A comparative study of physiological feature selection methods for emotion recognition

Andreas De Lille

Supervisors: Prof. dr. ir. Joni Dambre, Dr. ir. Pieter van Mierlo
Counsellor: Ir. Thibault Verhoeven

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Computer Science Engineering

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik Van de Walle
Faculty of Engineering and Architecture
Academic year 2015-2016

UNIVERSITEIT GENT

Contents

Nomenclature

ANOVA Analysis of Variance

BCI Brain Computer Interface

CSP Common Spatial Patterns

DASM Differential Asymmetry

DCAU Differential Caudality

DE Differential Entropy

DEAP Dataset for Emotion Analysis using Physiological Signals

EEG Electroencephalography

ELM Extreme Learning Machine

ERP Event Related Potential

GELM Graph regularized Extreme Learning Machine

GSR Galvanic Skin Response

KNN k-nearest neightbors

LDA Linear Disciminant Analysis

LDS Linear Dynamic System

LR Logistic Regression

MEG magnetoencephalography

MLP Multilayer Perceptron

MRMR Minimal Redundancy Maximal Relevance

OCR Optical Character Recognition

OOB Out of Bag

PCA Principal Component Analysis

PSD Power Spectral Density

RASM Rational Asymmetry

RCAU Rational Caudality

RF Random Forests

SAM self-assessment manikins

SBS Sequential Backward Selection

SFS Sequential Forward Selection

1

Introduction

This chapters introduces the masterthesis. It starts by explaining what a brain computer interface is and how it works. After that, emotion recognition is explained and basic concepts of machine learning are introduced. The last section of this chapter covers the goal of the thesis.

1.1 Brain computer interfaces

A Brain Computer Interface (BCI), creates a direct neural link between the brain and the computer[1], that tries to recognize patterns and based on the extracted information, performs actions. A BCI removes the need for physical actions, i.e. typing or moving a mouse, for the transfer of information. The neural link provided by the BCI is made of two important components. The first component is the extraction component, which extract brain signals from the brain. The second component is the computer that interprets signals and performs actions based on the outcome.

1.1.1 Electroencephalography (EEG)

Different technologies exist to analyse brain activity, the most convenient method is Electroencephalography (EEG), since it is a non-invasive method. Non-invasive methods, in contrast to invasive methods require no surgery; in the case of EEG, they simply measure electrical activity using electrodes placed on the scalp.

Electrical activity in the brain is caused when an incoming signal arrives in a neuron. This triggers some sodium ions to move inside the cell, which in turn, causes a voltage rise[2]. When this increase in voltage reaches a threshold, an action potential is triggered in the form of a wave of electrical discharge that travels to neighbouring neurons. When this reaction occurs simultaneously in a lot of neurons, the change in electrical potential becomes significant, making it visible to the EEG surface electrodes. EEG can thus only capture synchronized activity of many, many neurons, which explains its low spatial resolution capabilities.

Signals originating from the cortex, close to the skull, are most visible, while signals originating deeper in the brain cannot be observed directly. Even for signals originating close to the cortex, EEG is far from precise as the bone between the the cortex and electrodes distorts the signal. Additionally, other artifacts like eye and muscle movement add a lot of noise to the signal,

which explains why EEG signals are very noisy signal from nature. Noise removal techniques are therefor advised. Note that even though EEG data contains a lot of noise and has a low spatial resolution, it still provides significant insight into the electrical activity of the cortex while offering excellent temporal resolution[3].

Note that EEG only records electrical activity of the brain, other methods like magnetoencephalography (MEG) use magnetic fields to measure brain activity. Since MEG is more prone to noise from external magnetic signals, i.e. the earth's magnetic field and electromagnetic communication, a magnetic shielded room is required, making this method very expensive and not mobile. As a result, this method was not explored during this thesis.

EEG measures electrical activity with electrodes that are placed on the scalp. To ensure that experiments are replicable, standards for locations of electrodes have been developed. One of these systems is the 10/20 system, an internationally recognized methods to describe location of scalp electrodes[4]. The numbers 10 and 20 refer to the distances between the electrodes, which are either 10% or 20% of the total front-back or left-right distance of the skull. Each site is identified with a letter that determines the lobe and a number that determines the hemisphere location.

- **F:** Frontal
- **T:** Temporal
- **C:** Central
- **P:** Parietal
- **O:** Occipital

Note that no central lobe exists; the C letter is only used for identification purposes. The letter z indicates that the electrode is placed on the central line. Even numbers are use for the right hemisphere, while odd numbers are used for the left hemisphere. Figure 1.1 shows a picture of a 23 channel 10/20 system. Note that the 10/20 system does not require a fixed number of channels, some experiments may use a different set of channels, but they all follow the same naming convention.

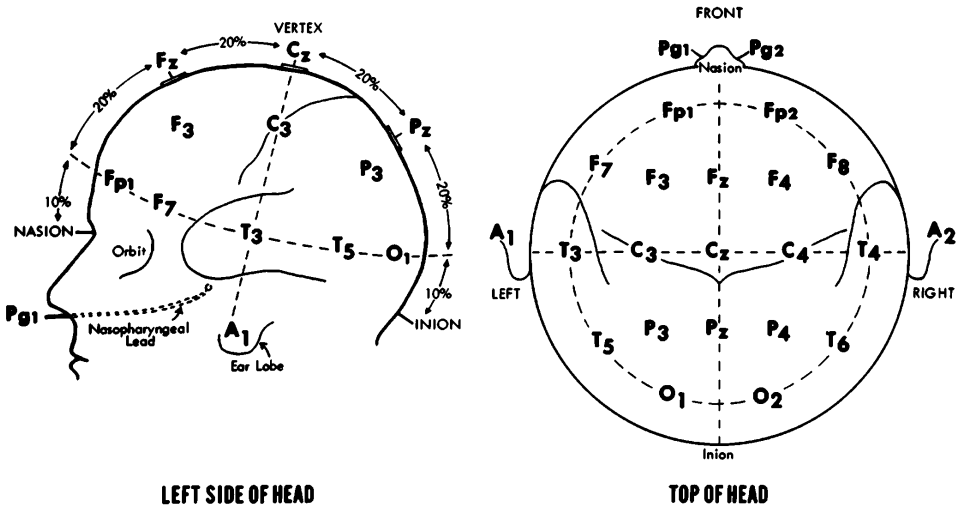

Figure 1.1: The electrode placement of a 23 channel system[5].

Two different types of EEG channels exist, monopolar and dipolar. A monopolar channel records the potential difference of a signal, compared to a neutral electrode, usually connected to an ear lobe of mastoid. A bipolar channel, on the otherhand, is obtained by subtracting two monopolar EEG signals, which improves Sound to Noise ratio (SNR) by removing shared artifacts[6].

In the frequency domain, brain waves are usually split up into different bands[7, 8], each band has a different medical interpretation. These wavebands are:

1. **Alpha:** 8-13Hz, indicate how relaxed and/or inactive the brain is.
2. **Beta:** 13-30HZ, indicate a more active and focused state of mind.
3. **Gamma:** 30-50Hz, relate to simultaneous processing of information from different brain areas.
4. **Delta:** 0-4hz, these waves are generated during dreamless sleep and meditation.
5. **Theta:** 4-8Hz, occurs during dreaming.

Most muscle and eye artifacts have a frequency around 1.2Hz. Artifacts caused by nearby power lines, have a frequency around 50Hz[2]. To remove most of this noise, a bandpass filter is usually applied to filter out frequencies below 4Hz and above 40-45Hz.

1.1.2 Person specific classification versus cross-subject classification

There exists two types of BCI applications. The first type is a person specific BCI, where the BCI interface is calibrated for a single subject. The second type is a general BCI interface that works 'cross-subject', meaning that it should be able to work for different persons. It is much harder to achieve good results for a cross-subject BCI, as EEG data is very personal from nature[9]. While transfer learning has provided good results in imaginary motion recognition in the past, little research has been done for transfer learning in the context of emotion recognition. Person specific classifiers are often used, due to the fact that finding person independent EEG features is still an ongoing topic of research[9].

1.2 Emotion recognition

Psychology makes a clear distinction between physiological behavior and the conscious experience of an emotion, called expression[2]. The expression consists of many parts, including the facial expression, body language and voice concern[10]. Unlike expression, the physiological aspect of an emotion, e.g. heart rate, skin conductance and pupil dilation, is much harder to control. This makes emotion recognition based on physiological signals more robust to social masking[11], the process where an individual masks or hides their emotions to conform to social pressures. To really know one's emotions, it seems, one has to research the physiological aspect of the emotion.

1.2.1 Valence/Arousal classification model for emotion

Before emotions can be recognized, a classification model is needed. A simple way of achieving this is using several discrete emotions, e.g. anger, joy, sad and pleasure. A more convienent model to classify emotions is the bipolar arousal-valence model[2, 12], that places emotions in a two dimensional space. The main advantage of using a continuous multidimensional model, is that all emotions are modelled in its space, even when no particular discrete label can be used to define the current feeling. Figure 1.2 shows the mapping of different emotions for this model.

Arousal indicates how active a person is and ranges from inactive/bored to active/excited. The valence indicates if the emotion is perceived as positive or negative. Even though arousal and valence describe emotions quite well, a third dimension, dominance, can also be added. The third dimension, dominance, indicates how strong the emotional feeling was and ranges from a weak feeling to an empowered, overwhelming feeling. The dominance component can aid to filter out samples of strong feelings, since feelings with low dominance are less likely to show significant effects.

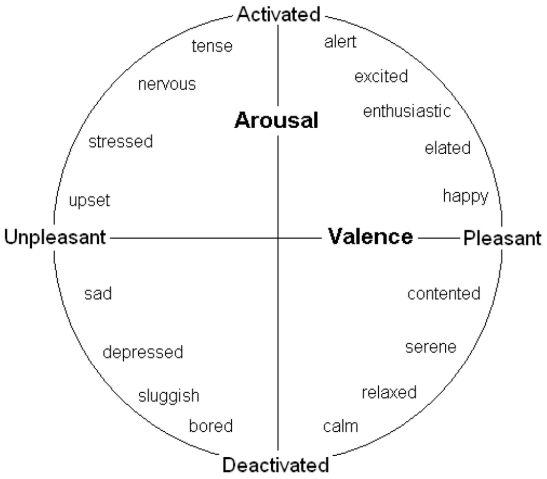

Figure 1.2: The arousal - valence model maps emotions in a two dimensional plane.

1.2.2 Possible applications for emotion recognition

Emotion recognition has many different applications, e.g. as an improvement for the P300 speller or marketing analysis. The P300 speller is a very well-known, academic application of BCI and an active topic of research. It uses EEG data to enable patients with a locked in syndrome to communicate[13]. The basic version uses a six by six grid of characters, each row and column is flashed in a random order while the subject silently counts the number of flashes of a certain character, as shown in figure 1.3. This procedure, where a train of stimuli with some infrequent occurring target stimuli is applied, is called the oddball paradigm[14]. It is known that this technique triggers an increase in the potential difference in the EEG around the parietal lobe. When a potential difference in the brain occurs as a reaction to an event, it is referred to as en event related potential (ERP) . The P300 ERP occurs roughly 300 milliseconds after the stimuli is flashed, hence its name[15]. The presence or absence of the P300 waveform is used by the P300 speller to determine what character the subject was focusing on, which basically allows the subject to spell text.

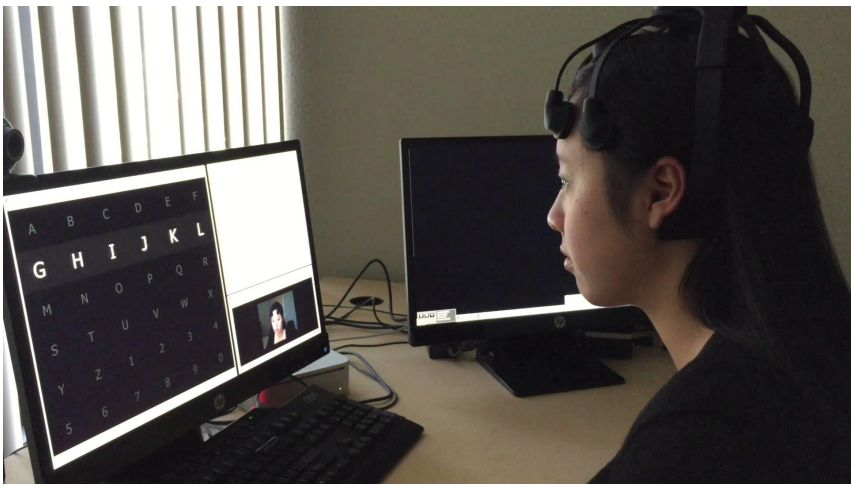

Figure 1.3: Different parts of the P300 speller, found at [16].

Research with visual stimuli on healthy subjects, has shown that emotion has an effect on the auditory P300 wave[17]. Both the P300 peak amplitude and area was highest when viewing neutral pictures and descended further, in decreasing order, for sadness, anger and pleasure. The latency of the P300 ERP speller was shortest or neutrality and in increasing order longer for pleasure, anger and sadness. It is expected that a visually triggered P300 wave, will also be influenced by emotion. Having a good emotion recognition system, might therefore improve the detection of P300 waves. Additionally knowing a subject's emotional state can help detecting when a subject gets frustrated, e.g. because of mistakes he makes.

An improvement in performance is not the only advantage an emotionally aware P300 speller has. Contrary to what subjects might think, the P300 speller is unable to read the mind and know what a person is thinking about[13]. The P300 speller provides no more than a means of communication that the subject can use. Should he chose to ignore the instructions and focus his attention elsewhere, then the recordings become useless. Nevertheless, ethical questions often remain unanswered. Knowing how the subject feels, can help him communicate more humane on one hand, while providing more insight for ethical issues, on the other hand, e.g. "How does the

subject think about the P300 speller recording and analyzing his brain activity?". Information about the subject's emotional state can help answer some of these ethical questions. Integrating the results from this thesis with the P300 speller, is an opportunity for future research.

Another application for emotion recognition is in the field of marketing and customer satisfaction research. Discovering how a persons feels about a product is often tricky. Questionnaires is one way to go, but they might contain a lot of noise. Being able to 'read' the emotion straight from a subject's mind, is expected to give more accurate results as it avoids any form of social masking.

1.3 Machine learning

Machine learning is the missing link between the EEG data the emotion recognition. Machine learning is a very broad domain, as a result, this discussion will be limited to an introduction of the basic machine learning concepts with the focus on the application of machine learning and machine learning techniques.

One possible definition for Machine learning is: "the science of getting computers to act without being explicitly programmed". To do so, machine learning uses pattern recognition to find patterns or structure in the data. A simple example of machine learning is the Optical Character Recognition (OCR), where a computer recognises characters in pictures.

To further explain how machine learning works, have a look at the following example. Suppose one has a price list of houses that are for sale combined with their total area. Logic sense dictates us that a bigger house will have a higher asking price than a smaller house. The total area is a characteristic of the house that helps us in predicting the price. In the context of machine learning, the characteristic 'total area', will be called a feature as the asking price of a house is correlated to the total area.

One possible way of predicting the asking price of a house is machine learning. Machine learning works in several steps, first you need to train your machine learning algorithm with a list of asking prices and the corresponding area of the house. This process is called training or fitting and gives the machine learning component an idea to what the corresponding price is for an area. The outcome might be a coefficient, suppose one square meter is worth 1000 Euro, than the predicted asking price will be 1000 x the total area.

Even though, this might already give some reasonable results, the algorithm will probably not be accurate enough for real life usage. This is due to the fact that the area of the house is only one feature that determines the price. Other features, like the number of bedrooms or the location of the house, were not taken into consideration. Adding additional features, gives more insight into the data, e.g. a house with 5 bedrooms is more expensive than a house with only 3 bedrooms. Having more features is thus likely to improve the performance of the machine learning algorithm.

Machine learning algorithms are responsible for finding the relation between features and the predicted value. There are many of these machine learning algorithms, one way to group these algorithms, is to look at their produced output. In the asking price examples above, the output is a price, which is (more or less) a continuous value. Machine learning problems that require

the output of a continuous value, are called regression problems. In the OCR example from above, where characters are recognized in a picture is a classification problem, as the outcome is a character out of a limit set of different characters.

Another way to group algorithms is on their training data. In the asking price examples above, the training data consists of labelled results; the correct asking price is given for each area. This type of machine learning, where the correct outputs for the train data are given, is referred to as supervised machine learning. The alternative is unsupervised machine learning, which often results in finding groups of similar data points (clustering), without knowing the actual labels. Note that the combination of supervised and unsupervised data, known as semi supervised learning, is also possible. Imagine a dataset with 5000 webpages that need to be grouped into 10 distinct categories, e.g. science, nature, cooking, ... Only 100 of the 5000 pages in the train set are labelled. An approach to solve this problem could be to first cluster the pages in similar groups using unsupervised learning. As soon as a group contains a single labelled page, all pages in the group can be labelled accordingly, as clustering returns groups of similar samples. Semi supervised learning has the advantage that one can also use unlabelled data, which is often easier and cheaper to obtain, unlike labelled data which is usually quite rare; if there was a fast and easy way to label the data then there would not be a need for machine learning.

Machine learning can be used for emotion recognition to find patterns in features extracted from physiological signals. The output of the machine learning algorithm is a prediction of the subject's emotional state. The general process of machine learning is as follows, the process starts with gathering EEG data, from which features are extracted. These features are then fed to a machine learning algorithm, which outputs a prediction.

To recognise emotion in the brain, features need to be extracted from the physiological signals. In this thesis two categories of physiological features are observed. non-EEG features and EEG features. Non-EEG features are physiological signals like heart rate, skin conductivity, respiration rate, ... These features are more convenient to obtain as they do not require one to mount an EEG cap to a subjects scalp.

EEG features, on the other hand, are features extracted from the EEG measurements of the subjects. Unfortunately, the literature does not fully agree on a specific set of features nor does it agree on what channels and/or waveband are most important for emotion recognition. This problem is addressed in this thesis, by ranking a large set of features using different features selection methods. However the literature does agree on certain things; the right hemisphere of a subject is generally speaking, more active during negative emotions than the left hemisphere, which is in turn more active during positive emotions[12, 18, 19]. The features are discussed in more depth in 2.1.

1.3.1 Over and underfitting / high bias and high variance

Over and underfitting is a common problem in many machine learning projects. An algorithm that 'overfits the data' is able to recognise seen sample points very well. However when the algorithm is tested on unseen points, the performance might be a lot lower than expected.

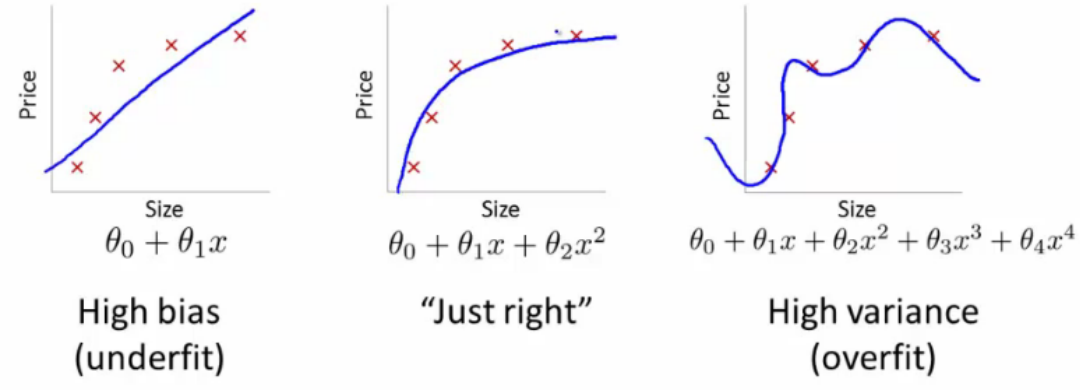

Figure 1.4: Overfitting versus underfitting[20].

Suppose the example in Figure 1.4, where one tries to find a good function to fit the given data points. Looking at the three proposed functions, one can easily see that the middle figure corresponds to the most logical generator function of the red points.

The figure on the left corresponds to an underfit, where the proposed function is not able to capture sufficient detail of the points. The function is not complex enough to approach the generator function. This is known as a high bias problem. A high bias problem has a high training error, as the function is not able to fit the points sufficiently, this is visible in Figure 1.5.

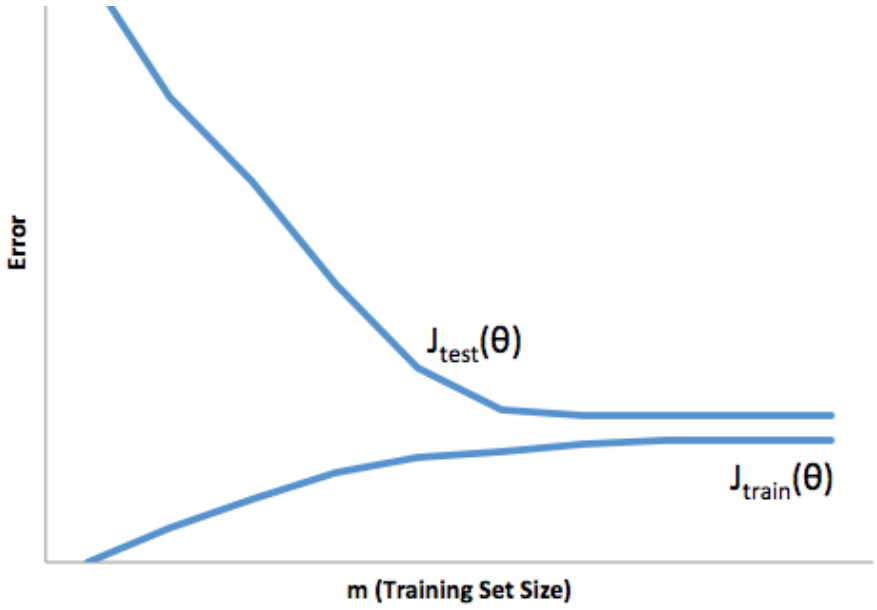

Figure 1.5: A high bias function is not complex enough to approach the generator function closely.

The function on the right corresponds to an overfit; the function fits or 'goes through' each point exactly, but one can see that the behaviour of the hypothesis function in between data points is

not what one would expect. This problem is known as a high variance problem. A high variance problem occurs when the train error is close to zero, so the algorithm fits the points well, but the test error is quite dramatic, which means that the algorithm will perform very badly when it gets new points.

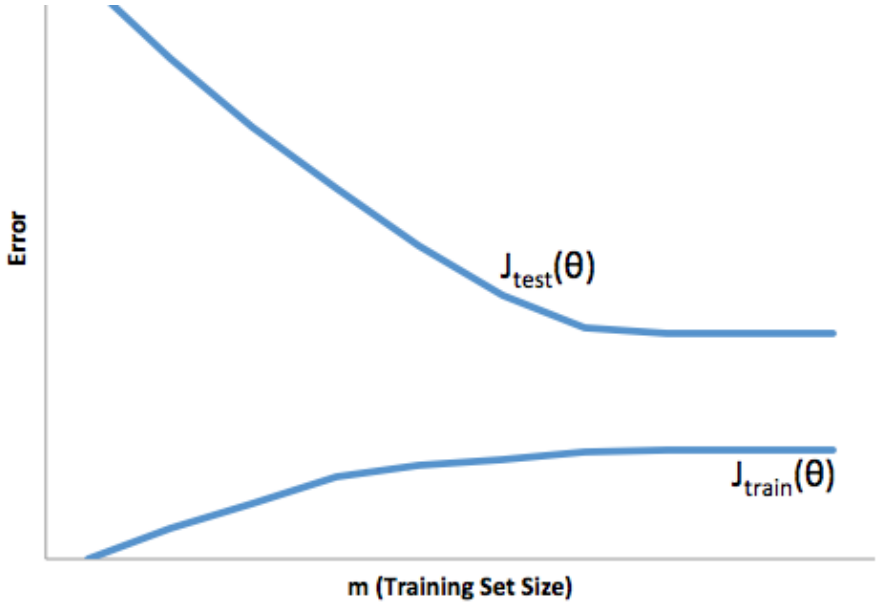

Figure 1.6: A High variance function is too complex and fits the data point too closely.

Another way to explain the bias variance trade-off is by an example. Suppose you have a dart board, as shown in Figure 1.7. Suppose the situation on the top left corner, this corresponds to a world class player that has perfect aim, and very little variation on his precision. The situation on the left bottom corresponds to a player that has very little variation on his precision, but that is consistently aiming too high. He is, in other words, biased to hit higher than needed. The pictures on the right side are different, there the person may or may not have a biased aim, but it is clear that he has a lot of variation in the precision of his aim.

In the context of machine learning, the low bias corresponds to having a hypothesis set that is close to the generator function, which allows you to get quite close. However you still have to pick the right function from that set, which is hard to do if you don't have enough data. If you are not able to take the best solution from the hypothesis set, you have a high variance problem; the solution is right in front of you, but you are not able to reach it precisely.

Figure 1.7: The bias variance explained using the dartboard example found at [21]

1.4 Goal of the thesis

The first goal is finding relevant features for emotion recognition in a person specific setting. This is already quite challenging as there are fuzzy boundaries and individual variation of emotion[22]. To do so, the output of different feature selection methods is compared. In a successful scenario, good features are found that, once given to a machine learning algorithm, can accurately predict the emotions of one person. In this stage some attention will also be spend on comparing non-EEG and EEG features to see which feature set contains the most information.

The second goal is finding features for emotion recognition in a cross-person setting. In this setting the features should generalise well across different persons, thus the algorithm should be able to recognize emotions from unseen persons. The comparison for non-EEG and EEG features will also be done here. Emotion recognition is harder in a cross-subject setting, since physiological signals are very personal[9].

The main problem is that there are already a lot of features known, but, as is often the case with EEG data, training data is expensive and limited. Using a lot of features will thus quickly result in overfitting. Using fewer features, does not only limit the risk of overfitting, it might speed up the algorithm and preparation time. Mounting EEG electrodes is a time consuming activity, using fewer electrodes limits this time.

Another point to note is that even though a simple limited subset of features might solve the overfitting problem, it will likely result in a performance drop as optimal features might have been left out. This problem is even more severe in a cross person setting, when considering that EEG data is person specific[9], features that work good for one person, might not work for another person. Finding a good set of features that work for all persons is a non trivial problem. One solution to this problem could be to use a large pool of possible features from

Table 1.1: Six different papers on emotion recognition, six different feature sets

study	**features used**
[23]	Alpha and beta power
[24]	PSD and asymmetry features
[25]	PSD
[26]	discrete wavelet transform of alpha, beta and gamma band
[2]	alpha/beta ratio, Fpz beta and alpha band power
[19]	PSD, RCAU, DCAU, DASM, RASM, DE

which a limited and person specific set of good features is selected. This allows the machine learning to use good features, while keeping the set of features limited in size. Another solution is to use dimensionality reduction, to project the feature space to a lower dimension, but this has the disadvantage that bad features still have influence and thus might confuse the classifier.

Additionally, a lot of different features are reported in the literature, as you can see in Table 1.1. This thesis tries to overcome this problem, by comparing a large set of features with different feature selection methods to the features reported in literature.

1.4.1 Dataset

One of the most used datasets in the context of emotion recognition is the Dataset for Emotion Analysis using Physiological Signals (DEAP) dataset[9]. This dataset consists of several parts, the first part is a rating of 120 music videos by 14 - 16 persons. Each video is rated for valence, arousal and dominance on a scale ranging from 1 to 9. This part of the dataset is not used during this thesis, because it contains no EEG recordings.

The next part of the dataset is the physiological experiment that contains emotional reactions of 32 subjects. The emotional reactions were triggered using music video excerpts; each subject watched 40 one-minute videos, while several physiological signals were recorded. These physiological signals consist of 32 channel 512Hz EEH and peripheral physiological signals. More concretely, this dataset contains following signals:

There also exists a preprocessed version of the physiological experiment database, where the EEG recordings were downsampled to 128Hz and noise and EOG artifact removal was performed. This dataset is used during this thesis.

Additionally facial video for 22 of the 32 subjects was recorded, so research in facial expressions is also possible with this dataset. These videos are also rated on 4 scales: arousal, valence, dominance and liking. The liking component indicates how much the person liked the video excerpt and should not be confused with the valence component; it inquires information about the participants' tastes, not their feelings, i.e. a person can like a video that triggers angry or sad emotions. However strong correlations were observed[9]. The liking rates are neglected, since they are not part of the emotion space.

For assessment of these scales, the self-assessment manikins (SAM), were used[9]. SAM visualizes the valence, arousal and dominance scales with pictures, each picture corresponds to a discrete

Table 1.2: The available signals in the DEAP dataset

Channel	Name	Category	Channel	Name	Category
1	Fp1	EEG	**21**	F8	EEG
2	AF3	EEG	**22**	FC6	EEG
3	F3	EEG	**23**	FC2	EEG
4	F7	EEG	**24**	Cz	EEG
5	FC5	EEG	**25**	C4	EEG
6	FC1	EEG	**26**	T8	EEG
7	C3	EEG	**27**	CP6	EEG
8	T7	EEG	**28**	CP2	EEG
9	CP5	EEG	**29**	P4	EEG
10	CP1	EEG	**30**	P8	EEG
11	P3	EEG	**31**	PO4	EEG
12	P7	EEG	**32**	O2	EEG
13	PO3	EEG	**33**	hEOG	non-EEG
14	O1	EEG	**34**	vEOG	non-EEG
15	Oz	EEG	**35**	zEMG	non-EEG
16	Pz	EEG	**36**	tEMG	non-EEG
17	Fp2	EEG	**37**	GSR	non-EEG
18	AF4	EEG	**38**	respiration belt	non-EEG
19	Fz	EEG	**39**	plethysmograph	non-EEG
20	F4	EEG	**40**	temperature	non-EEG

value. The user can click anywhere in between the different figures, which makes the scales continuous. All dimension are given by a float between 1 and 9. In this thesis, a preprocessing step scaled and translated these values to ensure they range between 0 and 1, a more convenient interval.

The used SAM figures are shown in Figure 1.8. The first row gives the valence scale, ranging from sad to happy. The second row shows the arousal scale, ranging from bored to excited. The last row represents the different dominance levels. The left figure represents a submissive emotion, while the right figure corresponds with a dominant feeling.

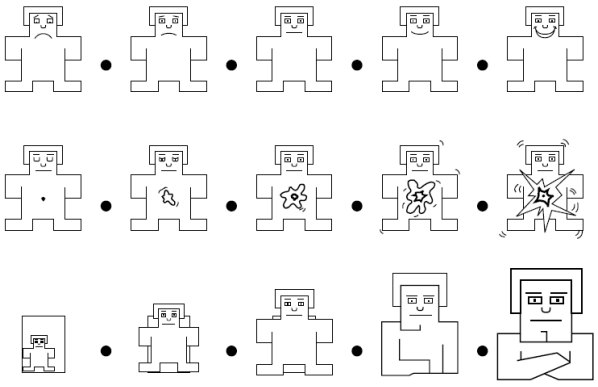

Figure 1.8: The images used for the SAM[9].

2

Methods

This chapter starts by giving an overview of the features found in literature for emotion recognition. Then an overview of some emotion recognition studies is given. The chapter ends by explaining the different feature selection methods.

2.1 Features

Usually, good features are needed to train a machine learning algorithm. In the context of this thesis, good features should be correlated with the subject's emotional state. Two categories of features are observed: EEG features and non-EEG features. Both categories are covered in the following sections.

2.1.1 EEG-features

EEG features are extracted from the electroencephalography measurements from the subject's scalp. The power spectral density (PSD) of a signal gives the distribution of the signal's energy in the frequency domain. By calculating the spectral density for different wavebands of the signal, one can determine how much power of each waveband is in the signal.

Differential entropy (DE) is defined as follows [19]

$$-\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{(x-\mu)^2}{2\sigma^2}) log(\frac{1}{\sqrt{2\pi\sigma^2}}) exp(\frac{(x-\mu)^2}{2\sigma^2}) dx$$

According to [27], the differential entropy of a certain band is equivalent to the logarithmic power spectral density for a fixed length EEG sequence, which simplifies the calculations significantly.

$$DE_{channel} = log(PSD_{channel})$$

The most known feature for valence recognition is the frontal asymmetry of the alpha power[3]. The right hemisphere is generally speaking, more active during negative emotion than the left hemisphere which is in turn more active during positive emotions[12, 18, 19]. The asymmetry can be calculated in different ways. First, one can calculate the differential asymmetry (DASM) , where the left alpha power is subtracted from the right alpha power.

$$DASM = DE_{left} - DE_{right}$$

Another way to measure the asymmetry is by division. The Rational Asymmetry (RASM) does exactly this and is given by:

$$RASM = \frac{DE_{left}}{DE_{right}}$$

With DE_{left} and DE_{right} being the left and right differential entropy respectively. Another reported feature in literature is the caudality, or the asymmetry in fronto-posterior direction[28], meaning the difference in power between the front and the back of the scalp. This can again be calculated in two ways. The first method is the differential Caudality (DCAU) is defined as:

$$DCAU = DE_{front} - DE_{post}$$

Another way to determine the Caudality is the Rational Caudality (RCAU) , which is defined as:

$$RCAU = \frac{DE_{front}}{DE_{post}}$$

With DE_{front} and DE_{post} being the frontal and posterior power respectively. Arousal is usually determined, by looking at the different wavebands[2]. Each waveband has their own medical interpretation, see 1.1.1. Alpha power corresponds to a more relaxed brain, while Beta power corresponds to a more active brain. The alpha / beta ratio therefore seems a good indicator for the arousal state of a person.

The Alpha/ Beta ratio is limited to comparing two wavebands. Other frequently used features are fractions of PSD. Where the fractions of waveband power is determined for a channel, given by:

$$frac_{band,channel} = \frac{power_{band,channel}}{power_{total,channel}}$$

These fractions give insight in the distributions of wavebands at different channel locations.

2.1.2 non-EEG features

The aforementioned EEG features are just one class of physiological features, the DEAP dataset contains several non-EEG, physiological measurements[9]. For each of these measurements the average, standard deviation, variation, median, minimum, maximum and the standard deviation are calculated.

The Galvanic Skin Response uses two electrodes on the middle and index finger of the subjects left hand to measure the skin resistance. It has been reported that the mean value of the GSR is related to the level of arousal[29, 9].

The respiration belt (RSP), indicates the user's respiration rate. Slow respiration is linked to relaxation (low arousal), while fast and irregular respiration patterns corresponds to anger or fear, both emotions with low valence and high arousal[9].

A plethysmograph is a measurement of the volume of blood in the subject's left thumb. This can be use to determine the blood pressure. Blood pressure offers valuable insight into the emotional state of a person as it correlates with emotion; stress is known to increase blood pressure[9].

The heart rate is not directly available in the DEAP dataset, but can be extracted from the plethysmograph, by looking at local minima and maxima[9]. This is clearly visible when looking at the plethysmograph's output, shown in Figure 2.1.

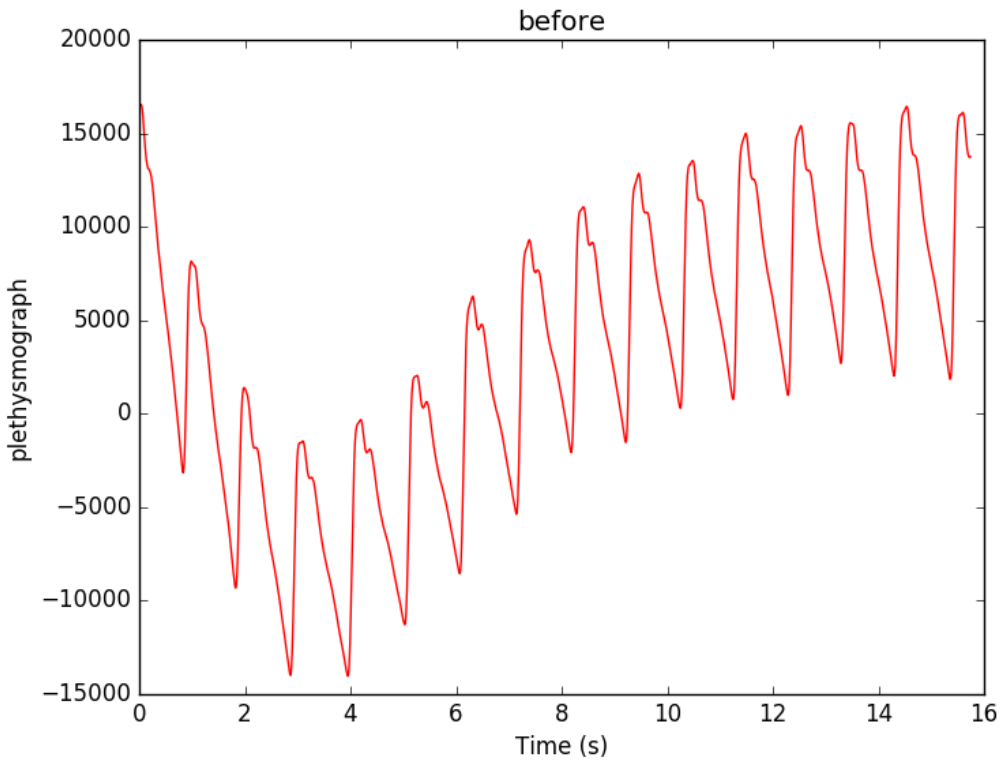

Figure 2.1: The plethysmograph before smoothing.

The heart rate extraction is done in two steps. First the plethysmograph signal is smoothed, by filtering out the high frequency components to avoid noise being selected as a local optima. In the second stage the local optima are located, this is shown in Figure 2.2

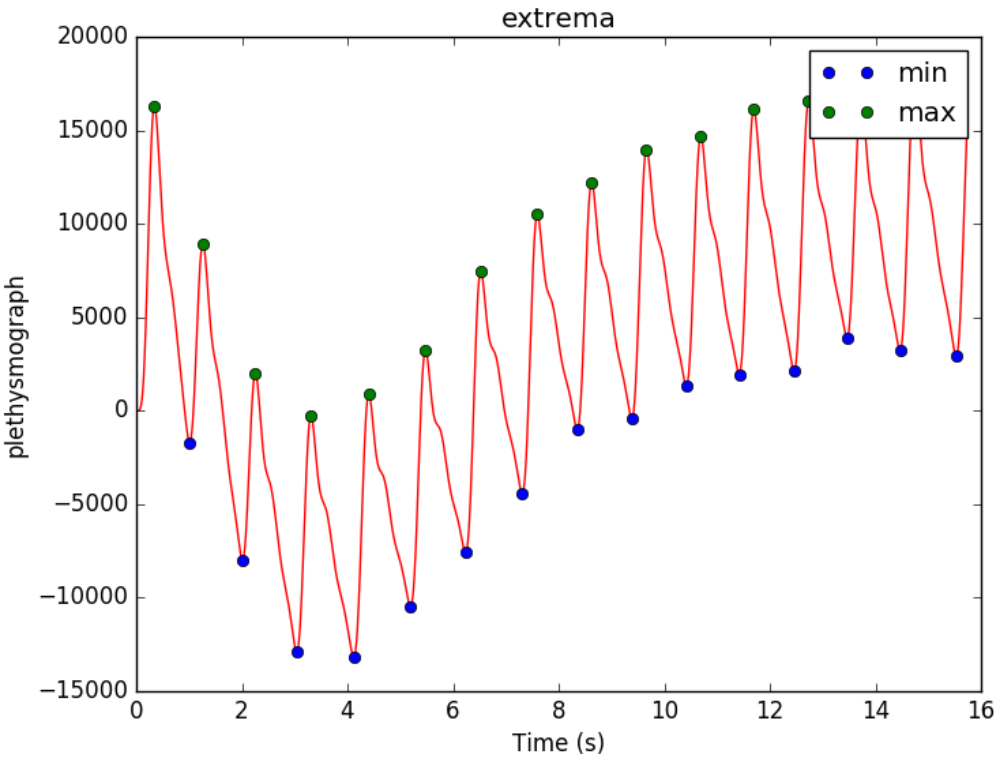

Figure 2.2: The local optima in the plethysmograph.

The combination of a local minima and maxima correspond to a heart beat, therefore the time between two consecutive local minima or maxima correspond to the time between two heart beats, known as the interbeat interval. Getting the average heart rate from the interbeat interval is straight forward. Lastly, the skin temperature of the subject is also available.

2.2 Emotion recognition Studies

This section will give a overview of similar studies and their conclusions. Some of these studies also did some research on cross-subject emotion recognition, but as the research for emotion recognition is still in its infancy[22] and subject independent features are hard to find [9] and as a result, the research is generally speaking, more focussed on person specific emotion recognition.

2.2.1 DEAP method

The first method of emotion recognition is the DEAP method, described in the DEAP paper[9], the paper that introduces the DEAP dataset used in this thesis. The research found that Valence shows the strongest correlations with the EEG signals. Additionally the study found correlations in all frequency bands, with an increase in power for the lower range wavebands for an increase in valence. These effects occur in the occipital regions of the brain, above the visual cortices, which might indicate that the subject is focussing on a pleasurable sound. A central decrease

in beta power was observed together with a occipital and right temporal increase in power for positive emotions. The research conclude that these observed correlations concur with other neurological studies, but that the absolute value of the correlations are seldom bigger than 0.1 for a cross person setting, which indicates that cross person emotion recognition is a non trivial problem. The absolute values of the person specific correlations were around 0.5.

The DEAP paper also presents their own classification method for person specific emotion classification. They start by performing feature selection using the Fisher's linear discriminant for feature selection. The Fisher's linear discriminant is defined as:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2}$$

With μ and σ being the mean and standard deviation of feature f. The Fisher's discriminant was calculated for each feature, before a threshold of 0.3 was applied to filter out irrelevant features. The used classifier was a Naive Bayes classifier, which assumes independence of features. The Naive Bayes classifier is a simple classifier that uses the following equation:

$$G(f_1, ..., f_n) = argmax_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

With F being the set of features and C the classes. $p(F_i = f_i | C = c)$ is estimated by assuming Gaussian distributions of features and modelling these from the training set.

2.2.2 Stable Emotion Recognition over Time

In [19], research is done to find EEG patterns for emotion recognition that are stable over time. EEG patterns are not only subject dependent, they are also dependent on the subjects mood and thus might vary in time. The paper starts by researching different EEG features: PSD, DE, DASM, RASM, DCAU, RCAU, these features are explained in 2.1 and are tested on the DEAP dataset. Afterwards, they develop a new dataset, where subjects have repeated trial sessions with some time in between.

Their machine learning set-up is as follows, first they perform feature extraction of the aforementioned features. Then feature smoothing is done using a Linear Dynamic system (LDS) , that can be expressed by:

$$x_t = z_t + w_t$$
$$z_t = Az_{t-1} + v_t$$

x_t denotes the observed variables or features, while z_t denotes the hidden emotion variables. A is a transformation matrix and w_t is Gaussian noise. The need for a linear dynamic system is supported by the assumption that emotion change gradually over time. The LDS filters out components that are not associated with emotional states.

The list of features at this point is too big and may contain uncorrelated features that might lead to performance degradation of the classifier. Two methods for this are compared, principal component analysis (PCA) and minimal redundancy maximal relevance (MRMR).

PCA uses an orthogonal transformation to create a lower dimensional feature space starting from the original higher dimensional feature space. It does so by minimizing the loss of information, i.e. the principal component should have the largest possible variance.

PCA cannot preserve original domain information like channel and frequency, therefore the paper also uses the MRMR method. MRMR uses mutual information in combination with maximal dependency criterion and minimal redundancy. The algorithm starts by searching features satisfying:

$$maxD(S,c), D = \frac{1}{|S|} \sum_{x_d \in S} I(x_d; c)$$

Where S is the feature subset to select. When two features are highly correlated, the maximal dependency is not likely to change when one of the correlated features is removed. This is expressed by the minimal redundancy condition.

$$minR(S), R = \frac{1}{|S|^2} \sum_{x_{di}, x_{dj} \in S} I(x_{di}, x_{dj})$$

The two conditions are then combined to from the Maximal Relevance Minimum Redundancy, which can be expressed as:

$$max\varphi(D,R), \varphi = D - R$$

Note that incremental search methods exists and are often used in practice. After performing the dimensionality reduction, the samples from the DEAP data set are classified in high / low valence and high/low arousal, giving a total of four classes. All values close to the separation border are removed from the training data, as they might confuse the classifier.

For the classification, three conventional and one newly developed pattern classifiers were compared. k-nearest neighbors (KNN) , logistic regression (LR), Support Vector Machines (SVM) and Graph regularized Extreme Learning Machine (GELM) .

Extreme Learning Machine (ELM) is a single layer feed forward neural network[30]. GELM is based on the idea that similar shapes should have similar properties and obtains better results for face recognition [31] and as the paper concludes, also for emotion classification.

The study found then performed a study on the different features and concluded that DE features are the most suitable EEG features, followed by the asymmetry features (RASM, DASM, DCAU and RCAU). The LDS smoothing was also found to be the better feature smoothing method.

2.2.3 EEG-based emotion recognition in music listening

This study[22] uses EEG features to recognize 4 different discrete emotions (joy, anger, sadness, pleasure) induced by music. They compared four different feature sets on 6 different wavebands: RASM and DASM of 12 channelpairs, raw PSD of the 24 channels and PSD of 30 channels (including 6 midline channels). The compared set of wavebands consists of: alpha, beta, gamma,

delta, tetha and all wavebands. These features were fed to two different classifiers, one Multilayer perceptron (MLP) and an SVM.

Their main results were that the DASM features worked better that the RASM features and even better than using the corresponding 24 PSD features. They also did research to person independent EEG features and found that their accuracy remained consistent. Note that while these results sound promising, they were unfortunately not performed on the DEAP dataset; performance of emotion recognition algorithms is known to vary a lot between datasets[11].

2.3 Comparing selected methods for feature extraction and classification

The paper[11] classifies four distinct emotions (joy, anger, sadness and pleasure) triggered by songs that were selected for each subject. The songs were selected by the subject himself, to help him in triggering memories that trigger the desired emotions. The four emotions are mapped in the valence-arousal model. The used features were typical statistical values of physiological signals (Skin Conductivity (SC), Electrocardiogram (ECG), Electromygraphy (EMG) and Respiration rate (RSP). They compared several techniques: Analysis of Variance (ANOVA) where the best d features were taken. Sequential forward selection (SFS) , where the algorithm starts with an empty feature set and then introduces a new feature in each iteration. Sequential backward selection (SBS) where a feature is removed in each iteration, were also tested. These feature selection methods were also compared to two dimensionality reduction methods: PCA and Fisher projection. The difference between dimensionality reduction and feature selection is that dimensionality reduction methods consider all information in the feature space, whereas feature selection methods take a subset of the information.

The newly formed feature space was then fed to three different classifiers: K-nearest neighbors, Multilayer perceptron and Linear discriminant function. The results indicated that is was easier to classify arousal than valence, which indicate that non-eeg features might be features for arousal classification. SFS in combination with fisher seemed to give the best classification performance, closely followed by LDF and ANOVA, a less computationally intensive method.

The paper also concludes that joy was characterized by a faster heart rate, while sadness was identified by low SC and EMG signals. There was also a higher breathing rate for negative valence emotions. They reported limited similarities for the selected features between subjects.

2.3.1 Research component in this thesis

It is clear that from the aforementioned papers, some research has already been done for emotion recognition. The contribution of this thesis is threefold, first compare a bigger ranger of feature selection methods on a bigger set of physiological features. Second, compare EEG features to non-EEG features to see how much information can be retrieved from the EEG signals compared to the non-EEG signals, which are usually easier to obtain. Third, perform the feature selection methods in a cross-subject setting to see which feature generalise well across subjects. It is also important to note that the feature selection will be performed on the DEAP dataset, so that it can serve as a benchmark. This is important as performance of emotion recognition algorithms based on physiological signals often varies a lot for different datasets[11].

2.4 Feature selection methods

Feature selection is the process of selecting good features from a set of features. The need for this is twofold: first reducing the number of features, is a protection mechanism against overfitting. This is important when a smaller dataset is used. Second, reducing the number of features can speedup the learning process of a learning algorithm as fewer parameters need to be optimized. Additionally, in the context of research, looking at which features are important might give more insight in how emotion is processed by the brain, i.e. knowing what feature are relevant can help neuroscientists understand the working of the brain better. There is also a practical use of feature selection, limiting the physiological signals to fewer channels, can help the setup time. Mounting an EEG cap to a subject is a time consuming process. Using fewer electrodes can make the system more convenient to use as it would save time.

2.4.1 Independent Metrics

These feature selection methods select features based on statistical tests or another machine-learning independent metric.

Pearson Correlation

The Pearson correlation coefficient measures the linear relationship between two variables. The output is a value r, that lies between -1 and 1, corresponding to perfect negative correlation and perfect positive correlation respectively. A correlation value of 0 means that there is no correlation.

More formally[32], the Pearson product-moment coefficient of correlation, r between variables X_i and Y_i of datasets X and Y is defined as:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

with

$$SS_{xy} = \sum_i (X_i - \tilde{X})(Y_i - \tilde{Y})$$

and

$$SS_{xx} = \sum_i (X_i - \tilde{X})^2$$
$$SS_{yy} = \sum_i (y_i - \tilde{Y})^2$$

The Pearson correlation coefficient is fast and simple to calculate, but has some major shortcomings. First off, it can only see linear relation ships and will not see the correlation between a value x and x^2.

In the context of this thesis, whether the correlation is positive or negative is not important; a learning algorithm needs features that have a significant correlation. As a result the absolute value of the r value is reported as this allows for a more convenient comparison of correlations.

Normalized Mutual Information

Mutual information is a more robust option for correlation estimation. The mutual information, I, of two variables X and Y is defined as [33]:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log(\frac{p(x,y)}{(x)p(y)}$$

Using the mutual information directly for feature ranking might be inconvenient because it does not lie in a fixed range. Fortunately, normalized variants of the mutual information score exists. The normalized mutual information, NMI, of variables X and Y is given by:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{(H(X)H(Y))}}$$

With $H(X)$ and $H(Y)$ being the Shannon entropy of variable X and variable Y, defined as:

$$H(X) = \sum_{i \in X} p_i log(\frac{1}{p_i}) = -\sum_i p_i log(p_i)$$
$$H(Y) = \sum_{i \in Y} p_i log(\frac{1}{p_i}) = -\sum_i p_i log(p_i)$$

Distance Correlation

The Pearson correlation coefficient might give a correlation of zero for dependent variables, as shown in Figure 2.3.

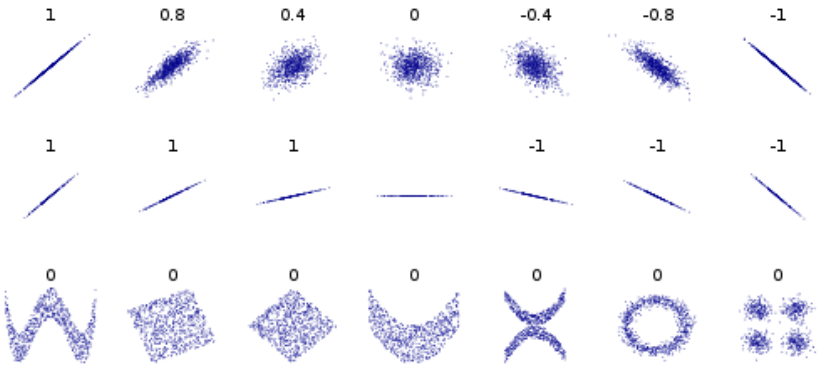

Figure 2.3: Pearson correlation coefficients for different sets of (x,y) points. Note that many coefficients are zero, while there clearly is some correlation. Source: Wikipedia

The distance covariance, sometimes referred as the Brownian covariance, addresses this problem[34]. Its main idea is that a good measurement for dependence is the 'distance' between the join distribution f_{XY} and the product of the marginal distributions f_X and f_Y weighted by a weight function W. This gives the following theoretical function:

$$||f_{XY} - f_X f_y||_W$$

The result is that the distance correlation metric gives very different results, as you can see when comparing the distance correlation outputs in Figure2.4 with the pearson correlation outputs in Figure 2.3.

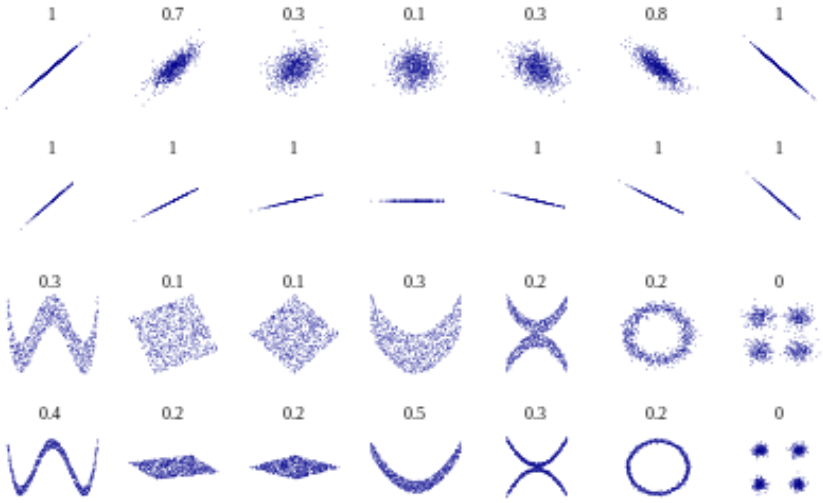

Figure 2.4: Distance correlation coefficients for different sets of (x,y) points. Note the difference with the pearson correlation coefficients in Figure 2.3. Source: Wikipedia

Without going further into the theory, the distance correlation between two variables X and Y, each with n data points can be calculated as follows. First compute all pairwise Euclidean distances for both variables.

$$[D_x]_{j,k} = ||X_j - X_k||$$
$$[D_Y]_{j,k} = ||Y_j - Y_k||$$
$$j, k = 1, 2, ..., n$$

The result is two n by n distance matrices D_x and D_y. Next, both matrices are centered:

$$S_x = C_n D_x C_n$$
$$S_y = C_n D_y C_n$$

Finally, the covariance is computed.

$$\nu^2(X, Y) = \frac{1}{n^2} \sum_l \sum_k [S_x]_{k,l} [S_y]_{k,l}$$

This metric is a covariance metric, which means that it is not normalized. The distance correlation is the normalized version of the distance covariance and is defined as:

$$dCor(X, Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X) dVar(Y)}}$$

With $dCov(X, Y)$ being the aforementioned distance covariance, $dVar(X)$ and $dVar(Y)$ are the distance standard deviations. The distance correlation has the disadvantage that is much slower than mutual information or Pearson correlation, but in return, the distance correlation is able to detect more complex relationships between two variables.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical test to analyse differences between groups. The idea is that the total variance, found in the samples consists of two parts. The first part is the variance within a single group, the second part is the variance between groups.

Suppose you want to test the influence of caffeine on the reaction speed[1]. To do so, you take two groups of each 10 persons. The first group has to drink a large coffee, the second group is the control group that only drinks water. Next the reaction times of all persons in both groups are measured. From these results it is possible to calculate the total variance as well as the variance within each group and the variance between the groups.

If the variance within each group is much larger than the variance in between the groups, one concludes that the groups are similar. The reaction time is thus dependent on the person and not on the caffeine. However should the variance between the groups be much bigger than the variance within each group, than one concludes that the variance in reaction time is caused by the caffeine and not by personal difference.

2.4.2 Machine Learning Methods

These methods select features by applying an arbitrary machine learning technique and looking at the coefficients of the features. The idea is that features with high coefficients have more influence on the end results than features with a lower coefficients and are therefore more important. Again absolute values are used.

Linear Regression

A first method is simple linear regression, where a linear combination of features is searched that produces a good estimate of the output value. Linear regression can achieve good results given that the data doesn't contain a lot of noise and the features are (relatively) independent. When the set of features contains correlated features, the model becomes unstable. As a result, small changes in input data might lead to huge differences in output coefficients. for example assume the 'real output' is given by $Y = X_1 + X_2$ and the dataset contains output in the form of $Y = X_1 + X_2 + \epsilon$ with ϵ being some random noise. Further more assume that X_1 and X_2 are linearly correlated, meaning that $X_1 \approx X_2$. The suspected output of the model should be $Y = X_1 + X_2$, but since noise is added the algorithm might end up with arbitrary combinations of X_1 and X_2, e.g. $Y = -X_1 + 3X_2$. the result will rate one feature much higher than another one, while in reality they are of equal importance. This is due to the noise; by maximizing the performance, the algorithm will minimize the influence of noise on the output, which result in unstable behaviour.

Lasso Regression

Lasso regression uses L1 regularization, that adds a penalty $\alpha \sum_{i=1}^{n} |w_i|$ to the loss function. the result is that the coefficients of weak features are forced to zero, as each non-zero feature adds

[1]This example was based on the following video: https://www.youtube.com/watch?v=ITf4vHhyGpc

to the penalty. This form is regularization is thus quite aggressive, it removes weak features completely. The problem with this is, again, stability; coefficients can vary significantly, even for small changes in training data, when there are correlated features.

Ridge Regression

Ridge regression uses L2 regularization, which add a L2 norm penalty to the loss function, given by $\alpha \sum_{i=1}^{n} w_i^2$. Where the L1 norm forces the coefficients to zero, the L2 regularization forces the coefficients to be spread out more equally. The result is that correlated features tend to get similar coefficients, as this minimizes the loss function, which in turn results in a more stable model.

SVM

A Support vector machine (SVM) is a well known and proven method for machine learning. It has been used in several emotion recognition studies. An SVM works in essence by creating a hyperplane that separates two classes. Shown in Figure 2.5 is a simple line separating the red from the blue balls.

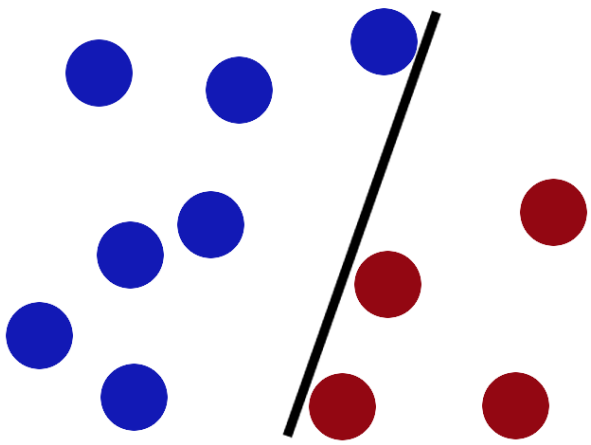

Figure 2.5: One possible separation border.

This is one possible solution, but note that an SVM will always search for a decision boundary that maximizes the boundary between the two classes, shown in Figure 2.6.

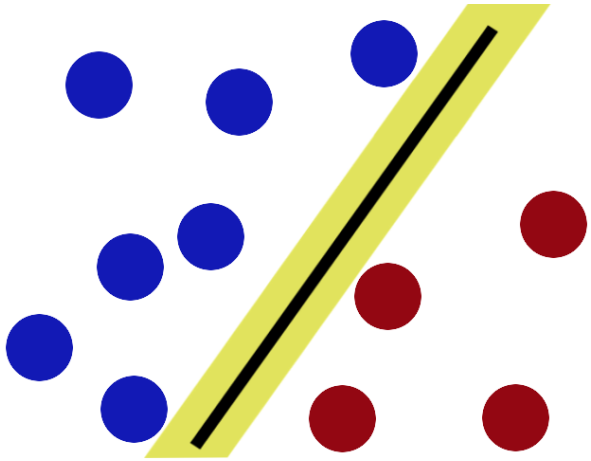

Figure 2.6: A separation with maximal boundary.

This all works well, as the balls are separable using a single straight line. This is not always the case though; shown in Figure 2.7 is a scenario where it is not possible to separate the red balls from the blue ones using a single straight line.

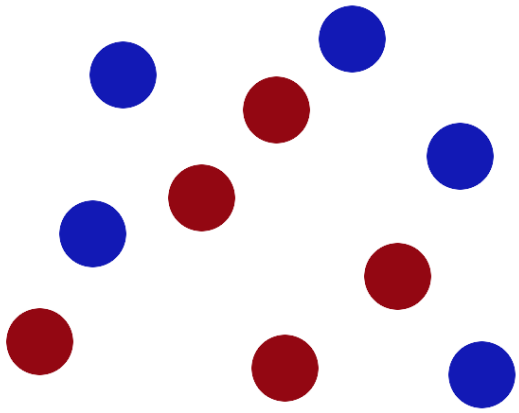

Figure 2.7: There exists no possible line that can separate the red balls from the blue ones.

A solution for this is to transform the input space to the feature space, where it is possible to separate the balls using a hyperplane, this is shown in Figure 2.8.

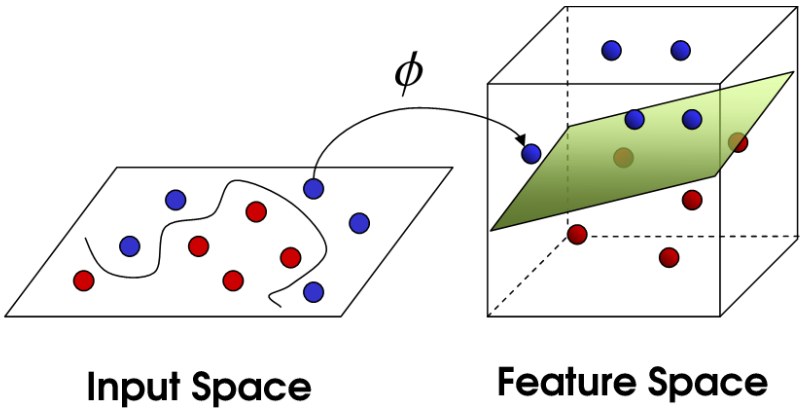

Figure 2.8: Transformation to a new features space where the balls can be separated by a hyperplane.

Back in the original feature space the separation boundary might look like Figure 2.9.

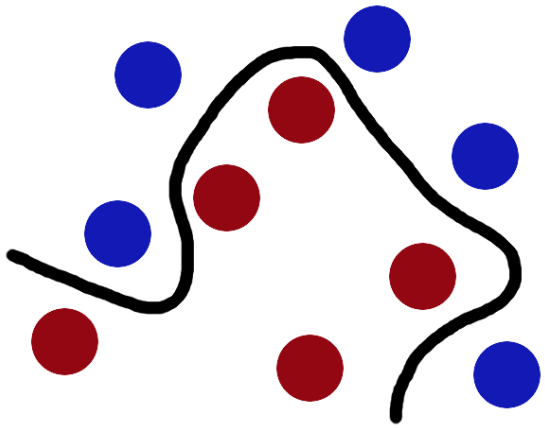

Figure 2.9: Separation boundary in the original feature space.

Random Forests

A random forest (RF) is an efficient learning algorithm based on model bagging and aggregation ideas[35]. The Random forests work by creating different decision trees. On their own, decision trees are very prone to overfitting. Random forests solve this problem by creating an aggregation of trees.

The word random in random forest indicates that some randomness is included. Each tree in a random forest looks at a random subset of the samples and a random subset of the features. This principle is shown in Figure 2.10. This random subset of samples is called the bootstrap sample and is selected out of N samples, by picking N samples with replacement. This results, on average, in 2/3 of the samples being selected (with some doubles). The other 1/3 of the

samples are then used as out of bag (oob) set. Averaging the performance of each tree on the out of bag set, offers an indication of the generalisation of the random forest.

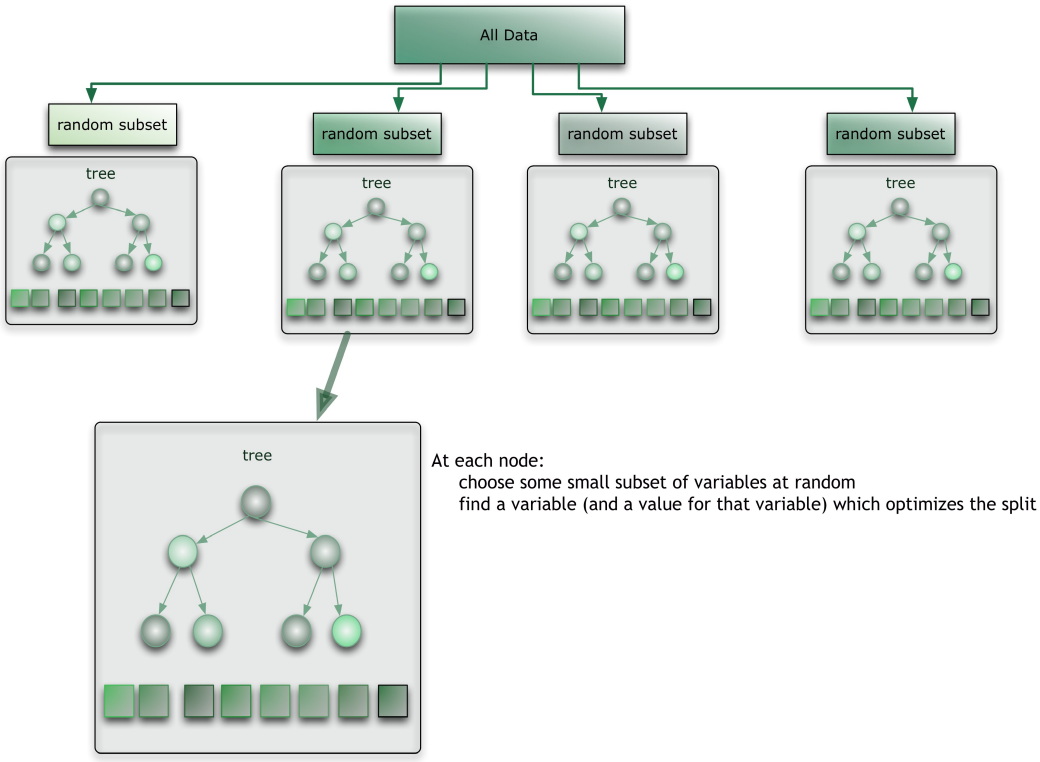

Figure 2.10: The structure of a random forest, found at [36]

To understand which features are good, one needs to understand the internal workings of a decision tree. Suppose the following example[2], where one tries to find an algorithm to predicted whether or not a person will play tennis on a given day. Suppose the training data is given by Table 2.1 and a prediction for the 15^{th} sample needs to be made.

[2]This example is based extensively on this youtube video: https://www.youtube.com/watch?v=eKD5gxPPeY0

Table 2.1: suppose the following training examples for a decision tree.

Day	Outlook	Humidity	Wind	Play tennis
1	sunny	high	weak	no
2	sunny	high	strong	no
3	overcast	high	weak	yes
4	rain	high	weak	yes
5	rain	normal	weak	yes
6	rain	normal	strong	no
7	overcast	normal	strong	yes
8	sunny	high	weak	no
9	sunny	normal	weak	yes
10	rain	normal	weak	yes
11	sunny	normal	strong	yes
12	overcast	high	strong	yes
13	overcast	normal	weak	yes
14	rain	high	strong	no
15	rain	high	weak	?

A decision tree will take a feature and split the data based on the possible outcomes of this feature. In case the features are continuous values, ranges are selected. In some cases the leafs will be pure, meaning that all samples in this leaf belong to a single class. The pure leaves in Figure 2.11 are displayed in green. In case the leave is not pure, another split is needed. Note that not all random forests split until all leaves are pure; random forest can be limited in depth, in that case the output is chose by a majority voting of the samples.

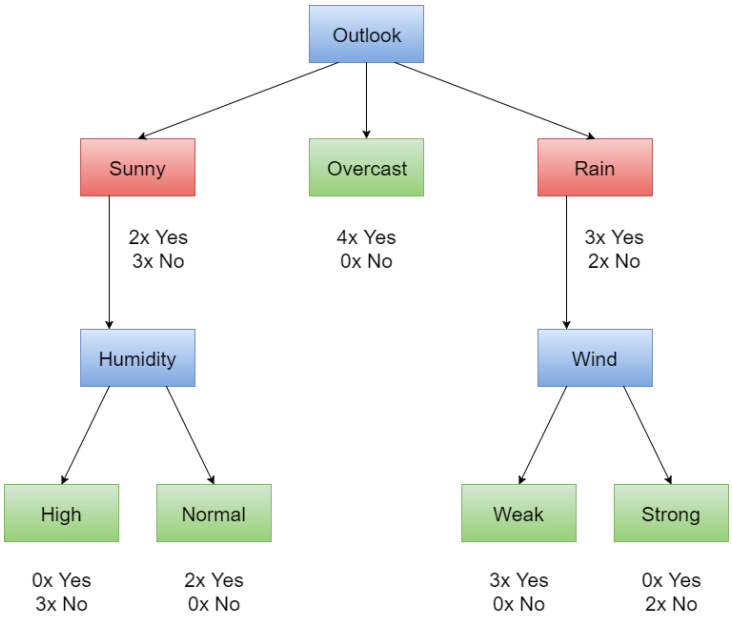

Figure 2.11: A decision tree for the data in Table 2.1

Once the tree is constructed it becomes clear that the predicted output of sample 15 is 'Yes'. This is obtained simply by following the tree branches. Even though the features are selected at random, they have influence on the accuracy. Good features will reduce the impurity significantly, thus the impurity reductions are a good indication for how important a feature is.

Since the importance is averaged over different nodes and different trees, it is also capable of detecting combinations of features that work well. One feature may not be important on its own, but might be a very good feature when combined with other features. Suppose the following example in Table 2.2:

label	**feature A**	**feature B**
Happy	+	+
Happy	-	-
Sad	-	+
Sad	+	-

Table 2.2: Some feature are not significant on its own, but a might be part of a combination of features.

It is clear that feature A and B are very important when it comes to predicting whether or not a person is happy or sad. When both features have the same sign, the person is happy, otherwise he is not. Combinations of features are often not found by feature selection methods as they look for correlations between a single feature and the output.

This problem does not occur for random forest though, as combinations of features are also 'tested' in the sense that a tree might split on them in different stages. Once the combination of features occurs randomly in a decision tree, the impurity will drop significantly, which will result in higher importance rankings.

2.4.3 Dimensionality Reduction methods

The algorithms described below perform a dimensionality reduction, often by projecting a high dimensional to a lower dimensional features space. Looking at coefficients of these trained models, indicate what features are important.

Common Spatial Patterns

Common Spatial Patterns (CSP) is a supervised technique that has its origin in the optimization of motor imagery BCIs[37]. It is a common technique in BCI research[38, 39, 40]. CSP creates linear combinations of the original EEG channels that maximize the variance for one class while simultaneously minimizing the variance of the other class [38]. One disadvantage of using CSP is that the default version can only distinguish between 2 classes, though one can easily aggregate multiple CSP models to create one-vs-one and one-vs-all models, similarly to the one-vs-one and one-vs-all SVMs.

The input for a CSP filter is a set of N labelled samples $E_j (j = 1...N)$, with dimension N_{ch} x T_j, with N_{ch} being the number of EEG channels and T_j the number of samples in a single trial[37].

First the train data is split into two classes, before computing the covariance matrices of both classes.

$$\Sigma_1 = \sum_{j \in C_1} X \frac{E_j E_j^T}{trace(E_j E_j^T)}$$
$$\Sigma_2 = \sum_{j \in C_2} X \frac{E_j E_j^T}{trace(E_j E_j^T)}$$

Note that the average of E_j is expected to be zero, because a bandpass filter is applied that makes the DC component of the signal zero. The next step is to calculate the composite covariance matrix.

$$\Sigma = \Sigma_1 + \Sigma 2$$

Next the covariance matrix is diagonalised by calculating the eigenvalues and eigenvectors of Σ.

$$V^T \Sigma V = P$$

The eigenvalues are then found on the diagonal of P, each eigenvalue corresponds to an eigenvector found in the columns of V.

The next step is the whitening transformation.

$$U = P^{\frac{1}{2}} V^T$$

Which results in

$$U \Sigma U^T = 1$$

Next the following two matrices are calculated:

$$R_1 = U \Sigma_1 U^T$$
$$R_2 = U \Sigma_2 U^T$$

R_1 is then diagonalised

$$Z^T R_1 Z = D = diag(d1, ..., d_m)$$

The eigenvalues on the diagonal are then sorted, as larger eigenvalues correspond to higher importances. Next the filters are determined by:

$$W = Z^T U$$

The EEG channels can then be filtered as follows:

$$E^{CSP} = W E^{orig}$$

Since CSP filters create simple linear combination of incoming channels, they can also be used as feature selection mechanism,albeit in a limited fashion. The result of a CSP transformation are again a set of EEG channels, where each channel is a combination of the previous channels. The first and last row of the resulting matrix W shows the coefficients for which the variance is maximized between the two signals. Looking at those coefficients, one can determine which channels are of more importance than other and thus which channel locations have the most influence on emotion.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), is a machine learning technique often used in combination with CSP[38, 39, 40]. LDA looks for a projection of the data where the data is linearly separable, as shown in Figure 2.12. Looking at the coefficients of the LDA model, one can again determine the importance of the different features.

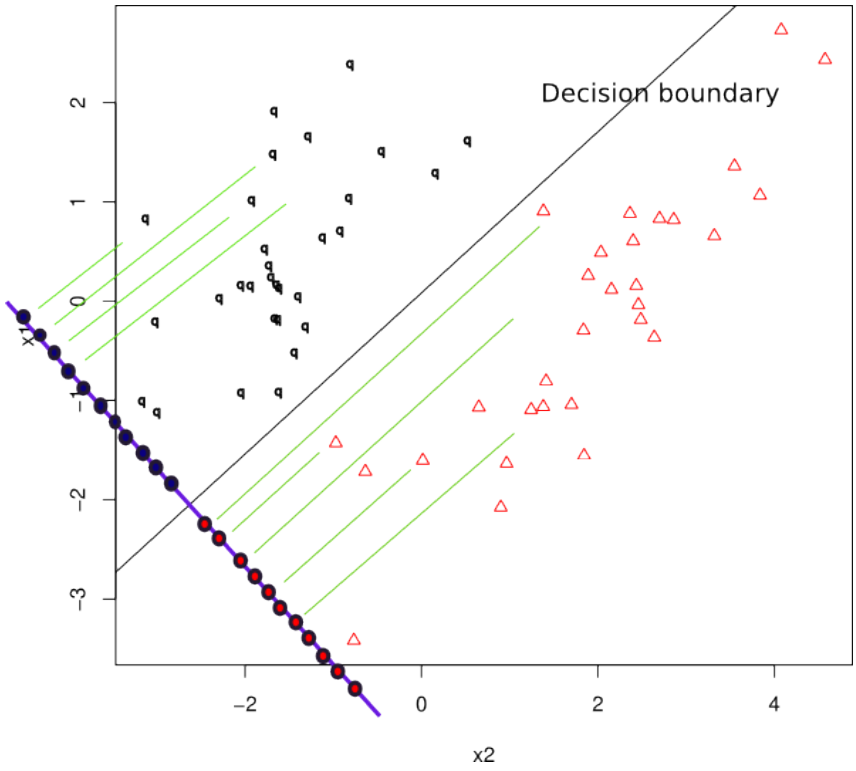

Figure 2.12: LDA finds a projection of the data where the separation of the data is clear.

Principal Component Analysis

Principal Component Analysis (PCA) is a technique to do dimension reduction. Intuitively, PCA can be seen as fitting an n-dimensional ellipsoid to the data. The Principal components are then the axes of the ellipsoid. Less variation in one direction, corresponds to a smaller axis. Removing that axis, will only remove a small fraction of the information, as there is only little variation in that direction. This is shown in Figure 2.13, where the ellipsoid covers a three dimensional features space. The ellipsoid has three axes: a,b and c. Intuitively, one can see that there is more variation (information) in the c ans b direction, while the a axis is relatively small.

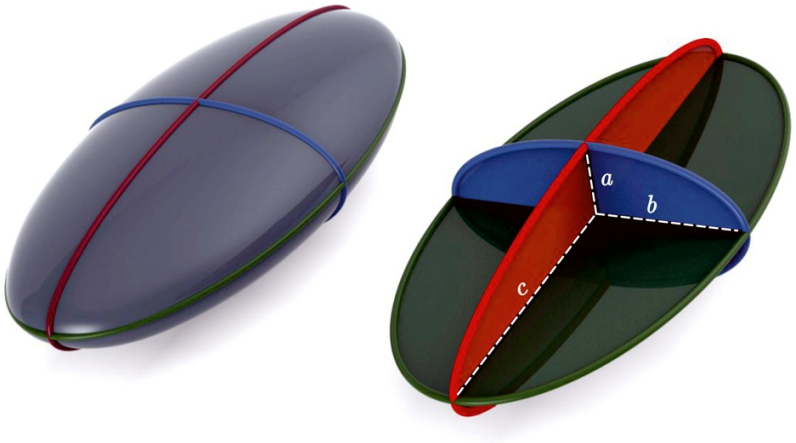

Figure 2.13: Suppose a three-dimensional feature space, where all points lie in the ellipsoid in the left.

Removing the a axis by projecting the data on the plane given by vectors b and c, will result in a two dimensional projection of the data in the form of an ellipse. This would be the black plane in Figure 2.13. This process can be repeated for higher dimensional features spaces. In other words, PCA will thus, without going into too much detail, start with an n-dimensional ellipsoid and iteratively remove the smallest axis in each iteration until the desired number of dimensions is obtained. Note that the ellipsoid should be adjusted in each step.

The major disadvantage of PCA is that the algorithm is unsupervised, meaning that it does not look at the corresponding labels of the given samples. Suppose the difference between two classes was clearly given by looking at the a axis in Figure 2.13. PCA would, in that case, result in a total loss of all information.

Advanced RF feature selection

One advanced method for feature selection is the two-step method using random forest, described in [35]. The paper states that there are two possible motivations for feature selection. The first motivation is to do interpretation, find out which features are important and use them for research. In the context of BCI, feature interpretation could help neuroscientist find out which parts of the brain are affected by an emotion, for example. The second motivation is to improve machine learning techniques, having fewer features will not only speed up training and prediction times, it also reduces the complexity, which often has a good influence on the generalisation property of a machine learning algorithm. Additionally in the context of BCI research and EEG data gathering, using fewer electrodes means less preprocessing time; mounting 32 electrodes to the brain of a subject is a time consuming task.

The selection procedure itself consists of two steps, in the first step data is fitted to a random forest and the importance values for each feature are determined, by taking the average and standard deviation of the importances over all trees. All features are then ranked based on their importance ranking, before features with small importance are cancelled.

Then, depending on the motivation of feature selection, one of two possible second steps is performed. For feature interpretation the second step starts by fitting a random forest using a single feature. The OOB is then averaged over multiple runs. The runs are needed because a random forest has an element of randomness; fitting the same data twice to a random forest, will not give you the same random forest. To get an accurate estimate of the performance of a random forest, it might be a good idea to fit the data several times. The average OOB score and its standard deviation is then used to determine an initial OOB score.

$$OOB_{init} = AVG(OOB) - STD(OOB)$$

The standard deviation is used to avoid noisy results, a result is only regarded as better, when it is better from a statistical point of view. Next features are added iteratively, when a larger features set has a better average OOB score (taking the standard deviation into account), the feature set is replaced by the larger feature set. Note that the whole set of features is always considered; it is not possible to leave a feature out and include the next feature.

The other possible second step is used for prediction, here the algorithm starts similarly, by determining an initial average OOB score and standard deviation. The idea behind the standard deviation is the same as with the interpretation step, noise removal and stability.

$$OOB_{init} = AVG(OOB) - STD(OOB)$$

The next part is different, in each iteration a feature is introduced. When the average OOB score and standard deviation of the feature are better, the feature is added to the feature set, otherwise it is neglected. This is a greedy forward selection algorithm, once a feature is selected it remains selected. The difference between the interpretation and prediction step is that here single features are added to the feature set, while step two-interpretation always takes the whole feature set as a replacement. Step two prediction on the other hand is able to select a distinct set of features out of the results from step one.

In the end the paper notes several observations, the step two-prediction method provides better OOB scores using fewer features. Additionally they mention that highly correlated features might confuse the algorithm, as correlated features have lower importances.

Bibliography

[1] H. Verschore, "A brain-computer interface combined with a language model: the requirements and benefits of a p300 speller," afstudeerwerk, Ghent University, June 2012.

[2] D. O. Bos, "Eeg-based emotion recognition," 2007.

[3] M.-K. Kim, M. Kim, E. Oh, and S.-P. Kim, "A review on the computational methods for emotional state estimation from the human eeg," *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 573734, p. 13, 2013.

[4] T. C. Technologies, *10/20 System Positioning manual*. Fortis Tower, 2012.

[5] unknown, "Electrode placement," 2015.

[6] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Time-frequency optimization for discrimination between imagination of right and left hand movements based on two bipolar electroencephalography channels," *EURASIP journal on Advances in Signal Processing*, vol. 2014, no. 38, 2014.

[7] K.-E. Ko, Hyun-Chang, and K.-B. Sim, "Emotion recognition using eeg signals with relative power values and bayesian network," *International Journal of Control, Automation, and Systems*, 2009.

[8] Brainworks, "What are brainwaves?," 2015.

[9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 18–31, Jan 2012.

[10] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1970–1973 vol.3, Oct 1996.

[11] J. Wagner, J. Kim, and e. AndrÃ©, "From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification," *IEEE*, 2005.

[12] Y. Lio, O. Sourina, and M. K. Nguyen, "Real-time eeg based human emotion recognition and visualization," 2010.

[13] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and clinical Neurophysiology*, vol. 70, no. 70, pp. 510–523, 1988.

[14] T. Verhoeven, "Brain-computer interfaces with machine learning: an improved paradigm for the p300 speller," afstudeerwerk, Ghent University, June 2013.

[15] N. V. Manyakov, N. Chumerin, A. Combaz, and M. M. V. Hulle, "Comparison of classification methods for p300 brain computer interface on diabled subjects," *Computational intelligence and neuroscience*, 2011.

[16] Cognionics, "Cognionics dry eeg p300 speller demo," 2015.

[17] Y. Morita, K. Morita, M. Yamamoto, Y. Waseda, and H. Maeda, "Effects of facial affect recognition on the auditory {P300} in healthy subjects," *Neuroscience Research*, vol. 41, no. 1, pp. 89 – 95, 2001.

[18] Y. Lio and O. Sourina, "Eeg databases for emotion recognition," *International Conference on Cyberworlds*, 2013.

[19] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *CoRR*, vol. abs/1601.02197, 2016.

[20] stackexchange, "When is a model underfitted," 2016.

[21] Fliptop, "bias, variance and overfitting overview," 2016.

[22] J. Kim and E. Andre, "Emotion recognition based on phys- Â´ iological changes in music listening," *IEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[23] D. O. Bos, "Eeg-based emotion recognition the influence of visual and auditory stimuli."

[24] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1798–1806, July 2010.

[25] L. Brown, B. Grundlehner, and J. Penders, "Towards wireless emotional valence detection from eeg," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2188–2191, Aug 2011.

[26] M. Murugappan, "Human emotion classification using wavelet transform and knn," in *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on*, vol. 1, pp. 148–153, June 2011.

[27] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for eeg-based emotion classification," pp. 81–84, Nov 2013.

[28] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Fusion of electroencephalographic dynamics and musical," 2014.

[29] P. LanG, M. Greenwald, M. Bradley, and A. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions.," *Psychophysiology*, vol. 30, pp. 261–273, May 1993.

[30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[31] Y. Peng, S. Wang, S. Long, and L. B.-L., "Discriminative graph regularized extreme learning machine and its application to face recognition," *Neurocomputing*, vol. 149, pp. 360–353, 2015.

[32] S. L. Crawford, "Statistical primer for cariovascular research," 2006.

[33] J. P. Pluim, A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions of Medical Imaging*, vol. 22, pp. 986–1003, August 2003.

[34] G. J. Szekely and M. L. Rizzo, "Brownian distance covariance," *The annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[35] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, pp. 2225–2236, 2010.

[36] Citizennet, 2016.

[37] S. Vercruysse, "Adaptieve common spatial patterns voor de classificatie van ingebeelde bewegingen," afstudeerwerk, UGent, 2011.

[38] A. Coone, "A study on different preprocessing and machine learning techniques for the detection of error-potentials in brain-computer interfaces," afstudeerwerk, Ghent university, June 2011.

[39] F. Lee, R. Scherer, R. Leeb, C. Neuper, H. Bischof, and G. Pfurtscheller, "A comparative analysis of multi-class eeg classification for brain computer interface," in *Proceedings of the 10th Computer Vision Winter Workshop*, pp. 195–204, 2005.

[40] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, "Current trends in graz brain-computer interface (bci) research," *IEEE Transactions on rehabilitation Engineering*, vol. 8, pp. 216–219, JUNE 2000.