schels, M., Palm, G., and Schwenker, F. (2012a). Multi-modal fusion based on classifiers using reject options and markov fusion networks. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1084–1088. IEEE. (Cited on pages 39, 88, and 109.)

Glodek, M., Schels, M., Palm, G., and Schwenker, F. (2012b). Multiple classifier combination using reject options and markov fusion networks. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 465–472. ACM. (Cited on pages 8, 39, 109, and 149.)

Glodek, M., Schels, M., and Schwenker, F. (2013b). Ensemble gaussian mixture models for probability density estimation. *Computational Statistics*, 28(1):127–138. (Cited on page 20.)

Glodek, M., Schels, M., Schwenker, F., and Palm, G. (2014). Combination of sequential class distributions from multiple channels using markov fusion networks. *Journal on Multimodal User Interfaces*, pages 1–16. (Cited on page 88.)

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., and Schwenker, F. (2011). Multiple classifier systems for the classification of audio-visual emotional states. In D'Mello, S., Graesser, A., Schuller, B., and Martin, J.-C., editors, *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction (ACII'11) - Part II*, LNCS 6975, pages 359–368. Springer. (Cited on pages 8, 116, and 149.)

Gnjatović, M. and Rösner, D. (2010). Inducing genuine emotions in simulated speech-based human-machine interaction: The NIMITEK corpus. *IEEE Transactions on Affective Computing*, 1(2):132–144. (Cited on page 61.)

Gray, H. M., Ambady, N., Lowenthal, W. T., and Deldin, P. (2004). P300 as an index of attention to self-relevant stimuli. *Journal of Experimental Social Psychology*, 40(2):216 – 224. (Cited on page 82.)

Hady, M. F. A. (2011). *Semi-Supervised Learning with Committees: Exploiting Unlabeled Data Using Ensemble Learning Algorithms*. PhD thesis, Universität Ulm. (Cited on pages 39, 56, and 60.)

Hady, M. F. A., Schels, M., Schwenker, F., and Palm, G. (2010a). Semi-supervised facial expressions annotation using co-training with fast probabilistic tri-class svms. In Diamantaras, K. I., Duch, W., and Iliadis, L. S., editors, *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN'10)*, LNCS 6353, pages 70–75. Springer. (Cited on page 151.)

Hady, M. F. A. and Schwenker, F. (2010). Combining committee-based semi-supervised learning and active learning. *Journal of Computer Science and Technology*, 25(4):681–698. (Cited on page 58.)

Hady, M. F. A., Schwenker, F., and Palm, G. (2010b). Semi-supervised learning for tree-structured ensembles of RBF networks with co-training. *Neural Networks*, 23(4):497–509. (Cited on page 57.)

Hajek, P. (2010). Fuzzy logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Stanford, fall 2010 edition. (Cited on page 40.)

Hansen, L., Liisberg, C., and Salamon, P. (1997). The error-reject tradeoff. *Open Systems & Information Dynamics*, 4:159–184. (Cited on page 38.)

Hartigan, J. (1975). *Clustering algorithms*. Wiley. (Cited on page 16.)

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752. (Cited on page 78.)

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1992). Rasta-plp speech analysis technique. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, volume 1, pages 121–124. (Cited on page 78.)

Hild, K. E., Kurimo, M., and Calhoun, V. D. (2010). The sixth annual mlsp competition. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 107 –111. (Cited on page 86.)

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844. (Cited on page 50.)

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780. (Cited on pages 87 and 117.)

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257. (Cited on page 28.)

Hoyle, D. (2011). Accuracy of pseudo-inverse covariance learning — a random matrix theory analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1470–1481. (Cited on page 158.)

Hrabal, D., Rukavina, S., Limbrecht, K., Walter, S., Hrabal, V., Gruss, S., and Traue, H. (2012). Emotion identification and modelling on the basis of paired physiological data features for companion systems. In *Proceedings of the 7th Vienna International Conference on Mathematical Modeling (MATHMOD 2012)*. (Cited on pages 6, 97, 119, and 169.)

Huang, H. P. and Liu, Y. H. (2002). Fuzzy support vector machine for pattern recognition and data mining. *International Journal of Fuzzy Systems*, 4:826–835. (Cited on page 32.)

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Upper Saddle River, USA, 1 edition. (Cited on page 76.)

Huang, X. and Lee, K.-F. (1993). On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(2):150–157. (Cited on page 106.)

Iscan, Z. (2010). Mlsp competition, 2010: Description of second place method. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 114 –115. (Cited on page 86.)

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80. (Cited on page 88.)

Jain, A., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: a review. *Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37. (Cited on pages 23 and 175.)

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323. (Cited on page 13.)

Japkowicz, N. (2000). Learning from imbalanced data sets: A comparison of various strategies. In *Proc. AAAI Workshop Learning from Imbalanced Data Sets*, pages 10–15. (Cited on pages 83, 100, and 136.)

Joachims, T. (2006). Transductive support vector machines. In Chapelle, O., Schölkopf, B., and Zien, A., editors, *Semi-Supervised Learning*, pages 105–117. The MIT Press. (Cited on page 60.)

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32:241–254. (Cited on page 14.)

Johnstone, T. and Scherer, K. R. (1999). The effects of emotions on voice quality. In Ohala, J. J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A. C., editors, *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2029–2032. Linguistic Society of America. (Cited on page 111.)

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214. (Cited on page 51.)

Kächele, M. (2011). Adaptive classifier fusion architectures for multimodal emotion recognition. Master's thesis, Ulm University. (Cited on pages 68, 69, 70, 72, and 178.)

Kamath, C. and Ananthapadmanabhayuyu, T. V. (2007). Modeling QRS complex in dECG. *IEEE Transactions on Biomedical Engineering*, 54(1):156–158. (Cited on page 68.)

Kamel, M. S. and Wanas, N. M. (2003). Data dependence in combining classifiers. In *Proceedings of the 4th international conference on Multiple classifier systems*, LNCS, pages 1–14. Springer. (Cited on page 46.)

Kanade, T., Cohn, J., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE. (Cited on pages 5, 81, 87, and 175.)

Karmakar, C., Khandoker, A., Voss, A., and Palaniswami, M. (2011). Sensitivity of temporal heart rate variability in poincaré plot to changes in parasympathetic nervous system activity. *BioMedical Engineering OnLine*, 10:1–14. (Cited on page 70.)

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 193–196. (Cited on pages 6 and 65.)

Kestler, H. A., Schwenker, F., Hoher, M., and Palm, G. (1995). Adaptive class-specific partitioning as a means of initializing RBF-networks. In *IEEE international conference on Systems, Man and Cybernetics*, volume 1, pages 46–49. (Cited on page 96.)

Kierkels, J. J. M., Soleymani, M., and Pun, T. (2009). Queries and tags in affect-based multimedia retrieval. In *Proceedings of the international conference on Multimedia and Expo*, ICME'09, pages 1436–1439. IEEE. (Cited on page 4.)

Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):2067–2083. (Cited on pages 66, 68, 71, 72, and 97.)

Kim, J. C., Rao, H., and Clements, M. A. (2011). Investigating the use of formant based features for detection of affective dimensions in speech. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 369–377, Berlin, Heidelberg. Springer-Verlag. (Cited on page 117.)

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239. (Cited on pages 45, 46, and 138.)

Kohavi, R. and Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In Saitta, L., editor, *Proc. 13th International Conference on Machine Learning*, pages 275–283. (Cited on page 49.)

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21(1-3):1–6. (Cited on page 21.)

Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., and Schwenker, F. (2013). Fusion of fragmentary classifier decisions for affective state recognition. In Schwenker, F., Scherer, S., and Morency, L.-P., editors, *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, volume 7742 of *LNAI*, pages 116–130. Springer. (Cited on pages 6 and 96.)

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240. (Cited on page 18.)

Krikler, D. M. (1990). The QRS complex. *Annals of the New York Academy of Sciences*, 601(1):24–30. (Cited on page 68.)

Krogh, A. and Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7:231–238. (Cited on page 38.)

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann. (Cited on page 35.)

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, Hoboken, NJ. (Cited on pages 37, 39, 42, 43, 45, 46, 47, and 51.)

Kuncheva, L. (2002a). Switching between selection and fusion in combining classifiers: an experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 32(2):146 –156. (Cited on pages 51, 83, and 136.)

Kuncheva, L. (2002b). A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286. (Cited on page 46.)

Kuncheva, L. (2003a). Combining classifiers: Soft computing solutions. In Pal, S. and Pal, A., editors, *Pattern Recognition: From Classical to Modern Approaches*, chapter 15, pages 427–449. WorldScientific. (Cited on page 50.)

Kuncheva, L., Bezdek, J. C., and Duin, R. P. W. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314. (Cited on page 47.)

Kuncheva, L. I. (2003b). That elusive diversity in classifier ensembles. In Perales, F., Campilho, A., Blanca, N., and Sanfeliu, A., editors, *Pattern Recognition and Image Analysis*, volume 2652 of *Lecture Notes in Computer Science*, pages 1126–1138. Springer. (Cited on page 49.)

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207. (Cited on pages 48, 49, and 132.)

Labbé, B., Xilan, T., and Rakotomamonjy, A. (2010). Mlsp competition, 2010: Description of third place method. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 116 –117. (Cited on page 86.)

Landgrebe, T. C. W. and Duin, R. P. W. (2008). Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):810–822. (Cited on page 105.)

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273. (Cited on page 4.)

Leiva, J. and Martens, S. (2010). Mlsp competition, 2010: Description of first place method. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 112 –113. (Cited on page 86.)

Lichtenauer, J., Valstar, M., Shen, J., and Pantic, M. (2009). Cost-effective solution to synchronized audio-visual capture using multiple sensors. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '09, pages 324–329. IEEE Computer Society. (Cited on page 75.)

Lin, C. F. and D., W. S. (2002). Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 13:464–471. (Cited on page 32.)

Ling, C. X., Du, J., and Zhou, Z.-H. (2009). When does co-training work in real data? In *Proceedings of the Conference on Advances in Knowledge Discovery and Data Mining*, pages 596–603. (Cited on page 57.)

Lisetti, C. L. and Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 2004(4):1672–1687. (Cited on page 4.)

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 298–305. IEEE. (Cited on pages 78 and 79.)

Littlewort, G. C., Bartlett, M. S., and Lee, K. (2009). Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803. (Cited on pages 81 and 178.)

Malik, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065. (Cited on page 68.)

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface'05 audio-visual emotion database. In *International Conference on Data Engineering Workshops*, page 8. IEEE. (Cited on page 5.)

Martinetz, T. and Schulten, K. (1991). A NNeural-GasNNetwork Learns Topologies. *Artificial Neural Networks*, I:397–402. (Cited on page 21.)

McKeown, G. (2011). Chatting with a virtual agent: The SEMAINE project character spike. `https://www.youtube.com/watch?v=6KZc6e_EuCg`. (Cited on pages 74 and 178.)

McKeown, G., Valstar, M., Cowie, R., and Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pages 1079–1084. IEEE. (Cited on page 73.)

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292. (Cited on page 3.)

Meudt, S., Glodek, M., Schels, M., and Schwenker, F. (2013). Multi-view video based tracking and audio-visual identification of persons in a human-computer-interaction scenario. In *Proceedings of IEEE International Conference on Cybernetics (CybConf'13)*, pages 116–121. Springer. (Cited on pages 8 and 149.)

Mewett, D., Reynolds, K., and Nazeran, H. (1999). Recurrence plot features of ECG signals. In *Proceedings of the First Joint 21st Annual Conference on Engineering in Medicine and Biology and the 1999 Annual Fall Meeting of the Biomedical Engineering Society BMES/EMBS*, volume 2, page 913. (Cited on page 70.)

Mietus, J. E., Peng, C. K., Goldsmith, R. L., and Goldberger, A. L. (2007). The pNNx files: re-examining a widely used heart rate variability measure. *Heart*, 8:378–380. (Cited on page 68.)

Minsky, M. and Papert, S. (1972). *Perceptrons: An Introduction to Computational Geometry*. Mit Press. (Cited on page 23.)

Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia object image library (coil-100). Technical Report CUCS-006-96, Columbia University. (Cited on pages 157, 159, and 181.)

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 86–93. ACM. (Cited on page 57.)

Nigam, K., McCallum, A., and Mitchell, T. (2006). Semi-supervised text classification using EM. In Chapelle, O., Schölkpf, B., and Zien, A., editors, *Semi-Supervised Learning*, pages 33–55. The MIT Press. (Cited on pages 54 and 187.)

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134. (Cited on page 54.)

Osuna, E., Freund, R., and Girosi, F. (1997). Support vector machines: Training and applications. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA. (Cited on page 101.)

Palm, G. and Glodek, M. (2013). Towards emotion recognition in human computer interaction. In Apolloni, B., Bassis, S., Esposito, A., and Morabito, F. C., editors, *Neural Nets and Surroundings*, volume 19 of *Smart Innovation, Systems and Technologies*, pages 323–336. Springer Berlin Heidelberg. (Cited on page 61.)

Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236. (Cited on page 68.)

Pan, S., Tao, J., and Li, Y. (2011). The casia audio emotion recognition method for audio/visual emotion challenge 2011. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 388–395, Berlin, Heidelberg. Springer-Verlag. (Cited on page 117.)

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of The Cambridge Philosophical Society*, 51:406–413. (Cited on page 24.)

Picard, R., Vyzas, E., and Healey, J. (2001a). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191. (Cited on page 97.)

Picard, R. W., Vyzas, E., and Healey, J. (2001b). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191. (Cited on pages 2 and 66.)

Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., Miller, G., Ritter, W., Ruchkin, D., Rugg, M., and Taylor, M. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, 37(02):127–152. (Cited on page 84.)

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 88(6):2297–2301. (Cited on page 69.)

Pincus, S. M. and Goldberger, A. L. (1994). Physiological time-series analysis: what does regularity quantify? *American Journal of Physiology*, 266(4 Pt 2):H1643–H1656. (Cited on page 69.)

Platt, J. (1999a). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods*, pages 185–208. MIT press. (Cited on pages 32 and 100.)

Platt, J. (1999b). Probabilistic outputs for SVMs and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, pages 61–74. (Cited on pages 32, 38, 40, and 100.)

Poh, N., Heusch, G., and Kittler, J. (2007). On combination of face authentication experts by a mixture of quality dependent fusion classifiers. In *Proceedings of the 7th international conference on Multiple classifier systems*, LNCS, pages 344–356. Springer-Verlag. (Cited on page 38.)

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151. (Cited on page 28.)

Qiu, G. (2002). Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition*, 35(8):1675–1686. (Cited on page 97.)

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106. (Cited on page 34.)

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers. (Cited on page 34.)

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. (Cited on page 88.)

Rabiner, L. R. and B.-H., J. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Eaglewood Cliffs, NJ. (Cited on page 109.)

Raileanu, L. E. and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93. (Cited on page 33.)

Ramirez, G. A., Baltrušcaitis, T., and Morency, L.-P. (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 396–406, Berlin, Heidelberg. Springer-Verlag. (Cited on page 116.)

Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology*, 278(6):H2039–H2049. (Cited on page 69.)

Riedmiller, M. (1994). RPROP - Description and Implementation Details. Technical report, Technische Universität Karlsruhe. (Cited on page 28.)

Robinson, D. W. and Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166–181. (Cited on page 78.)

Ros, F., Pintore, M., Deman, A., and Chrétien, J. R. (2007). Automatical initialization of RBF neural networks. *Chemometrics and Intelligent Laboratory Systems*, 87(1):26–32. (Cited on page 96.)

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *Proceedings of the IEEE Workshop on Application of Computer Vision*, pages 29–36. (Cited on page 54.)

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington, D. C. (Cited on page 24.)

Rudnicki, M. and Strumiłło, P. (2007). A real-time adaptive wavelet transform-based QRS complex detector. In *Proceedings of the 8th International Conference on Adaptive and Natural Computing Algorithms*, pages 281–289. (Cited on page 68.)

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel distributed processing: explorations in the microstructure of cognition*, volume 1, pages 318–362. MIT Press. (Cited on page 28.)

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178. (Cited on page 62.)

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172. (Cited on page 3.)

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805–819. (Cited on page 3.)

Ruta, D. and Gabrys, B. (2005). Classifier selection for majority voting. *Information Fusion*, 6(1):63 – 81. (Cited on page 48.)

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012). Last minute: a multimodal corpus of speech-based user-companion interactions. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC'12)*, pages 2559–2566. European Language Resources Association (ELRA). (Cited on pages 6 and 61.)

Sansone, C., Kittler, J., and Roli, F., editors (2011). *Multiple Classifier Systems*, volume 6713 of *LNCS*. Springer. (Cited on page 42.)

Sayedelahl, A., Fewzee, P., Kamel, M. S., and Karray, F. (2011). Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, pages 407–414. Springer-Verlag. (Cited on page 117.)

Sayers, B. (1973). Analysis of heart rate variability. *Ergonomics*, 16(1):17–32. (Cited on page 68.)

Schäfer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764. (Cited on page 158.)

Schels, M., Glodek, M., Kächele, M., and Schwenker, F. (2014a). Audio-visual classification of user states in human-computer interaction. *Pattern Recognition Letters*. Manuscript submitted for publication. (Cited on pages 8 and 149.)

Schels, M., Glodek, M., Meudt, S., Scherer, S., Schmidt, M., Layher, G., Tschechne, S., Brosch, T., Hrabal, D., Walter, S., Palm, G., Neumann, H., Traue, H., and Schwenker, F. (2013a). Multi-modal classifier-fusion for the recognition of emotions. In Rojc, M. and Campbell, N., editors, *Coverbal synchrony in Human-Machine Interaction*, chapter 4, pages 73–98. CRC Press. (Cited on pages 3, 8, 149, 151, and 175.)

Schels, M., Glodek, M., Meudt, S., Schmidt, M., Hrabal, D., Böck, R., Walter, S., and Schwenker, F. (2012a). Multi-modal classifier-fusion for the classification of emotional states in woz scenarios. In *Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]*, Advances in Human Factors and Ergonomics Series, pages 5337–5346. CRC Press. (Cited on pages 6, 8, 39, 92, 108, 149, and 151.)

Schels, M., Glodek, M., Palm, G., and Schwenker, F. (2013b). Revisiting AVEC 2011 — an information fusion architecture. In Esposito, A., Squartini, S., and Palm, G., editors, *Computational Intelligence in Emotional or Affective Systems*, Smart Innovation, Systems and Technologies, pages 385–393. Springer. (Cited on pages 8, 109, and 149.)

Schels, M., Kächele, M., Glodek, M., Hrabal, D., Walter, S., and Schwenker, F. (2014b). Using unlabeled data to improve classification of emotional states in human computer interaction. *Journal on Multimodal User Interfaces*, 8(1):5–16. (Cited on pages 8 and 151.)

Schels, M., Kächele, M., and Schwenker, F. (2012b). Classification of emotional states in a Woz scenario exploiting labeled and unlabeled bio-physiological

data. In F. Schwenker and E. Trentin, editors, *Partially Supervised Learning (PSL 2011)*, volume 7081 of *LNAI*, pages 138–147, Heidelberg. Springer. (Cited on pages 8 and 151.)

Schels, M., Scherer, S., Glodek, M., Kestler, H. A., Palm, G., and Schwenker, F. (2013c). On the discovery of events in eeg data utilizing information fusion. *Computational Statistics*, 28(1):5–18. (Cited on pages 8 and 152.)

Schels, M., Scherer, S., Glodek, M., Kestler, H. A., Schwenker, F., and Palm, G. (2010). A hybrid information fusion approach to discover events in EEG-data. In Kestler, H. A., Binder, H., Lausen, B., Klenk, H.-P., Schmid, M., and Leisch, F., editors, *Proceedings of Statistical Computing 2010*, Nr. 2010-05. Ulmer Informatik Bericht. (Cited on pages 8 and 152.)

Schels, M., Schillinger, P., and Schwenker, F. (2011). Training of multiple classifier systems utilizing partially labeled sequences. In *Proceedings of the 19th European Symposium on Artificial Neural Networks (ESANN'11)*, pages 71–76. Ciao - i6doc.com. (Cited on pages 8 and 151.)

Schels, M. and Schwenker, F. (2010). A multiple classifier system approach for facial expressions in image sequences utilizing GMM supervectors. In Ercil, A., editor, *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, E4109, pages 4251–4254. IEEE. (Cited on pages 8, 32, 41, and 149.)

Schels, M., Thiel, C., Schwenker, F., and Palm, G. (2009). Classifier fusion applied to facial expression recognition: An experimental comparison. In Dillmann, R., Vernon, D., Nakamura, Y., Schaal, S., Ritter, H., Sagerer, G., and Buss, M., editors, *Human Centered Robot Systems*, volume 6 of *Cognitive Systems Monographs*, pages 121–129. Springer. (Cited on pages 5, 8, 39, 50, and 149.)

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):693–727. (Cited on pages 4 and 175.)

Scherer, S. (2011). *Analyzing the User's State in HCI: from Crisp Emotions to Conversational Dispositions*. PhD thesis, Universität Ulm. (Cited on page 6.)

Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., Tschechne, S., Schwenker, F., Neumann, H., and Palm, G. (2012a). A generic framework for the inference of user states in human computer interaction: How patterns of low level behavioral cues support complex user states in HCI. *Journal on Multimodal User Interfaces*, 6(3-4):117–141. (Cited on pages 8, 35, and 149.)

Scherer, S., Glodek, M., Schwenker, F., Campbell, N., and Palm, G. (2012b). Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems*, 2(1):4:1–4:31. (Cited on page 88.)

Scherer, S., Kane, J., Gobl, C., and Schwenker, F. (2013). Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech & Language*, 27(1):263–287. (Cited on page 32.)

Scherer, S., Schels, M., and Palm, G. (2011). How low level observations can help to reveal the user's state in hci. In D'Mello, S., Graesser, A., Schuller, B., and Martin, J.-C., editors, *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction (ACII'11) - Part II*, LNCS 6975, pages 81–90. Springer. (Cited on pages 8 and 149.)

Scherer, S., Schwenker, F., and Palm, G. (2007). Classifier fusion for emotion recognition from speech. In *3rd IET International Conference on Intelligent Environments 2007 (IE07)*, pages 152–155. IEEE. (Cited on page 5.)

Scherer, S., Trentin, E., Schwenker, F., and Palm, G. (2009). Approaching emotion in human computer interaction. In *Proceedings of the first International Workshop on Spoken Dialog Systems (IWSDS'09)*, pages 156–168. (Cited on page 61.)

Schmidt, M., Schels, M., and Schwenker, F. (2010). A hidden markov model based approach for facial expression recognition in image sequences. In Schwenker, F. and Gayar, N. E., editors, *Proceedings of the 4th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR'10)*, LNAI, pages 149–160. Springer. (Cited on pages 5, 8, and 149.)

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press. (Cited on pages 29, 31, and 101.)

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, 12:1207–1245. (Cited on page 29.)

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., and Wöllmer, M. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165 –183. (Cited on page 73.)

Schröder, M., Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., and Sawey, M. (2000). "FEELTRACE": An instrument for recording perceived

emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 19–24. Textflow. (Cited on page 74.)

Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Interspeech*, pages 805–808. (Cited on page 106.)

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). AVEC 2011 – the first international audio visual emotion challenge. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, ACII'11, pages 415–424, Berlin, Heidelberg. Springer-Verlag. (Cited on pages 6, 35, 73, 75, and 185.)

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, I:119–131. (Cited on page 5.)

Schwenker, F., Dietrich, C., Thiel, C., and Palm, G. (2006). Learning decision fusion mappings for pattern recognition. *IICGST International Journal on Artificial Intelligence and Machine Learning (AIML)*, 6:17–21. (Cited on pages 45 and 47.)

Schwenker, F., Kestler, H., Palm, G., and Hoher, M. (1994). Similarities of LVQ and RBF learning-a survey of learning rules and the application to the classification of signals from high-resolution electrocardiography. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 646–651. IEEE. (Cited on page 97.)

Schwenker, F., Kestler, H. A., and Palm, G. (2001). Three learning phases for radial-basis-function networks. *Neural Networks*, 14:439–458. (Cited on page 97.)

Schwenker, F., Scherer, S., Schmidt, M., Schels, M., and Glodek, M. (2010). Multiple classifier systems for the recognition of human emotions. In Gayar, N. E., Kittler, J., and Roli, F., editors, *Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS'10)*, LNCS 5997, pages 315–324. Springer. (Cited on pages 8 and 149.)

Segalowitz, S. J., Bernstein, D. M., and Lawson, S. (2001). P300 event-related potential decrements in well-functioning university students with mild head injury. *Brain and Cognition*, 45(3):342 – 356. (Cited on page 84.)

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. (Cited on page 52.)

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton. (Cited on page 37.)

Siegert, I., Böck, R., and Wendemuth, A. (2012). The influence of context knowledge for multimodal annotation on natural material. In Böck, R., Bonin, F., Campbell, N., Edlund, J., de Kok, I., Poppe, R., and Traum, D., editors, *Joint Proceedings of the IVA 2012 Workshops*, pages 25–32. Otto von Guericke Universität Magdeburg. (Cited on page 96.)

Simson, M. (1981). Use of signals in the terminal QRS complex to identify patients with ventricular tachycardia after myocardial infarction. *Circulation*, 64(2):235–242. (Cited on page 68.)

Smyth, P. (1997). Clustering sequences with hidden Markov models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 648–654. The MIT Press. (Cited on page 108.)

Soleymani, M., Chanel, G., Kierkels, J., and Pun, T. (2008). Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *International Symposium on Multimedia*, pages 228–235. IEEE. (Cited on page 4.)

Sörnmo, L. and Laguna, P. (2005). *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Elsevier, Burlington, MA, 1 edition. (Cited on page 82.)

Spies, M. (1996). Wahrscheinlichkeit. In Strube, G., Becker, B., Freska, C., Hahn, U., Opwis, K., and Palm, G., editors, *Wörterbuch der Kognitionswissenschaft*. Klett-Cotta. (Cited on page 36.)

Stemmler, G. (1989). The autonomic differentiation of emotions revisited: Convergent and discriminant validation. *Psychophysiology*, 26(6):617–632. (Cited on page 4.)

Stemmler, G., Heldmann, M., Pauls, C. A., and Scherer, T. (2001). Constraints for emotion specificity in fear and anger: The context counts. *Psychophysiology*, 38(2):275–291. (Cited on page 66.)

Strauss, P.-M., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Traue, H., and Weidenbacher, U. (2008). The PIT corpus of german multiparty dialogues. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2442–2445. (Cited on page 6.)

Tax, D. M. J. and Duin, R. P. W. (2002). Using two-class classifiers for multiclass classification. In *16th International Conference on Pattern Recognition*, volume 2, pages 124–127. IEEE. (Cited on page 46.)

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*. Academic Press, Burlington, MA, 4-th edition. (Cited on pages 11, 13, 14, 15, 16, 21, 35, 104, and 179.)

Thiel, C. (2010). *Multiple Classifier Systems Incorporating Uncertainty*. PhD thesis, Universität Ulm. (Cited on pages 31, 32, 35, and 38.)

Thiel, C., Scherer, S., and Schwenker, F. (2007). Fuzzy-input fuzzy-output one-against-all support vector machines. In *Knowledge-Based Intelligent Information and Engineering Systems 2007*, pages 156–165. Springer. (Cited on pages 32 and 101.)

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 3–10. ACM. (Cited on page 106.)

Van Boxtel, A. (2010). Facial emg as a tool for inferring affective states. In *Proceedings of Measuring Behavior*, pages 104–108. (Cited on page 66.)

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York. (Cited on pages 29, 31, and 58.)

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Springer, New York. (Cited on page 29.)

Vapnik, V. N. (2006). Transductive inference and semi-supervised learning. In Chapelle, O., Schölkpf, B., and Zien, A., editors, *Semi-Supervised Learning*, pages 33–55. The MIT Press. (Cited on pages 59 and 177.)

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 1:511–518. (Cited on page 79.)

Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007). Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, pages 139–147, Berlin, Heidelberg. Springer-Verlag. (Cited on page 5.)

Vogt, T. and André, E. (2009). Exploring the benefits of discretization of acoustic features for speech emotion recognition. In *INTERSPEECH*, pages 328–331. ISCA. (Cited on page 97.)

Wagner, J., Andre, E., Lingenfelser, F., and Kim, J. (2011). Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218. (Cited on pages 41 and 42.)

Wagner, J., Vogt, T., and André, E. (2007). A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, pages 114–125, Berlin, Heidelberg. Springer-Verlag. (Cited on page 5.)

Wahlster, W., editor (2006). *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer. (Cited on page 61.)

Walter, S., Kim, J., Hrabal, D., Crawcour, S., Kessler, H., and Traue, H. (2013). Transsituational individual-specific biopsychological classification of emotions. *IEEE Transactions on Systems, Man, and Cybernetics*, 43(4):988–995. (Cited on pages 63, 65, 67, 106, and 177.)

Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H. C., and Schwenker, F. (2011). Multimodal emotion classification in naturalistic user behavior. In *Proceedings of the 14th International Conference on Human-Computer Interaction (HCII'11)*, pages 603–611. Springer. (Cited on pages 6, 8, 35, 67, 92, 96, and 149.)

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244. (Cited on page 16.)

Webb, A. R. (2002). *Statistical Pattern Recognition.* John Wiley and Sons Ltd., Chichester, England. (Cited on pages 16, 21, 22, 25, 27, 28, 30, 31, 32, 33, 35, 103, 104, 105, and 176.)

Wendemuth, A. (2004). *Grundlagen der stochastischen Sprachverarbeitung*. Oldenbourg Wissenschaftsverlag, München. (Cited on pages 76 and 109.)

Wendt, B. and Scheich, H. (2002). The "Magdeburger Prosodie Korpus" - a spoken language corpus for fmri-studies. In *Speech Prosody 2002*. SProSIG. (Cited on page 5.)

Wermter, S., Palm, G., and Elshaw, M., editors (2005). *Biomimetic Neural Learning for Intelligent Robots*. Springer, Heidelberg. (Cited on page 162.)

Wimmer, M., Schuller, B., Arsic, D., Rigoll, G., and Radig, B. (2008). Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proceedings of the 3rd International Conference on Computer Vision Theory and Applications*, pages 145–151. (Cited on page 41.)

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163. (Cited on page 73.)

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259. (Cited on page 48.)

Woods, K., Kegelmeyer, W.P., J., and Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *Transactions on Pattern Analysis and Machine Intelligence*, 19(4):405–410. (Cited on pages 50 and 51.)

Wright, R. A. and Dill, J. C. (1993). Blood pressure responses and incentive appraisals as a function of perceived ability and objective task demand. *Psychophysiology*, 30(2):152–160. (Cited on page 4.)

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163. (Cited on pages 87, 116, and 117.)

Zadeh, L. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100, Supplement 1(0):9–34. (Cited on page 40.)

Zhang, Z., Deng, J., and Schuller, B. (2013). Co-training succeeds in computational paralinguistics. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, pages 8505–8509. IEEE. (Cited on page 92.)

Zhang, Z. and Schuller, B. (2012). Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *INTERSPEECH*. ISCA. (Cited on page 92.)

Zhou, Z.-H. and Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77. (Cited on pages 83, 100, and 136.)

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. (Cited on pages 39 and 52.)

Acknowledgments

Ich bedanke mich bei allen, die direkt oder indirekt zur Entstehung dieser Dissertation beigetragen haben. Insbesondere danke ich:

- Prof. Dr. Günther Palm für die Betreuung und die Begutachtung dieser Dissertation.
- Dr. Friedhelm Schwenker für die Unterstützung bei der Anfertigung dieser Doktorarbeit
- Weiterhin danke ich den verschiedenen Organisationen die die Finanzierung meiner Arbeit geleistet haben, namentlich dem SFB Transregio 62 — *Eine Companion-Technologie für kognitive technische Systeme* und der Carl Zeiss Stiftung.
- Eine große Portion Dankbarkeit verdienen auch die diejenigen die sich zum Korrekturlesen breitschlagen ließen: Michael Glodek (fürs Grobe), Markus Kächele (fürs Feine), Wilhelm Schels und Friedhelm Schwenker.
- Natürlich danke ich auch allen Kollegen im Institut für Neuroinformatik für das Schaffen einer tollen Arbeitsatmosphäre die zur wissenschaftlichen Produktivität einlädt.
- Außerdem danke ich allen Koautoren der im Rahmen dieser Arbeit entstandenen wissenschaftlichten Publikationen für die interessanten und fruchtbaren Kooperationen. Namentlich sind das in zufälliger Reihenfolge: Michael Glodek, Markus Kächele, Stefan Scherer, Steffen Walter, Friedhelm Schwenker, Miriam Schmidt, Georg Layher, Tobias Brosch, Stephan Tschechne, Heiko Neumann, Günther Palm, David Hrabal, Ronald Böck, Kerstin Limbrecht, Harald C. Traue, Sascha Meudt, Hans A. Kestler, Mohamed Farouk Abdel Hady, und Christian Thiel.