Signal Processing for a Brain Computer Interface

By

Ruiting Yang

Thesis submitted for the degree of

Master of Engineering Science

School of Electrical & Electronic Engineering

Faculty of Engineering, Computer & Mathematical Sciences

The University of Adelaide

Adelaide, South Australia

2009

Contents

Declaration

NAME:.....Ruiting.....Yang..... PROGRAM:.....

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNATURE: DATE:

Acknowledgements

I would like to express my gratitude to those people who have given me support and assistance throughout my time as a master student.

First of all, I would like to thank my principal supervisor, Professor Doug Gray. It would have been impossible to finish this thesis without his constant guidance, encouragement, support and understanding.

Many thanks also go to my co-supervisor, Dr. Brian Ng who contributed more than I would have expected from a usual co-supervisor. It has been so important to get his academic advice and particular understanding in our regular meetings. His communication with the School of Molecular & Biomedical Science was crucial for the experiments in this thesis, and I will always appreciate his contributions.

Thanks to Dr. Michael Ridding from the School of Molecular & Biomedical Science, for sharing his experimental equipment and organizing the experiments. Similarly, thanks must go to the organizer of the BCI competition II for sharing their dataset, on which we relied for much of our research.

Thanks to my Chinese supervisor, Professor Mingyi He for his constant encouragement and kind help during these years. He was one of founders of the agreement of exchanging students between the University of Adelaide and Northwestern Polytechnical University, which brought me to Australia.

Thanks to Associate Professor Cheng Chew Lim, another founder of this program. I can clearly remember the interview in China, which actually has changed my life already. Thanks also to Associate Professor Michael Liebelt and the school committee, who decided to grant me a fee remission scholarship to study here.

Thanks to Mr. Matthew Trinkle, whose enthusiasm and patience has been an inspiration for me. His help and kindness made it much easier for me to fit into this unfamiliar country.

Thanks to other colleagues in Sensor Signal Processing Group, their kind help, suggestions and news shared during coffee break and Friday drinks

were always pleasant, and gave me a chance to know Australian culture better. Thanks also to other friends in Australia, who provided valuable support and companionship in my time here and without whom I would be much lonelier in Australia.

Finally, thanks must go to my parents and other family members, their consistent financial support and encouragement is the basis of my overseas study. Their love is always the force driving me forward.

Abstract

Brain computer interface (BCI) systems measure brain signal and translate it into control commands in an attempt to mimic specific human thinking activities. In recent years, many researchers have shown their interests in BCI systems, which has resulted in many experiments and applications. However, most methods are just based on a specific selected dataset or a typical feature. As a result, there are questions about whether some methods generalise well on other datasets. Therefore, the major motivation of this thesis is to compare various features and classifiers described in the literature.

Pattern recognition is considered as the core part of a BCI system in our research. In this thesis, a number of different features and classifiers are compared in terms of classification accuracy and computation time. The studied features are: time series waveform, autoregressive (AR) components, spectral components; these are used with different classifiers: such as template matching, nearest neighbour, linear discriminant analysis (LDA), Bayesian statistical and fuzzy logic decision classifiers.

In order to assess and compare these different features and classifiers, an extensive investigation was carried out on a public dataset (imagined left or right hand movement) from an international BCI competition and the results are reported in this thesis. The classification was done in a continuous fashion, to match a real time application. In this process, the average and best accuracy, as well as the computation time, were analysed and compared. The results showed that most classifiers achieved very high accuracies and short computation times for most features.

A BCI experiment based on imagined left or right hand movement was carried out at the University of Adelaide and some investigations on the data from this experiment are discussed. The result shows that the selected classifiers can work well with this new dataset without much additional preprocessing or modifications.

Finally, this thesis culminates with some conclusions based on our research, and discusses some further potential work.

Abbreviations

ALS:	Amyotrophic Lateral Sclerosis
ANN:	Artificial Neural Network
AR:	Autoregressive
BCI:	Brain Computer Interface
CSP:	Common Spatial Pattern
CSSD:	Common Subspace Decomposition
DFT:	Discrete Fourier Transform
ECoG:	Electrocorticography
EEG:	Electroencephalography
EM:	Expectation Maximization
EMG:	Electromyography
EOG:	Electrooculography
ERD:	Event Related Desynchronization
ERS:	Event Related Synchronization
FFT:	Fast Fourier Transformation
FMRI:	Functional Magnetic Resonance Imaging
GMM:	Gaussian Mixture Models
LDA:	Linear Discriminant Analysis
LMS:	Least Mean Square
LOO:	Leave One Out
MLP:	Multi-Layer Perceptron
PDF:	Probability Density Function

List of Figures

List of Tables

Publication

Ruiting YANG, Douglas A. GRAY, Brian W. NG, Mingyi HE, Comparative Analysis of Signal Processing in Brain Computer Interface, accepted by the 4th IEEE Conference on Industrial Electronics and Applications, Xi'an China 25-27, May, 2009

Chapter 1 Introduction and Literature Review

This chapter introduces background materials for the research reported in this thesis. It begins with a brief introduction to Brain Computer Interface (BCI) in section 1.1. This is followed by a summary of physiology literature related to this research in section 1.2 and a technical description of the key parts of a Brain Computer Interface, including signal acquisition, signal processing, application and feedback in section 1.3. Section 1.4 provides a literature review of BCI techniques, including experiments, features and classifiers in BCI research. Finally, section 1.5 states the major aim of this thesis.

1.1 Concept of Brain Computer Interface

A disabled person, such as a serious amyotrophic lateral sclerosis (ALS) patient, may lose the ability to exercise language and muscle functions, which are two common ways of human information output. The communication pathway to the outside world can be restored for these patients if their intentions can be translated from their brain signals into actions by the use of a machine. Brain Computer Interface research aims to provide the means of fulfilling this promise.

Brain Computer Interface research provides a new communication channel from humans to devices via a computer. At the first international meeting for BCI technology, it was agreed to define the term BCI as a system that does not depend on the brain's normal output pathways of peripheral nerves and muscles [1]. According to the definition, a Brain Computer Interface should be able to detect human intentions and translate them to the computer where suitable actions are carried out. Typically, a BCI system consists of several components: brain signals, signal acquisition, signal processing, operation of application and feedback presentation. Human intentions modulate the electrical brain signals which are detected and recorded by the signal acquisition block and then filtered by the signal preprocessing block. The signal processing block, which includes processes such as feature extraction and classification, subsequently analyses the captured signals and provides

the corresponding instructions to appropriate devices. During the operation of these devices, some feedback may be returned to the user(s). The block diagram of a BCI system is shown in Figure 1.1.

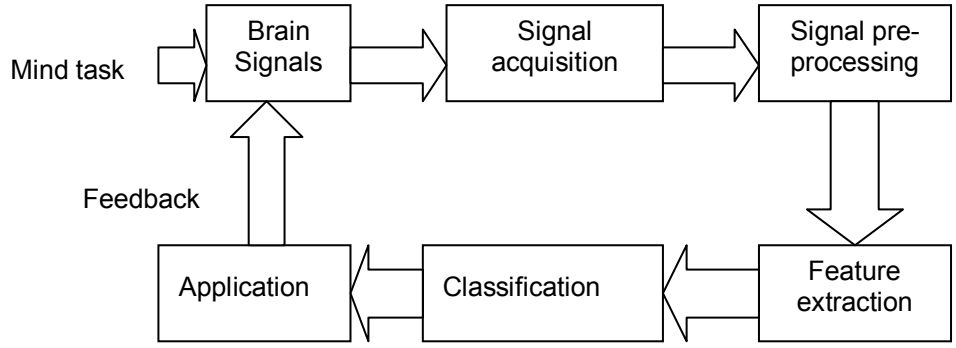

Figure 1.1. A typical Brain Computer Interface system.

BCI signal acquisition systems are broadly divided into two classes, defined in terms of the manner with which the brain signal is captured, and these are known as invasive and non-invasive methods. The invasive approach relies on using electrocorticography (ECoG) [1] from electrodes implanted inside the skull and it has been applied to some epilepsy patients [42] and monkeys [8]. The non-invasive approaches are based on using electroencephalography (EEG) from electrodes on the scalp or imaging techniques, such as functional Magnetic Resonance Imaging (fMRI) [12, 13]. For the BCI applications, the advantages of using the ECoG include higher signal to noise ratio [43], shorter training time [42] and being free of problems with muscular and ocular artifacts [45]. These advantages generally ensure greater classification accuracy. However, surgery and the subsequent recovery period pose great risks for the patient. Additionally, it is hard to find volunteers for such methods and the applications would be limited to a few patients with serious neurological problems or animals. So far, fMRI technology is too expensive and the equipment is not portable for a broad and practical use in BCI. Therefore, the non-invasive EEG is still the most preferable signal acquisition technique in current BCI research.

[1] Electrocorticography (ECoG) uses electrodes placed directly on the surface of brain to record electrical activity from the brain cerebral cortex.

1.2 Physiology background

1.2.1 The human brain

The average human brain only weighs about 1400 grams but it contains billions of neurons and the number of synapses in the brain, estimated to be on the order 10^{14} [68]. There are four main parts in the human brain: cerebral cortex, diencephalon, cerebellum and brain stem [41]. The most relevant part to BCI is the cerebral cortex and it can be divided into two hemispheres. From the topographic aspect, each hemisphere can be divided into frontal, parietal, occipital and temporal lobes. According to relevant functions, the hemisphere can be divided into several areas, such as auditory cortex, sensor motor cortex. These areas can be seen in Figure 1.2.

Figure 1.2. Main parts and functional areas of the brain [40]

1.2.2 Electroencephalography (EEG)

EEG reflects electrical activities continuously produced in the brain and is typically described in term of rhythmic brain waves, which are primarily grouped together according to their frequency. Some common EEG waves are named after Greek letters and shown here in Table 1.1.

Name	Frequency (Hz)
Delta(δ)	Up to 3
Theta(ɵ)	4~7
Alpha(α)	7~13
Beta(β)	14~26
Gamma(γ)	34~100

Table 1.1. Common EEG waves and their frequency range

Different brain waves have different signal intensities at a given location of the scalp and usually arise from different thinking activities. For example, the alpha rhythm (7-13 Hz) always exists in the occipital regions, when a healthy adult is relaxed or has his/her eyes closed. It is generally involved in perceptual, judgment and memory functions [18]. The beta rhythm (14-26 Hz) is prominent when people are excited or active. In addition to these common brain waves, another brain wave, mu rhythm, appears over the primary sensorimotor cortical areas, when a body movement is suppressed or when there is an imagined movement [4]. With a hand movement (or a mental imagery of the hand movement), it characteristically attenuates the signal in relevant positions on the contralateral side. The mu rhythm occupies the same frequency range as the alpha rhythm and peaks at about 10 Hz. Central beta rhythm is another kind of sensorimotor related brain wave, and peaks at 20 Hz. It is similar to the mu rhythm but with different locations and shorter recovery time after the signal attenuation [38].

In summary, EEG signals in the range 7~30 Hz form the main frequencies of interest for a motor cortex based BCI system, more specifically, 7~13 Hz for the mu rhythm and 14~26 Hz for the beta rhythm. The mu rhythm has the same frequency as alpha rhythm, so, the frequency band of mu rhythm is usually labeled as alpha band and similarly, the band of 14~26 Hz is labeled as beta band.

Each part of the human body has a corresponding region in the motor and somatosensory area of the neocortex (see Figure 1.3). For example, the left hand is represented laterally on the right hemisphere and the right hand on the left hemisphere. Specifically, they correspond to positions C4 and C3 in the International 10-20 System of Electrode Placement [32], which is shown in Figure 1.4. The mapping of brain area to function is reasonably consistent, but there is still some variance among individuals [67]. Subsequently, there is also some variance in the placement of electrodes, which has resulted in some BCI systems not being able to work consistently on every subject [31]. When a BCI system aims to provide solutions for a particular individual rather than a general group, the electrode placement can be optimised by an experienced technician. When a subject is relaxed and not engaged with one of his/her

limbs (real or imagined), the intensity of the idle brain signals in related positions lie in a specific range but are not zero. When the subject starts to move his/her arm or even tries to imagine a movement of the arm, there will be an attenuation to the idle rhythm around 7~30 Hz, specifically, 7~13 Hz for the mu rhythm and 14~26 Hz for the beta rhythm. As the attenuation effect is due to suppression of synchrony in neural system, it is termed as event related desynchronization (ERD). There is also an effect of enhancing signal intensities in these two frequency bands, which is termed event related synchronization (ERS) [46]. Imagining different movements can cause different ERD and ERS effects on the idle signal and produce different modulated signals. This is the main physiological cue used by BCI systems to identify a subject's thought.

Figure 1.3. Geometric mapping between body parts and motor cortex [39]

Figure 1.4. International 10-20 System of Electrode Placement [37]

Since the sampling rate of typical EEG acquisition equipment is much higher than the Nyquist frequency required for EEG signals (e.g., two datasets used in this thesis), the temporal resolution is usually good but the spatial resolution is poor due to the relatively large distance (a matter of centimeters) between electrodes and the EEG overlapping from different areas. The spatial resolution can be improved by using surface Laplacian filters [69], but this relies on using more electrodes in the signal acquisition. Both spatial and temporal data are available, so, the problem can be viewed as a space-time-processing problem although most researchers have treated the processing in the spatial and temporal domains separately. Usually, data are processed spatially and then analysed in the temporal domain (e.g., [4, 6]).

1.3 Key components of a Brain Computer Interface system

As discussed above, a BCI system consists of several components. Technically, the BCI system can be divided into 3 key parts: signal acquisition, signal processing and application. Additionally, feedback has been used in certain BCI systems and these then form closed loops.

1.3.1 Signal acquisition and pre-processing

The EEG signal can be recorded easily and with inexpensive equipment (specifically, several electrodes and an amplifier). In order to capture the proposed EEG signals, it is important to accurately place electrodes on the correct positions according to the International 10-20 System of Electrode Placement. When the signal is captured, an amplifier amplifies it with typically 60–100 dB of voltage gain. Such large gains are necessary since the amplitude of EEG signals is about 100 μV [41]. In this process, the captured analogue signal is sampled to produce a digital signal by the acquisition equipment. Usually, before further analysis, it is necessary to perform some pre-processing, such as filtering out unnecessary frequency components and removing artifacts. It is noted that artifacts measured by electromyography (EMG)[2] and electrooculography (EOG)[3] can seriously degrade the EEG signal.

[2] Electromyography (EMG) is a technique to measure the activation signal of muscles.
[3] Electrooculography (EOG) is a technique to measure the resting potential of the retina.

1.3.2 Signal Processing

When the EEG signals have been transmitted to a computer, signal processing algorithms convert them into control commands, which are sent to devices. The conversion algorithm typically has two stages: feature extraction and classification. Feature extraction aims to obtain the most appropriate features which reflect differences between various classes of brain signals, while classification aims at determining the class to which any particular brain signal belongs. If a subject's intentions could be identified accurately, it would be straight forward to issue relevant commands to the electromechanical devices that make the physical control movements. At this stage, the problem is one of controlling such devices and falls outside the scope of this thesis. Therefore, the core of a BCI system is considered as a pattern recognition system and the research in this thesis focuses on the feature extraction and classification algorithms used in BCI.

Normally, the recorded EEG signals can reflect relevant brain activities but it is currently impossible to identify every intention, because the understanding of neural activities is still limited and only a few patterns of intentions are known well. Additionally, thinking processes are always complex and changeable, which also makes the patterns complicated. However, it is unnecessary to know every intention, since the practical applications of BCI can be achieved by identifying a limited set of intentions. For example, in a gaming environment, the basic operations, such as turning left or right in a virtual car on a screen, can be achieved by imagining left or right hand movement. Therefore, in current research, it is sensible to focus on detecting some recognizable intentions, such as imaginary hand movements, rather than general thoughts.

1.3.3 Application

The classification process produces a trigger or control command for devices. However the control aspect can have an influence on the nature and specification of the signal processing block as different control strategies place different demands on the signal processing. In BCI research, offline classification is easier since a longer time is allowed to make decisions.

However, for many practical applications, real time operation is a key requirement, and in this case only short time delays are acceptable, since a control process should track a subject's thinking in a timely manner. Therefore, the classification must be finished promptly when online operation is required. Further, classification results may be used as continuous inputs to a control system rather than just a command.

1.3.4 Feedback

A BCI system is usually a closed loop system and two types of feedback are involved in, as shown in Figure 1.5. The two types of feedback result in a mutual learning process, which helps both a subject and computer adapt to each other. Bio-feedback is helpful for the user to acquire the skills of controlling his/her EEG response in a BCI system and machine feedback is essential to modify the classifier and training dataset. The bio-feedback may speed up the learning process and improve performance, because feedback would give subjects the motivation to stay concentrated for the length of the trial and would help subjects to correct their thinking model in reaction to a wrong classification. The performance of an application can be used as the feedback, for example, in a game scene, whether the movement of a cursor on screen follows the decoded intentions. On the other hand, there may be some harmful effects from the bio-feedback, such as provoking feelings of frustration, attempts to guess the next cue or intention and inducing EEG artifacts from visual stimulus.

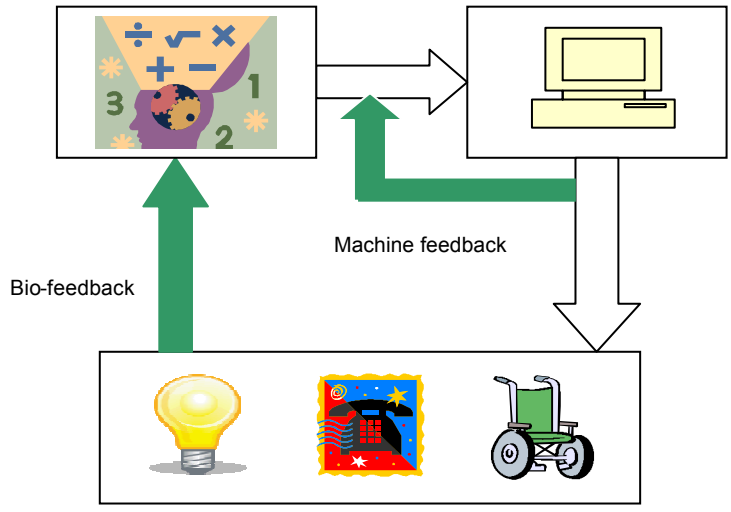

Figure 1.5. Feedbacks in a closed loop BCI system.

1.4 Literature review

Research in the field of BCI has a short history. In the 1970's, several scientists developed some simple BCI projects that were driven by electrical activity recorded from the head. Among them, the most successful project, which aimed to control the movement of a cursor, was carried out by Dr. Jacques Vidal [1]. In recent years, many groups around the world have investigated BCI systems, resulting in many experiments and applications such as cursor [4], wheelchair [29] and robot [7] control, and very high recognition rates have been achieved. However, it is hard to compare the performances due to different experimental arrangements and measurement techniques. Some of these experiments were performed on the paralyzed subjects [44], but the majority of them were performed on healthy users. In an international BCI competition [15], it became obvious that some of the submitted algorithms did not generalise well to other BCI datasets. Therefore, it invited doubts as to whether BCI systems could work generally rather than only for some specific data. On the other hand, extensive research [31] in Graz showed that using 99 subjects, 70% of the sessions were classified with an accuracy of 60% to 80%, which proves that the BCI research can be useful for universal users.

1.4.1 BCI experiments

For a BCI experiment, the number and position of electrodes is the initial issue to consider. Generally, using more electrodes requires greater effort in the preparation and recording of data. As a result, there is a trend that many researchers prefer to use only a few electrodes. Usually the recording method is to use bipolar electrodes or electrodes with a common reference electrode, typically located on an ear or an event irrelevant position. Typically, two electrodes were used at positions C3 and C4 as a bipolar channel and the difference between them was detected as the signal [47]. Some motor based BCI systems used positions C3, C4 and Cz to record three-dimensional EEG signals, as described in the literature from Graz [6, 28]. More electrodes were used by some researchers, such as the 13 electrodes used for a wheelchair application [29]. Almost all the researchers use the standard international 10-

20 System of Electrode Placement and many of them focus on the positions C3 and C4 when attempting to classify imaginary movements. Even when more electrodes were involved, they were either only used as the reference electrodes or even ignored in further processing [6].

There are several kinds of tasks designed for the BCI applications, such as imagining hand [28] and feet [9] movements. Most systems are based on the synchronous protocols, which usually limit the thinking in several seconds after a given cue [3, 11, 17]. In contrast, the asynchronous protocols were used in some research, which allows the subjects to imagine some given task but without the use of a cue [4, 7, 9].

1.4.2 Feature selection

Determining suitable features is a key precursor to a successful BCI system. Many features have been investigated in BCI research, such as amplitude of EEG signals [11], band powers [7, 9, 31, 52], phase synchronization [3], matched filter outputs [4], and regressive parameters [27-29]. In practical feature extraction of EEG signals, the following issues need to be considered:

Normally, the EEG signal or BCI features are noisy. Low pass filtering is used extensively, since most sources of noise have many higher frequency components while the signals do not. Artifact removal is usually performed manually during the experiment, such as visual inspection or monitoring the eye movements (e.g. blinks) with an electrode [6]. The method of linear regression has been used to remove the artifacts by automated correlation [48, 49]. Instead of removing the artifacts, some methods aim to transform the signal to a new space, such as the common spatial pattern (CSP) technique [33] or separate the task related EEG from the artifacts, such as common subspace decomposition (CSSD) [10,11]. These methods are based on the Karhunen-Loève transform [23], which will be discussed in the next chapter.

Another issue is the high dimensionality of BCI features. BCI feature vectors often have high dimensionality since numerous electrodes are used. Actually, the useful features are generally extracted from only a subset of all available channels. Besides judging and placing electrodes at correct

positions in an experiment, some methods have been considered to reduce the number of features. For example, 9 channels were removed while the remaining 19 channels were kept and used as features in reference [11]. The authors compared the difference of EEG amplitudes in the same channels between a subject's thinking activities corresponding to different classes and rejected the irrelevant channels. In another study [3], the number of electrodes was reduced from 22 to 7 by selecting some typical geometric positions on the subjects' head. In another study [6], a reference was obtained for an electrode by applying a Laplacian filter to the 4 nearest neighbours. In doing so, the dimension of signal was reduced by 4.

BCI features should contain some time information as thinking activities and response are related to specific events in time. To reflect these activities, features extracted from EEG signals must be able to change over the period of each session or trial. To deal with a time course of the EEG signal, two main approaches utilising temporal information have been proposed. The first method is to divide the time course into several segments with a fixed time window and to extract features from one of them or a combination of different segments. The most common examples are to estimate spectrum from the samples in a fixed time window [3, 4, 14, 35, 53] and moving the window during the process. The other method is to deal with the time course with a changeable size window and work as a dynamic classification [27-29]. The extreme example is to use all previous time information at every sample [34]. This has been used to achieve satisfactory results; adaptive classification (see 1.4.3) in [16, 17] is based on updating the classifier through renewing the feature space as new temporal information becomes available.

1.4.3 Classification

Classifiers have been investigated intensively by machine leaning researchers and several kinds of classifiers have been applied to BCI systems. Four basic categories are reviewed here: nearest neighbour classifiers, neural networks, linear classifiers and Bayesian statistical classifiers.

1.4.3.1. Nearest neighbour classifiers

The nearest neighbour classifiers have been used in BCI systems but usually in the role of a baseline for comparison purposes. Basically, they assign a feature vector to a class according to the one or *k* nearest neighbours. Usually, those neighbours are determined by Euclidean distance. An example is the work of Blankertz et al [21]; EEG signals were low pass filtered and cut off at 5 Hz or without any pre-processing, then, signal amplitudes in different channels were used as features and the *k* nearest neighbour approach was used as the classifier, but the performances were not satisfactory. However, in study [22], the normalized signal amplitude feature was simplified by a Karhunen–Loève transform [23] and good performance was achieved by the combination of the simplified feature and the nearest neighbour classifier. Another study [29] used the averaged correlation coefficient as the feature and the nearest neighbour classifier to obtain nearly 80% for the recognition rate.

In the above examples, only the Euclidean distance was used. Euclidean distance considers that equiprobable classes have the same diagonal covariance matrix, while the Mahalanobis distance considers the situation for non-diagonal covariances [20]. There are scarce examples of the Mahalanobis distance based classifiers in BCI literature. Babiloni et al applied the combination of spectral features and a Mahalanobis distance based classifier to a dataset which involved 8 subjects and achieved very high accuracy [24].

1.4.3.2. Neural network

Artificial neural network (ANN) is a popular non-linear classifier. When composed of sufficient neurons and layers, a neural network can approximate any continuous function and any number of classes can be classified by it [20]. This makes the Multi-layer perceptron (MLP) a very flexible classifier that can adapt to a great variety of problems. Anderson et al [14] have applied a neural network classifier with AR coefficients to BCI research. However, it was found to be necessary to make some improvements to both of the AR model and classifier. Meanwhile, other attempts have been made to apply neural

networks to BCI, such as the research in Oxford and Graz [25, 26]. However, computation of the MLP classifier is complex and runs the risk of being trapped in local optima [20]. In addition, the MLP is a universal approximator, so it tends to suffer from over training. Especially, the non-stationary nature of the EEG data will make it very difficult for the neural networks to achieve the optimal result. Therefore, neural network classifiers are not popular in BCI research and will not be discussed further in this thesis.

1.4.3.3. Linear classifiers

Linear classifiers have been used in BCI research extensively, due to their simplicity and low computational requirements [20]. Almost all the features can be used in conjunction with linear classifiers, such as phase synchronization [3], band powers [9, 31], common spatial pattern [2, 6], AR coefficients [27, 28, 31] and most of them obtained really good results.

Linear support vector machine has also been used in BCI systems as a linear classifier and also demonstrate very good performances, as shown in references [21, 30].

1.4.3.4. Bayesian statistical classifiers

Bayesian statistical classifiers form another broad kind of classifiers used in BCI research. Researchers in Berlin [34] used amplitudes at frequencies of 10 and 20 Hz and estimated a Gaussian model for each class to compute the probability of a feature associated with the class. Similarly, the researchers in the IDIAP research institute [7, 17] divided the band powers into narrow sub-band powers and used these as features and inputted them into Gaussian classifiers to achieve satisfactory results. However, in order to enhance the accuracy with non-stationary EEG features, a higher number of Gaussian prototypes were selected for each class, which increased the computation times and required some sophisticated methods to judge from results of several Gaussian prototypes.

The hidden Markov model is a type of stochastic modeling appropriate for non-stationary stochastic sequences, with statistical properties that undergo distinct random transitions among a set of stationary processes [20]. It has also been applied to BCI research [35] where it was shown that the

classification obtained using this approach is optimal at the end of the trial. This is an advantage, compared with other BCI systems where the optimal time is not known in advance.

1.4.4. Adaptive classification

A classifier trained on data from one subject will probably not work very well for a new subject, perhaps not even for a new session with the same subject. One study [16] has successfully applied adaptive classification to BCI. In order to suit a new situation, the hyperplane of a linear classifier was shifted after applying an updated retraining step. The results showed that surprisingly simple adaptive methods in combination with an offline feature selection scheme can significantly increase BCI performance. In this approach, the parameters of the classifier were flexible enough to follow the dynamic EEG data. In a study [17], the authors applied supervised online learning in the initial training phase and the Gaussian classifier was modified by recalibrating the centre and covariance of each Gaussian prototype. Essentially, the adaptive classification reduces the error (or some similar cost functions) through retraining and updating the feature space.

1.5 Major aims of this thesis

From the literature above, many kinds of classifiers have been applied to BCI systems, but most investigations have been based on specific selected datasets or typical features and classifiers. As a result, there are still outstanding questions about (1) what constitutes the best set of features to use and (2) how generalisable are these classifiers to other features or datasets. Therefore, the major aim of this thesis is to compare some of the features typically used in BCI research and to compare various classifiers described in the literature on the same dataset to enable valid comparisons to be carried out.

Chapter 2 Feature Extraction

The signal processing block is fundamental to BCI when it is viewed as a pattern recognition system. This chapter concentrates on the first part of signal processing: feature extraction. Some basic concepts about feature extraction are introduced in section 2.1. The investigated features, including time series waveform, spectral components, autoregressive (AR) components and eigenvector components, are discussed in sections 2.2 to 2.5, respectively.

2.1 What is feature and feature extraction?

It is common to classify similar objects in the same class, and two different classes can be distinguished according to their differences. In order to identify the class of a given object, it is important to extract some properties which can reflect the similarities in the same class as well as differences between classes. In pattern recognition, features are measured or derived properties from the object (or process) of interest, which contain distinctive information allowing different type of objects (or processes) to be clearly differentiated.

Different thinking activities often result in different patterns of EEG signals, but the differences between the recorded signal waveforms are not always immediately obvious on inspection. In particular, the signals of interest can be hidden in a highly noisy environment or the EEG signals may consist of a superposition of a large number of simultaneously active brain sources that are typically distorted by artifacts such as EOG and EMG. Indeed, the signal itself may not always be stable. Therefore, it is crucial for a BCI system to extract a suitable feature set which distills the required inter-class discrimination information in a manner that is robust to various contaminants and distortions.

Choosing good discriminating features is the key to any successful pattern recognition system and as discussed above, it is usually hard for a BCI system to judge a thinking activity just using raw data, which are very noisy. Therefore, the raw data must be transformed to a reduced representative set

of features for use by the classifier. The process is shown in Figure 2.1. The process of mapping the original measurements into more effective features is generally called feature extraction or feature selection [50]. A desirable characteristic of any feature set is reduced representation size, which implies low dimensional features, but the reduction must not be at the cost of removing relevant information that discriminates between the classes. In BCI research, both signal processing intuition and physiology knowledge should be involved in the feature extraction.

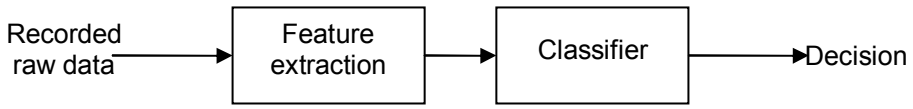

Figure 2.1. The process of pattern recognition

2.2 Time series waveform template

Recall from Chapter 1 that multi-channel EEG signals are recorded using different electrodes. The signal in each channel is recorded as a sequence of data points sampled at regular times; this is termed the time series and describes the shape and form of a signal in the time domain. Both temporal and spatial signal processing can be applied to the multi-channel time series, which contain both temporal and spatial information.

Due to experimental protocols, BCI time series usually include the idle signal and the attenuation caused by ERD and enhancement caused by ERS. As discussed earlier, different thinking activities have different effects on EEG signals and may cause different distortions to the waveform, so it is hypothesized that the differences between classes should be reflected in the shape of the time series waveforms, and hence the time series waveform was the first feature tested in this research.

The average time series waveforms of training trials in one of the datasets for each class (imagining left or right hand movement) in two channels (C3 and C4) are shown in Figure 2.2. It shows that the waveforms of both channels are similar before 4 seconds, when the subject was relaxed (more details are described in Chapter 4). However, there are some differences

between the waveforms of two thinking activities, especially in the period of 6-8 s. In the first class, i.e., imagining left hand movement, the signal in C4 is suppressed more than in C3 and the waveforms in C3 and C4 are very different, while in the second class, i.e., imagining right hand movement, the signals in both channels are similar, but there is a trend that the signal in C4 is stronger than the synchronous signal in C3 during the period 7-8 s. Using the time series waveforms to distinguish the pattern of different thinking activities relies on matching them with some extracted templates from training data; more details of how this was done will be introduced in the next chapter.

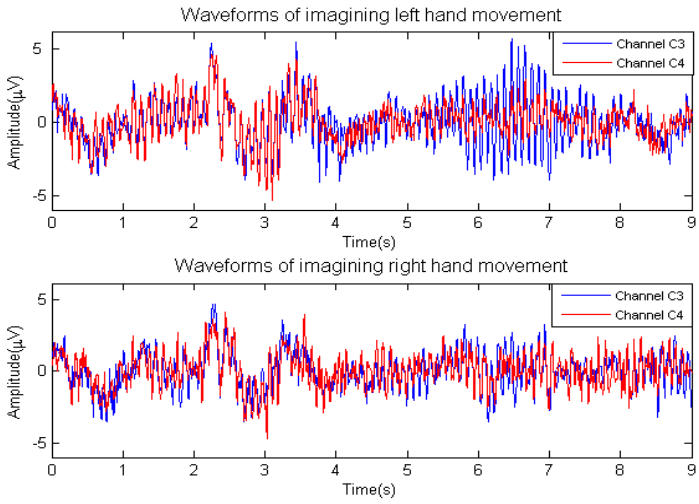

Figure 2.2. Averaged time series waveforms of two thinking activates(imaginary left/ right hand movements). The waveforms were recorded from electrodes C3 and C4 and 70 trials were included in each class.

2.3 Autoregressive components

A random time series signal $y(t)$ can often be described by an autoregressive model of order p in the following form:

$$y(t) = a_1 y(t\text{-}1) + a_2 y(t\text{-}2) + a_3 y(t\text{-}3) + \cdots + a_p y(t\text{-}p) + e(t), \qquad (2\text{-}1)$$

where $e(t)$ is a white noise process with zero mean and variance σ^2 [54]. Using more samples can provide a more accurate estimate of an AR model. In BCI research, it is supposed that different thinking processes produce different signals and the discriminative information can be captured by comparing the relevant AR model parameters.

2.3.1 AR coefficients

The AR model expresses the signal characteristics through the AR coefficients a_i, $i = 1 \cdots p$. There are several classical algorithms for estimating the AR coefficients, such as the Yule-Walker, Burg, covariance and forward-backward approaches [54]. It was found that all these algorithms provided very similar results for the EEG signals analysed, when the above different AR algorithms were used to estimate the AR parameters of a 4-th order process using 512 random data points. The average AR coefficients were [*-1.6555, 1.4132, -0.8577, 0.3450*], and the variance of these coefficients across the different models were insignificantly small, i.e., $10^{-3}\times$ [*0.1664, 0.6861, 0.5732, 0.0922*].

For multi-channel EEG data, a possible BCI system can produce AR coefficients for each channel and combine them as a feature vector. Denote the EEG signals captured from electrodes C3 and C4 as $x_{C3}(t)$ and $x_{C4}(t)$, respectively. The AR coefficients are extracted from these channels independently using the least mean square (LMS) algorithm and placed into the vectors $\boldsymbol{a}_{C3}(t)$ and $\boldsymbol{a}_{C4}(t)$, where the first element is the first coefficient and the last element the last coefficient. The two $p\times1$ vectors $\boldsymbol{a}_{C3}(t)$ and $\boldsymbol{a}_{C4}(t)$ are concatenated to form a feature vector $\boldsymbol{f}(t)$ of dimension $2p\times1$. The process is shown in Figure 2.3.

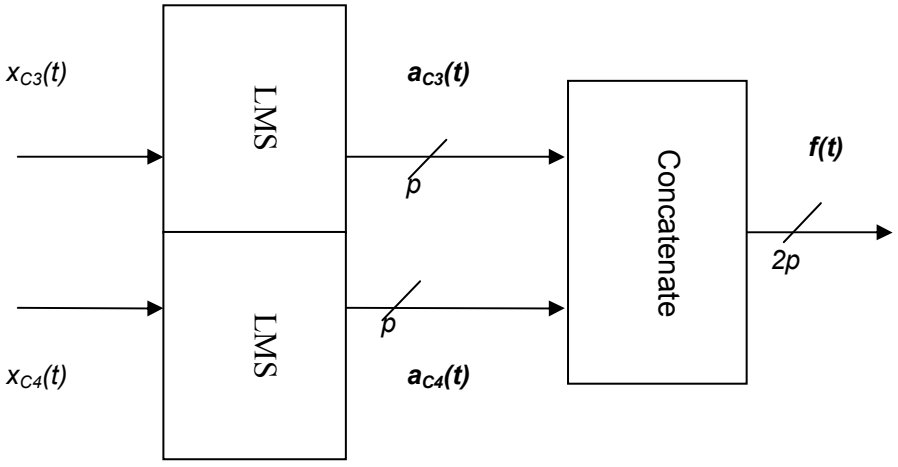

Figure 2.3. Extracting AR coefficients feature

2.3.2 AR poles

Applying the z-transform to (2-1), gives

$$Y(z) = H(z)E(z) = \frac{E(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}} \quad . \tag{2-2}$$

An AR model's transfer function contains poles in the denominator and trivial zeros at *z=0* in the numerator, so it is referred to as an "all-pole" model. The poles p_i, are obtained by solving the roots of the AR coefficient polynomial in the denominator of *H(z).* Since the AR coefficients are real, the roots must be real or occur in complex conjugate pairs. Each pair of complex conjugate poles has a one to one relationship with the AR spectral peak in the *z* domain [55], where the sharpness of the peak is determined by the distance of the poles to the unit circle (see Figure 2.4). Specifically, the closer the poles are to the unit circle, the bigger is the amplitude of the spectral peak. A *p*-th order AR model has *m* peak frequencies where $m = \frac{p}{2}$, when *p* is even and $m = \frac{P+1}{2}$, when *p* is odd. Assuming that f_s is the sampling rate of the EEG signal, the AR pole is represented by $p_i = a_i + jb_i = | p_i | e^{j\phi_i}$, the phase ϕ_i is obtained using $\phi_i = tan^{-1}(\frac{b_i}{a_i})$, the amplitude is $| p_i |$, and from reference [55] the spectral peak frequency f_i is given by

$$f_i = \frac{\phi_i}{2\pi} \times \frac{f_s}{2} \quad . \tag{2-3}$$

The amplitude and phase of the AR poles can be chosen as a feature vector, since they reflect the two main aspects of signal: intensity and frequency.

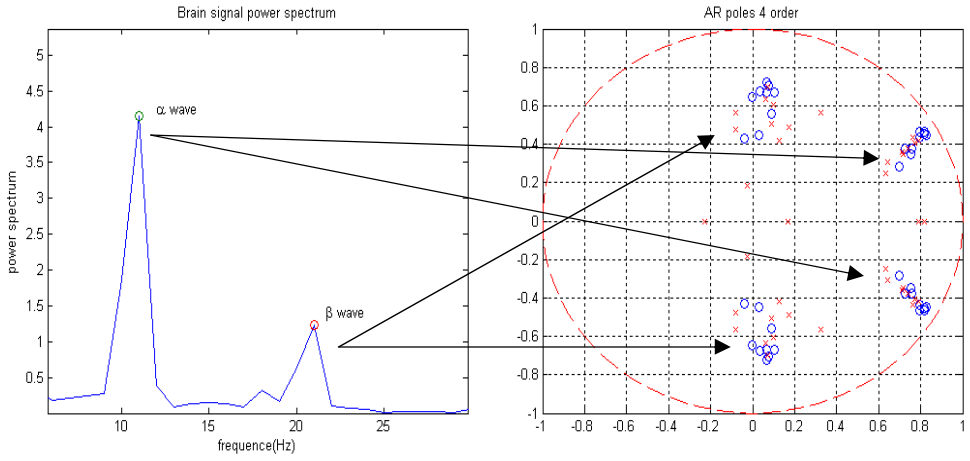

Figure 2.4. Spectral peaks and AR poles

2.3.3 Optimal AR order

Actually, before calculating the AR coefficients, the order of the model, *p*, needs to be selected. The order of the model is a trade-off between accuracy and simplicity. A higher order gives a potentially more accurate model but produces a greater number of parameters increasing the dimension of the feature space and reducing the ability of the classifier to generalize.

For a signal *y(t)* with *q* samples, the ratio, *R*, between square error and average signal power, defined as,

$$R = \frac{\sum_{t=p+1}^{q} e^2(t)}{\sum_{t=p+1}^{q} y^2(t)} \tag{2-4}$$

is a measure of how accurately the model matches the processed data. *R=1* means that the signal is a white noise process and *R=0* means the signal could be modelled exactly by an AR process. Two AR models were estimated for each of the 70 EEG time series in each class (corresponding to imaginary left or right hand movements); a separate model is estimated for each channel (C3 or C4). For the signals in the same class and the same channel, average *R* values for different AR orders are shown in Figure 2.5. If the curve has a concave shape the optimal order can be obtained as the minimum point. Figure 2.5 shows the presence of a "knee" in each curve when the order is equal to 4.

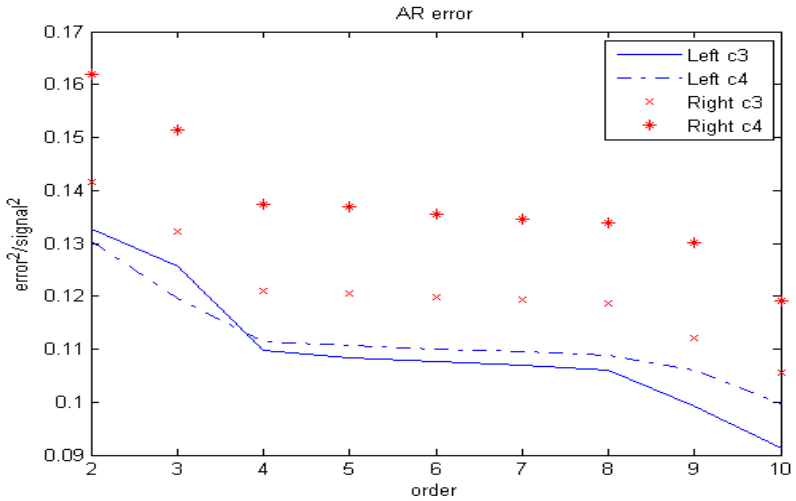

Figure 2.5. Different *R* values with different AR order in channel C3 or C4 and class of imaginary left or right hand movement

Additionally, from physiological considerations, it is known that only the brain signals in the alpha and beta bands are related to imaginary hand movements. Every brain wave has a peak value of its frequency spectrum and every spectral peak corresponds to a pair of AR poles (see Figure 2.3). Subsequently, 2 spectral peaks correspond to 2 pairs of AR poles, therefore, the optimal AR order can be chosen equal to 4. Further analysis carried out has showed that choosing a higher AR order just gave more poles located near the origin and did not improve classification accuracy.

2.4 Spectral components

The EEG time series contains all the information but also mixes the information of interest with some redundant noise. As discussed in Chapter 1, the EEG signals in channels C3 and C4 are similar to each other when the subject is relaxed, but the imaginary hand movements cause contralateral attenuations, which are reflected in significant differences between the signals in channels C3 and C4. Specifically, the alpha (7~13 Hz) and beta (14~26 Hz) bands are the two most prominent frequency bands where the modulation of signals happens. Therefore, it is instructive to analyse the EEG signals in the frequency domain. The most common way is to use the Discrete Fourier Transform (DFT) to generate the power spectrum, which gives a plot of the portion of a signal's power (energy per unit time) falling within given frequency bins, where the bin size is determined by the length of data analysed.

2.4.1 Alpha and Beta band power

The alpha and beta bands are the two most prominent frequency bands in BCI research, where the signals are modulated by thinking activities. Besides the effect from blinks and emotion, the main effect is the contralateral attenuation effect from mu and central beta rhythms. There are some common ways to obtain the band powers, such as band pass filtering and squaring the samples; using the Fast Fourier Transformation (FFT) to transform the signal weighted by a Hamming window to the frequency domain and to calculate the signal power in the frequency bands of interest. Compared to the FFT method, the first method keeps amplitude but ignores phase information.

Denoting the Fourier coefficients as $X_v=a_v+jb_v$ and the resolution of the FFT as Δf, the phase $\psi = tan^{-1}(\frac{b_v}{a_v})$ and the amplitude $A = | X_v |$. Varying the length of the Hamming window changes the frequency resolution of the amplitude and phase information obtained from the transform. For example, choosing ¼ second as the length of the Hamming window and taking the FFT of a signal with sampling rate f_s, the complex coefficient X_v would contain the phase information of every frequency bin, spaced 4 Hz apart. The power in the frequency bands of interest are given by

$$\text{Alpha band power } P_\alpha = \sum_{v\Delta f \geq 7Hz}^{v\Delta f \leq 13Hz} | X_v |^2 , \tag{2-5}$$

$$\text{Beta band power } P_\beta = \sum_{v\Delta f \geq 14Hz}^{v\Delta f \leq 26Hz} | X_v |^2 . \tag{2-6}$$

In this research, Δf was set to 0.25 Hz, i.e., data of length 4 seconds was Fourier transformed. The alpha band powers in both channels C3 and C4 for different imagined movements are shown in Figure 2.6. As shown, there is a significant difference between the alpha band powers of the two classes. Therefore, the band powers in both channels C3 and C4 can be combined to form a very low dimensional feature vector. It is reported in a study [53] that the phase information combined with band powers was used as a feature and better performance was achieved, but in our research, no improvement was found when using the phase information, so it was not used further.

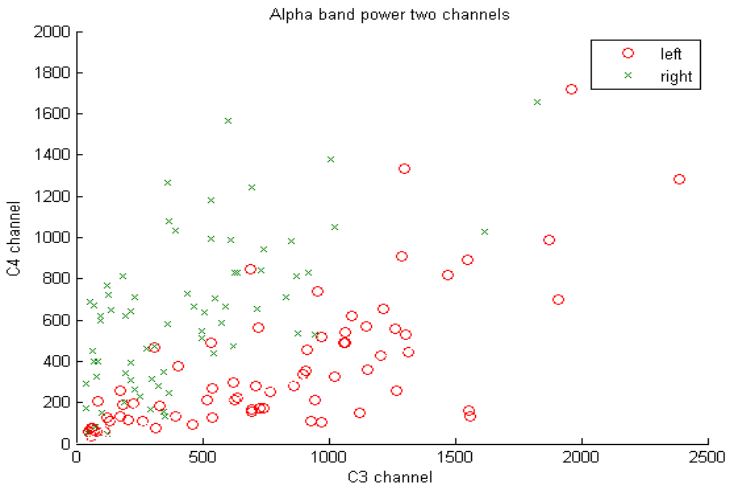

Figure 2.6. Distribution of the alpha band powers in channels C3 and C4. The red circle and green cross are symbols for alpha band powers of 70 imagined left and right hand movements respectively.

2.4.2 Spectrum density peak

In spectrum analysis, an alternative to using the total energy in selected frequency bands is to choose the peak value of power in these bands as the feature containing the discriminative information for determining motor activities. The peak value usually appears at approximately 10 Hz in the alpha band and 20 Hz in the beta band. This spectral peak for the alpha band is defined as:

$$P_{\alpha,max} = \max_{7Hz<v\Delta f<13Hz}(|X_v|^2), \tag{2-7}$$

and similarly for the beta band

$$P_{\beta,max} = \max_{14Hz<v\Delta f<26Hz}(|X_v|^2). \tag{2-8}$$

2.4.3 Asymmetry ratio

The signals in channels C3 and C4 are asymmetric during imagining hand movements, especially in the alpha and beta frequency bands, which is the main source of distinguishable information utilised in BCI. However, there is no direct explanation for the exact amount of change of signal intensities. A tentative idea is to work out whether there is a ratio, independent of signal absolute intensities, between the two channels which can indicate a corresponding task. To this end, we use the asymmetry ratio defined in [52] as:

$$R_{\alpha,asy} = \frac{(P_{\alpha,c3} - P_{\alpha,c4})}{(P_{\alpha,c3} + P_{\alpha,c4})}, \tag{2-9}$$

where $P_{\alpha,C3}$ and $P_{\alpha,C4}$ are the alpha band powers (or spectrum density peaks) in channels C3 and C4 respectively. Similarly $R_{\beta,asy}$ can be defined and used. The asymmetry ratio is based on the band power, therefore, it can be considered as a supplementary feature to the band power.

In a dataset where a subject was asked to start imagining at 3 s and each trial lasted for 9 s (more details are described in Chapter 4), the averaged asymmetry ratios of 70 trials in each class (imagining left or right hand movement) are calculated, and averaged asymmetry ratios for different classes over a period of 9 s are shown in Figure 2.7. Obviously, during the relaxation phase, the difference between asymmetry ratios of different thinking

activities is very small, but during the imagining hand movement, the asymmetry ratios can reach ±50% and the difference between classes becomes very significant.

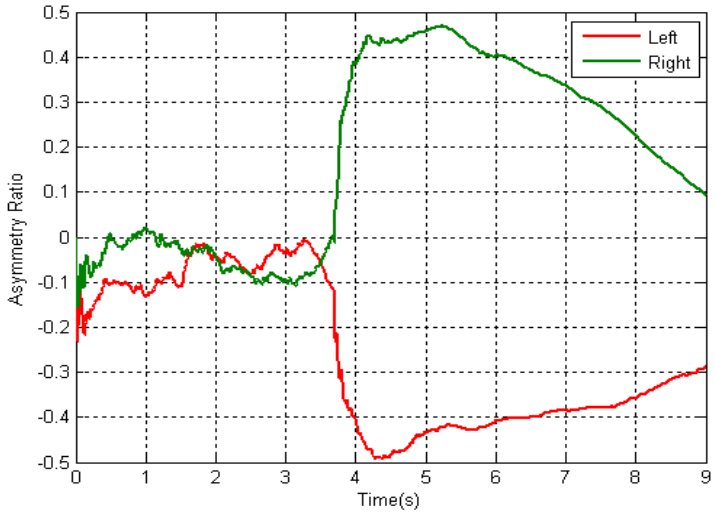

Figure 2.7. Averaged asymmetry ratios over time in different classes. Asymmetry ratios of 70 trials in each class were averaged, the red and green lines are for imagining left and right hand movement respectively.

2.5 Eigenvector elements analysis

2.5.1 Introduction

The combined time series of *m* samples from *n* electrodes can be denoted as a matrix,

$$\boldsymbol{X} = \begin{bmatrix} x_1(1) & \cdots & x_1(m) \\ \vdots & \ddots & \vdots \\ x_n(1) & \cdots & x_n(m) \end{bmatrix}. \tag{2-10}$$

$\boldsymbol{X}$ is fully characterized by its distribution function, but it is difficult to determine this distribution function. Instead of the raw time series or an estimated distribution function, it is preferable to use less complex but more computable variables such as the covariance matrix.

The covariance matrix indicates the dispersion of the signal distribution. It is defined by

$$\boldsymbol{\Sigma} = \boldsymbol{E}\{(\boldsymbol{X} - \boldsymbol{M})(\boldsymbol{X} - \boldsymbol{M})^T\} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}, \tag{2-11}$$

where $\boldsymbol{E}$ is the mathematical expectation operator and

$$\boldsymbol{M} = \begin{bmatrix} M_{\boldsymbol{1}} & \cdots & M_{1} \\ \vdots & \ddots & \vdots \\ M_{n} & \cdots & M_{\boldsymbol{n}} \end{bmatrix}, \tag{2-12}$$

where $\boldsymbol{M_i}$ is the mathematical expectation of the signal from the *i*-th electrode. In practice, $\boldsymbol{M_i}$ is estimated by averaging the signal values over *m* samples from the same electrode, i.e.

$$\hat{M}_i = \frac{1}{m}\sum_{t=1}^{m} x_i(t), \tag{2-13}$$

and

$$\hat{c}_{ij} = \frac{1}{m}\sum_{t=1}^{m} (x_i(t) - \hat{M}_i)(x_j(t) - \hat{M}_j) \ . \tag{2-14}$$

Thus, the covariance matrix is a symmetric matrix, while its diagonal components are the variances of individual random variables, and the off-diagonal components are the covariances of pairs of random variables [50]. It quantifies the extent of linear relationships between the elements. In this case, it describes any linear relationship between the signals from different electrodes, so it quantifies spatial properties of the brain signals.

Any matrix can be viewed as a linear transformation and its eigenvector has the property of only changing in scale under this transform, and the value of this change is determined by the corresponding eigenvalue. Therefore, the eigenvectors reflect the basic components in a system and the eigenvalues indicate the relative prominence of the components.

2.5.2 Principal Eigenvector

For an EEG signal $\boldsymbol{X}$ captured from *n* electrodes, its covariance matrix $\boldsymbol{\Sigma}$ is an *n* by *n* square matrix. The eigenvalues λ of this square matrix can be derived from the characteristic polynomial and the relevant eigenvector $\boldsymbol{v}$ can be obtained by solving the linear equation $(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\ \boldsymbol{v}=0$, where $\boldsymbol{I}$ is the identity matrix. Practically, the eigenvalues and eigenvectors can be estimated by QR decomposition algorithm [56].

The eigenvector, with the largest corresponding eigenvalue, is termed the principal eigenvector of $\boldsymbol{\Sigma}$. The principal eigenvector describes the dominant

spatial variation of a given thinking activity, so, a hypothesis is that captured EEG signals from different thinking activities have different principal eigenvectors but the principal eigenvectors of EEG signals from the same class are similar. Therefore, the principal eigenvector can be used as a feature and the classification can be carried out by comparing the similarity between the principal eigenvectors from the test trials and those from the training trials.

2.5.3 Common spatial pattern

In the 2-class problem of distinguishing between imaginary left and right hand movements, brain signal time series with length N samples are recorded by n electrodes. The recorded data set for each trial is denoted as the n by N matrix $\boldsymbol{X}$. For each trial, an $n \times n$ covariance matrix can be computed from the n-channel time series. All the covariance matrices obtained from trials of the same class are averaged and the result is denoted as $\boldsymbol{R}_L$ or $\boldsymbol{R}_R$, for the classes of imagined left/right hand movements, respectively. A mixture of the covariance matrices of both classes, is denoted as $\boldsymbol{R}$, which is defined by

$$\boldsymbol{R}=\boldsymbol{R}_L+\boldsymbol{R}_R. \tag{2-15}$$

Since the covariance matrices $\boldsymbol{R}_L$ or $\boldsymbol{R}_R$ are symmetric, then so is $\boldsymbol{R}$.

Common spatial pattern (CSP) [50, 51] is a technique for analysing multi-channel data belonging to 2-class problems. It is based on the K-L decomposition, which aims to project the signal onto a subspace where differences between classes are highlighted and similarities are minimized. This is achieved by a signal decomposition using an n by n matrix $\boldsymbol{W}$ to project the recorded raw signal $\boldsymbol{X}$ to $\boldsymbol{X}_{csp}$, which lives in a new space, as follows:

$$\boldsymbol{X}_{csp}=\boldsymbol{W}^T\boldsymbol{X}. \tag{2-16}$$

The matrix $\boldsymbol{W}$ is chosen such that it simultaneously diagonalizes two covariance matrices, i.e.

$$\boldsymbol{W}^T\boldsymbol{R}_L\boldsymbol{W}=\boldsymbol{\Lambda}_L \tag{2-17}$$

$$\boldsymbol{W}^T\boldsymbol{R}_R\boldsymbol{W}=\boldsymbol{\Lambda}_R \tag{2-18}$$

where the $\boldsymbol{\Lambda}_L$ and $\boldsymbol{\Lambda}_R$ are diagonal matrices. In order to find the matrix $\boldsymbol{W}$, $\boldsymbol{X}$ is first whitened by

$$\boldsymbol{Y} = \theta^{-\frac{1}{2}}\boldsymbol{\Phi}^T\boldsymbol{X}\ , \tag{2-19}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ are the eigenvalue matrix and normalised eigenvector matrix of $\boldsymbol{R}$, respectively. Then $\boldsymbol{R}$ and $\boldsymbol{R}_R$ are transformed to

$$\theta^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{R}\boldsymbol{\Phi}\theta^{-1/2} = \boldsymbol{I}\ , \tag{2-20}$$

and

$$\theta^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{R}_R\boldsymbol{\Phi}\theta^{-1/2} = \boldsymbol{K}\ , \tag{2-21}$$

where $\boldsymbol{I}$ is the identity matrix, and in general, $\boldsymbol{K}$ is not a diagonal matrix.

Now compute $\boldsymbol{\psi}$ and $\boldsymbol{\Lambda}_R$, which are the normalised eigenvector matrix and eigenvalue matrix of $\boldsymbol{K}$, respectively. It follows that,

$$\boldsymbol{\psi}^T\boldsymbol{I}\boldsymbol{\psi}=\boldsymbol{I}, \tag{2-22}$$

$$\boldsymbol{\psi}^T\boldsymbol{K}\boldsymbol{\psi}=\boldsymbol{\Lambda}_R. \tag{2-23}$$

Finally, the matrix $\boldsymbol{W}$ is equal to $\boldsymbol{\Psi}^T\theta^{-1/2}\boldsymbol{\Phi}^T$. In this process, the matrix $\boldsymbol{R}_L$ is diagonalised because

$$\boldsymbol{\Psi}^T\theta^{-1/2}\boldsymbol{\Phi}^T\boldsymbol{R}_L\boldsymbol{\Phi}\theta^{-1/2}\boldsymbol{\Psi} = \boldsymbol{\Psi}^T\theta^{-1/2}\boldsymbol{\Phi}^T(\boldsymbol{R}-\boldsymbol{R}_R)\boldsymbol{\Phi}\theta^{-1/2}\boldsymbol{\Psi} = \boldsymbol{I} - \boldsymbol{\Lambda}_R = \boldsymbol{\Lambda}_L \tag{2-24}$$

It is derived in reference [50] that the matrix $\boldsymbol{\Psi}^T\theta^{-1/2}\boldsymbol{\Phi}^T$ is the eigenvector matrix of $\boldsymbol{R}^{-1}\boldsymbol{R}_R$. In practice, the series of computations to obtain $\boldsymbol{W}$ can be replaced by solving the generalized eigenvalue problem for matrices $\boldsymbol{R}$ and $\boldsymbol{R}_R$; that is, solving the problem $\boldsymbol{Rv}=\lambda\boldsymbol{R}_R\boldsymbol{v}$, which can be estimated by a QZ decomposition [56].

In summary, the covariance matrices for the transformed observations satisfy:

$$\boldsymbol{\Lambda}_L = \boldsymbol{W}^T\boldsymbol{R}_L\boldsymbol{W}, \tag{2-25}$$

$$\boldsymbol{\Lambda}_R = \boldsymbol{W}^T\boldsymbol{R}_R\boldsymbol{W}, \tag{2-26}$$

$$\boldsymbol{\Lambda}_L+\boldsymbol{\Lambda}_R=\boldsymbol{I}. \tag{2-27}$$

After the above transformations, recorded datasets corresponding to these two classes can effectively be distinguished by eigenvalues contained in the matrices $\boldsymbol{\Lambda}_L$ and $\boldsymbol{\Lambda}_R$. This is evident since the individual eigenvalues λ_{Lj} and λ_{Rj}, inside the matrices $\boldsymbol{\Lambda}_L$ and $\boldsymbol{\Lambda}_R$, satisfy,

$$\lambda_{Lj}+\lambda_{Rj}=1 \quad j= 1, 2\ldots n. \tag{2-28}$$

Through the CSP transform, $\boldsymbol{X}_{csp}=\boldsymbol{W}^T\boldsymbol{X}$, the variance of the spatially filtered signal is maximized for one class while it is minimized for the other class. As discussed above, the eigenvector with the largest eigenvalue for class 1 would correspond to the smallest eigenvalue for class 2 and vice versa.

The two classes do not share common important features, and different distributions of features are obtained for different classes. For a signal $\boldsymbol{X}$ of unknown class label, we can apply the CSP transform, with the transformation matrix $\boldsymbol{W}$ obtained from the training set. The diagonal elements of the transformed data's covariance matrix $\boldsymbol{\Sigma}_{csp}=\boldsymbol{W}^T\boldsymbol{\Sigma}\boldsymbol{W}$ can be used as features or further features can be obtained from the new signal $\boldsymbol{X}_{csp}=\boldsymbol{W}^T\boldsymbol{X}$ by other feature extraction methods, such the calculation of the alpha band power. For an example, in one of the datasets (more details are described in Chapter 5), the diagonal elements $\boldsymbol{\Sigma}_{CSP}(1,1)$ and $\boldsymbol{\Sigma}_{CSP}(3,3)$ are combined to form a two-dimensional feature. The scatterplot from these features is shown in Figure 2.8.

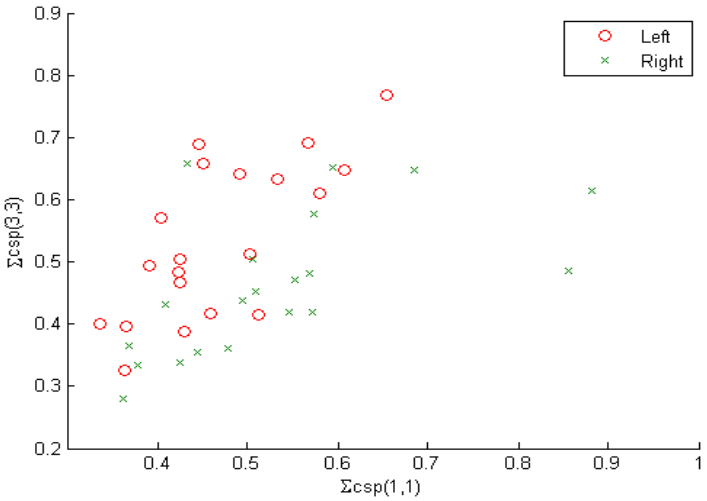

Figure 2.8. Distribution of CSP features. The red circle and green cross are symbols for CSP features of 20 imaginary left and right hand movements respectively.

A limitation of the CSP feature is that it not suitable to multi-class problems. This is not an issue in this thesis as the main focus is on a 2-class problem. However, a generalization to the multi-class problem can be addressed by considering all possible pairs of 2-class problems separately [33], and their results fused by various voting schemes to form a final decision.

Chapter 3 Classification

This chapter provides a detailed description of the core part of a Brain Computer Interface: the classifier. The role of the classifier in the BCI system is to identify a subject's intentions from a finite number of predefined choices. Following feature extraction, a suitable classifier needs to be designed. This is usually achieved through machine training, which utilises part of the data and their known corresponding labels. This information is used to train the classifier into recognising how the different categories are distributed throughout feature space. In our approach, features extracted from the BCI signals were divided randomly into two parts, one for training and the other for assessing the classification accuracy. Generally, it is assumed that the training and test data have similar properties and underlying distributions, which is a fundamental precondition of feasibility in classification.

In order to establish an effective classification process within wider BCI systems, an investigation was performed to test several types of classification algorithms. The classification methods that have been used in this project are template matching, nearest neighbour, linear discriminant analysis, Bayes statistical classifier and fuzzy logic decision classifiers; details will be described in the subsequent sections. Finally, some attempts to improve each classifier are also described in this chapter.

3.1 Template matching

An EEG signal time series waveform describes the shape and form of a signal in the time domain. It is assumed that the EEG time series contains sufficient information for discriminating between different thinking activities. The primary idea is to survey the similarity between training time series and test time series, and the test time series data is classified into a class whose member waveforms bear the greatest similarity with it.

A measure of the similarity between two time series is the linear cross correlation between them. For multi-channel EEG data, signals captured from

electrodes C3 and C4 are denoted as $x_{C3}(t)$ and $x_{C4}(t)$, respectively; when concatenated together, the trial is marked as

$$\boldsymbol{X} = \begin{bmatrix} x_{C3}(1) & x_{C3}(2) & \cdots & x_{C3}(t) \\ x_{C4}(1) & x_{C4}(2) & \cdots & x_{C4}(t) \end{bmatrix}, \tag{3-1}$$

Similarly, another trial can be marked as

$$\boldsymbol{Y} = \begin{bmatrix} y_{C3}(1) & y_{C3}(2) & \cdots & y_{C3}(t) \\ y_{C4}(1) & y_{C4}(2) & \cdots & y_{C4}(t) \end{bmatrix}. \tag{3-2}$$

The correlation coefficients of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are expressed by a 2 by 2 matrix $\boldsymbol{R_{xy}}$. The signal in each channel is used as a variable and $\boldsymbol{R_{xy}(i,j)}$ stands for the correlation coefficient of signals from the i-th channel in $\boldsymbol{X}$, $\boldsymbol{x_i}$ and the j-th channel in $\boldsymbol{Y}$, $\boldsymbol{y_j}$. The matrix $\boldsymbol{R_{xy}}$ is derived from the covariance matrix $\boldsymbol{C_{xy}}$, which returns a 2-by-2 matrix containing the estimated pairwise covariance coefficient between each pair of rows in $\boldsymbol{X}$ and $\boldsymbol{Y}$. Mathematically,

$$\boldsymbol{C_{xy}}(i,j)=\boldsymbol{E}\{(\boldsymbol{x_i}-M_{xi})(\boldsymbol{y_j}-M_{yj})\}, \tag{3-3}$$

where $\boldsymbol{E}$ is the mathematical expectation operator and M_{xi} and M_{yj} are the means of $\boldsymbol{x_i}$ and $\boldsymbol{y_j}$, respectively and

$$\boldsymbol{R_{xy}}(i,j) = \frac{\boldsymbol{C_{xy}}(i,j)}{\sqrt{\boldsymbol{C_{xy}}(i,i)\boldsymbol{C_{xy}}(j,j)}}. \tag{3-4}$$

Therefore, the diagonal elements of correlation matrix represent the similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For example, $\boldsymbol{R_{xy}(1,1)}$ shows the similarity between the time series x_{C3} and y_{C3}.

3.1.1 Build the template for each class

The template matching approach relies on comparing the similarity between a test signal and a set of template waveforms in each class to judge to which class a signal belongs. Therefore, it is important to build a suitable database of template waveforms for each class. Three methods were considered.

(a): The simplest way assumes that a given class of thinking processes, such as an imaginary left hand movement, follows a single template in the time domain and different templates exist for the other classes. To construct the single representative template for one class, all the waveforms of the training trials in the same class are averaged. If there are m trials in a class and the trials are recorded as $\boldsymbol{X_i}$ (i=1,2,⋯,m), then the template of the class is

$$\boldsymbol{T}=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{X}_i\,, \tag{3-5}$$

In this way, the common (average) properties in each class are obtained and utilised but the differences amongst trials in the same class (i.e. intra-class variations) are averaged out. This way of template estimation has a short computation time but, as will be shown in Chapter 4, produces low accuracy. This can be attributed to the fact that, in reality, real thinking processes are more complicated and greater dynamics exist than can be represented by a simple template waveform. Differences exist between different subjects' EEG signals; even the same subject will, on different occasions, be influenced by the environment and some factors such as emotion and fatigue.

(b): The second idea is a combination of template matching and the nearest neighbour method. It considers that every trial in the training set is potentially useful as an individual template, and some similarities should exist amongst trials in the same class. When the classifier is in use, the correlations between the waveform being tested and every waveform in the training database are calculated and a decision is made according to the nearest neighbour rule (see next section). This method utilises information from all the individual training trials and, as will be shown in Chapter 4, provides better results than the single template classifier. However, the computation time is much longer, because many more correlation coefficients need to be calculated and compared.

(c): Instead of simple averaging or having highly redundant templates, the third way is a trade off between accuracy and computation time. In such an approach, the templates are built according to the signal amplitude level of EEG signals in pre-stimulus brain state. Usually, subjects are supposed to be relaxed before imagining given activities, and it is found that, during this period, the signal amplitudes vary amongst different trials. It has been demonstrated in a number of studies that the differences in pre-stimulus brain state from trial to trial do influence the subsequent response [61]. Observations of the signal waveforms of training trials show that, when signals from different trials but from the same class have similar signal intensities in the pre-stimulus brain state, subsequent ensuing variations of time series are

very similar. So, using the signal intensity in the pre-stimulus brain state as a measured quantity, the training trials can be divided into several subgroups. All the training trials in each subgroup are averaged and the result is used as the template for this subgroup. For example, the alpha band filtered time series in channels C3 and C4 were divided into 4 subgroups according to their pre-stimulus signal intensity levels and Figures 3.1 and 3.2 show the templates in each channel for different subgroups of each class, respectively. Then, the correlations between the test waveform and every template in the same class are calculated and, similarly to approach (b), the largest correlation is used as the correlation between the test waveform and the class.

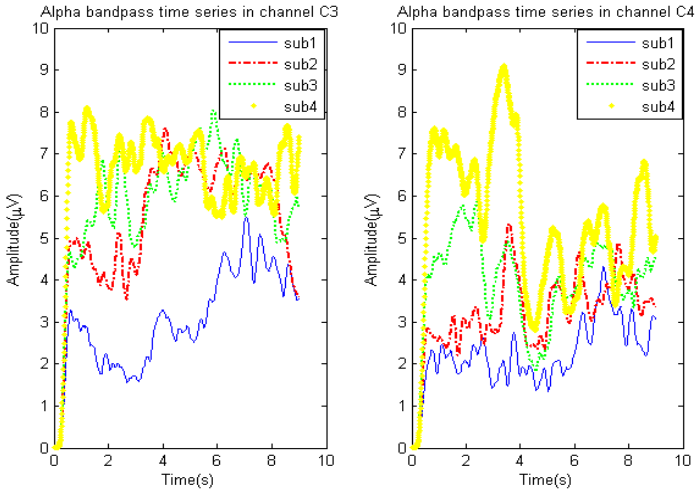

Figure 3.1. Averaged time courses (imaginary left hand movement) of the absolute amplitudes. Data are shown for alpha band (7-13 Hz) at electrode C3 and C4. Each of the four subgroups showing different pre-stimulus activity is averaged over 17 trials.

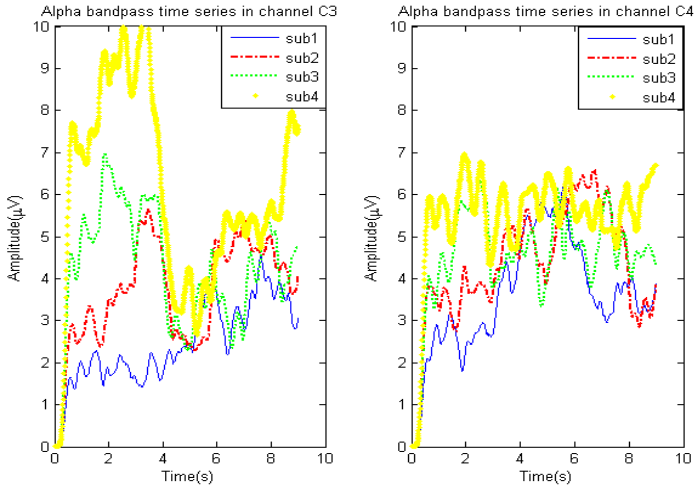

Figure 3.2. Averaged time courses (imaginary right hand movement) of the absolute amplitudes. Data are shown for alpha band (7-13 Hz) at electrode C3 and C4. Each of the four subgroups showing different pre-stimulus activity is averaged over 17 trials.

3.1.2 Using cross correlation sequence to improve classification

The amplitude of the EEG signal at an electrode of interest (such as C3 or C4) is usually obtained by comparing the signal to an EEG reference signal, for example, the signal at electrode Fz. This implies that the sign of EEG signal can be positive or negative and the correlation coefficients between two signals may be negative. The range of correlation coefficients is [-1, 1], the correlation equal to 1 is in the case of identical signals, -1 in the case of the same signal but with exactly opposite phase, and values between -1 and 1 indicates the degree of correlation. Thus the relevant parameter of interest is the absolute value of the correlation coefficient.

Due to the variable human response time, there might be time delays between signals recorded from different trials resulting in a small correlation coefficient between two very similar EEG time series. Recall that the correlation coefficient is the zero-th lag of the cross correlation sequence and this arbitrary time offset can be compensated for by calculating the cross correlation sequence for all the possible lags and finding the lag with the highest absolute value of the cross correlation sequence. The normalized cross-correlation sequence with *n*-th lags is

$$\boldsymbol{R}_{xy}(n) = \frac{\boldsymbol{E}\{\boldsymbol{x}_{i+n}\boldsymbol{y}_i\}}{\sqrt{\boldsymbol{E}\{\boldsymbol{x}_{i+n}\}\boldsymbol{E}\{\boldsymbol{y}_i\}}} \tag{3-6}$$

and for the best correlation

$$R_m = \max_n [|\boldsymbol{R}_{xy}(n)|] \,. \tag{3-7}$$

As will be shown in Chapter 4, this process enhances the classification accuracy but also increases the computation time.

3.2 Nearest neighbour classifier

In feature space, features corresponding to the different classes usually form separate clusters. Therefore, close neighbours in feature space are likely to have similar properties and belong to the same class. The nearest neighbour classifier is based on this principle. The training data is used to populate the hypothesized clusters of the training data in feature space and the distance of a test sample from every training sample is calculated. As the

labels of the training trials are assumed to be known, the test sample is assigned to the class to which the closest training sample belongs. The metrics used in this research were Euclidean, Manhattan and Mahalanobis distances. If an *n*-dimensional testing data is denoted as $\boldsymbol{X}=(x_1,x_2,\cdots x_n)^T$ and a training data is $\boldsymbol{Q}=(q_1,q_2,\cdots q_n)^T$, then the Euclidean distance is defined as:

$$D_e = \sqrt{\sum_{i=1}^{n}(x_i - q_i)^2}\ , \qquad (3-8)$$

the Manhattan distance is

$$D_m = \sum_{i=1}^{n} | x_i - q_i |, \qquad (3-9)$$

and the Mahalanobis distance is

$$D_M = \sqrt{(\boldsymbol{X}-\boldsymbol{Q})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{Q})}\ , \qquad (3-10)$$

where Σ is the covariance matrix of vector $\boldsymbol{X}$ which is estimated by

$$\boldsymbol{\Sigma} = \frac{1}{m-1}\sum_{p=1}^{m}(\boldsymbol{X}_p - \bar{\boldsymbol{X}})(\boldsymbol{X}_p - \bar{\boldsymbol{X}})^T. \qquad (3-11)$$

where *m* is the number of samples and $\bar{X}$ is the sample mean of the feature $\boldsymbol{X}_p$.

3.2.1 The nearest neighbour

In the first application of the nearest neighbour technique to BCI, only the closest feature in the training set was considered. For instance, the alpha band powers in channels C3 and C4 were combined and used as a 2-dimensional feature vector and the test data was classified as belonging to that class of the training sample with the shortest distance. Mathematically, denoting *c* classes of thinking activities as $\{\omega_1, \omega_2,\ldots,\omega_c\}$, and there are N_i training samples in *i-th* class. Taking the Euclidean distance as an example, for a test data ***X,*** the nearest neighbour discriminant function for class ω_i is expressed as*:*

$$g_i(\boldsymbol{X}) = \min_k(D_{ek})\ ,\ \ k=1,2,\cdots,N_i, \qquad (3-12)$$

where D_{ek} is the Euclidean distance between the test feature and the *k*-th

training sample in class ωi. The decision rule is that if

$$g_j(\boldsymbol{X}) = \min_i [g_i(\boldsymbol{X})] \text{ , } i=1,2,\cdots,c, \tag{3-13}$$

then $\boldsymbol{X}$ belongs to class ω_j.

3.2.2 *K*-nearest neighbour

With a limited number of training trials, only limited training samples are available to populate the feature space, and some of them may be polluted by noise and artifacts. So, the estimated feature space may be biased and errors in the recorded data may seriously influence the classification result, especially if the erroneous data happens to be in a position where some samples under test appear often. However, the chance of several similar errors occurring together is much lower. Therefore, instead of using only the closest sample, several (k) closest samples can be considered and the classifier is termed as: k-nearest neighbour. Again, the labels of the training data are assumed to be known, the test sample is classified to the class relevant to the most common label in the k closest training samples – a voting scheme is used to decide between competing choices and there are a variety of voting schemes that can be used. In order to avoid tied votes for 2-class problems, k is usually chosen to be an odd number.

A practical drawback of using a large k is that in certain regions of feature space, the density of training data may be so low that it is difficult to generalize the classifier decision, which could be easily changed if additional training data were available and the decision may be dominated by some non-representative but frequently occurring training samples.

3.2.3 Fast nearest neighbour algorithm

During the classification, in order to get high accuracy, distances of the test sample from every training sample are calculated, so, the computation time is very long. But in this process, some training samples which are far away from the test sample, have little effect on the classification result. If such samples could be excluded from the candidate feature space, the number of distance calculations is reduced and the computation time is shorter, without

incurring any loss in accuracy. Based on the idea above, feature space can be divided into several equally sized clusters. The distance from the feature under test to a representative sample for each cluster, such as the mean of the samples, is then calculated. Then the closest and second closest cluster would be selected and merged as the new feature space, where the nearest neighbour algorithm is carried out. A similar method of using subset is introduced in [63] and there is another method, the condensed nearest neighbour [62], which removes samples far away from decision boundaries.

3.3 Linear discriminant analysis classifier

Linear classifiers address a simple situation of where two classes are linearly separable so a linear discrimination function can be used to distinguish between the classes. Fisher's linear discriminant analysis (LDA) uses a plane or hyperplane (for high dimensional features) to separate the features representing two different classes. The decision hyperplane can be represented as:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0, \tag{3-14}$$

where $\boldsymbol{w}$ is known as the weight vector, $\boldsymbol{x}$ is the input feature vector and w_0 a predetermined threshold. The feature is assigned to one class or the other depending on the sign of $g(\boldsymbol{x})$. For multi-class problems, multiple planes or hyperplanes can be used together. There are many classical methods to compute $\boldsymbol{w}$ and w_0, such as the perceptron method, least squares methods [20], and these are not described in detail here. The advantages of the linear method are simplicity and low computational requirements. The classification computation can usually be finished in several steps of matrix calculations on a computer. Therefore, it is a good choice for online BCI systems, where a rapid response from limited computational resources is required.

3.3.1 LDA classifier in BCI

In current BCI research, most problems belong to the 2 class category and multi-class problems are dealt with by first transforming them into several 2 class sub-problems. For example, the most common and popular case is to

identify imagined left or right hand movement. Figure 3.3 shows how LDA is used for the alpha band power feature.

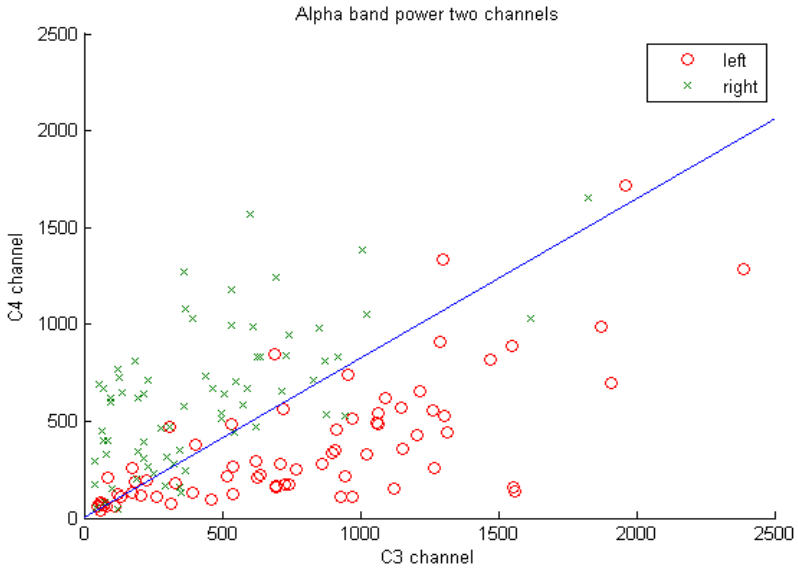

Figure 3.3. Using LDA for the alpha band power in channels C3 and C4. The red circle and green cross are symbols for alpha band powers for imagined left hand and right hand movements respectively. The straight line is an LDA classifier.

3.3.2 Improved LDA

Figure 3.3 shows the LDA can provide a satisfactory classification result, but obviously, there are still some incorrectly classified samples due to the hard border of LDA. An improved idea is to set up a piecewise linear relation to solve the problem where the feature space is split into several subsections and a separate LDA classifier is trained for each subsection. This way results in flexible multi-LDA classifiers. Every classifier works in a given area, such as a given signal intensity range in channel C3, and achieves locally optimal classification. A disadvantage of this method is that a sufficient number of samples are required in each subsection; otherwise, the piecewise classifier would not be general enough.

If the feature space can be divided into many tiny subsections and with enough samples in each, the piecewise LDA classifiers can approximate any smooth curve in the feature space. Instead of incorporating many piecewise LDA classifiers, the quadratic discriminant function can be used to estimate a curving classifier, for example:

$$g(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \quad (3\text{-}15)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and estimated covariance matrix for the *i*-th class

respectively. The test samples are assigned to one class or the other depending on the sign of $g(\boldsymbol{x})$.

3.4 Bayesian statistical classifier

The principle underlying the Bayesian statistical classifier is to calculate and compare probabilities of an observed feature $\boldsymbol{x}$ given statistical models for different classes y and assign $\boldsymbol{x}$ to the class associated with the highest probability. In the case of this BCI research, there are two classes corresponding to imaginary left or right hand movements, so $y \in \{L, R\}$. It is assumed that the two classes have equal *a priori* probabilities, i.e., $P(y) = 0.5$. The distribution of features for each class is assumed to be a mixture of the same number of Gaussian prototypes. Figure 3.4 shows a simple example to explain how to estimate the prototypes for each class. Closely spread samples in the same class are clustered together and assumed to belong to the same Gaussian distribution. Then, parameters of the distribution are estimated from these samples.

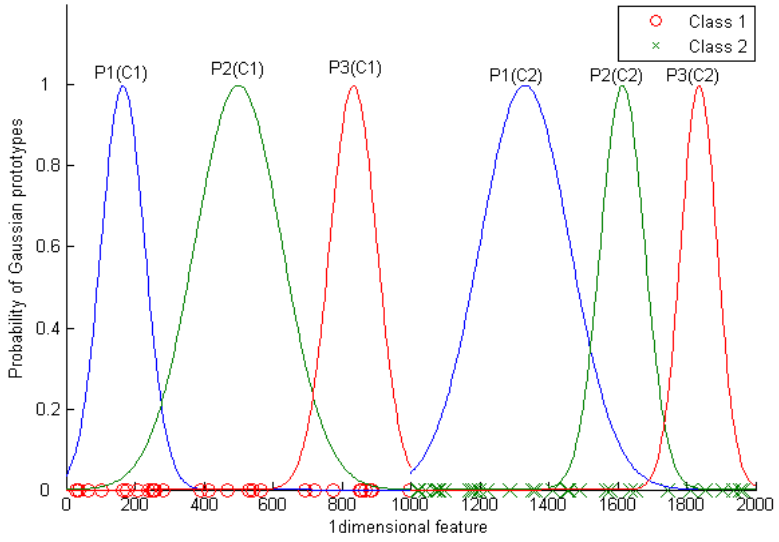

Figure 3.4. The same number of Gaussian prototypes for different classes. The red circle and green cross are symbols for the features in two classes respectively, while 3 Gaussian prototypes are estimated for each class.

Generally, using more prototypes can provide more accurate estimation of the underlying statistical model but requires an increased number of training samples and increases the computational load. The number of prototypes was

decided by comparing the accuracies of classification with different numbers of prototypes. For example, for the Graz dataset (more details are described in Chapter 4), classifications with 3, 4 and 5 prototypes were tested separately. Whilst the results were similar, classification with 4 prototypes in each class was the best.

There are two ways of estimating the Gaussian prototypes, the first is to divide the feature space evenly and use the statistical properties of samples (mean and variance) in each area to estimate the corresponding prototype. The second method uses a Gaussian Mixture Models (GMM). When the first procedure was applied to this BCI problem, the covariance between two channels was ignored and the two channels were considered as independent. This was justified since correlations between them did not influence the classification significantly. The probability density function (pdf) of the *n*-dimensional feature vector $\boldsymbol{x}$, conditioned on it being in the *i*-th prototype in class *y* was calculated by the following formula:

$$P(\boldsymbol{x}|c_i) = \frac{|\boldsymbol{\Sigma}_i|^{-1/2}}{(2\pi)^{n/2}} e^{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}, \quad (3\text{-}16)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the individual means and the estimated covariance matrix for the *i*-th prototype in class *y*. The probability density function of the *n*-dimensional feature vector $\boldsymbol{x}$, conditioned on it being in class *y*, is denoted $P(\boldsymbol{x}|y)$ and can be obtained by mixing the probabilities of $\boldsymbol{x}$ occurring in the underlying Gaussian prototypes $P(\boldsymbol{x}|c_i)$.

$$P(\boldsymbol{x}|y) = \sum_{i=1}^{M} w_i P(\boldsymbol{x}|c_i), \quad (3\text{-}17)$$

where M is the number of prototypes and w_i is the weight of each prototype and more details about it are described in the following sections. The classification can be achieved by computing the probability of a certain class given a data $\boldsymbol{x}$ using Bayes' theorem:

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})} = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x}|L)P(L) + P(\boldsymbol{x}|R)P(R)} = \frac{P(\boldsymbol{x}|y)}{P(\boldsymbol{x}|L) + P(\boldsymbol{x}|R)} \quad (3\text{-}18)$$

since $P(L) = P(R) = P(y) = \frac{1}{2}$.

3.4.1 Gaussian models from evenly divided feature space

The first method divides the feature space into several equally sized regions and assumes several Gaussian prototypes in the same class y are weighted equally. For each cluster c_i, the training samples in that cluster are used to estimate the parameters of the Gaussian prototype. After calculating $P(\boldsymbol{x}|c_i)$ according to (3-16) the desired probability $P(\boldsymbol{x}|y)$ is estimated using either

$$P(\boldsymbol{x}|y) = \max_i [P(\boldsymbol{x}|c_i)], \tag{3-19}$$

or

$$P(\boldsymbol{x}|y) = \frac{1}{M}\sum_{i=1}^{M} P(\boldsymbol{x}|c_i). \tag{3-20}$$

Taking the alpha band power feature in channels C3 and C4 as an example, as shown in Figure 3.5, the two-dimensional feature space is divided evenly. Then, a Gaussian prototype is set up and its parameters are obtained from samples in the corresponding sub-region. Every Gaussian prototype has equal weight and only works in its local area.

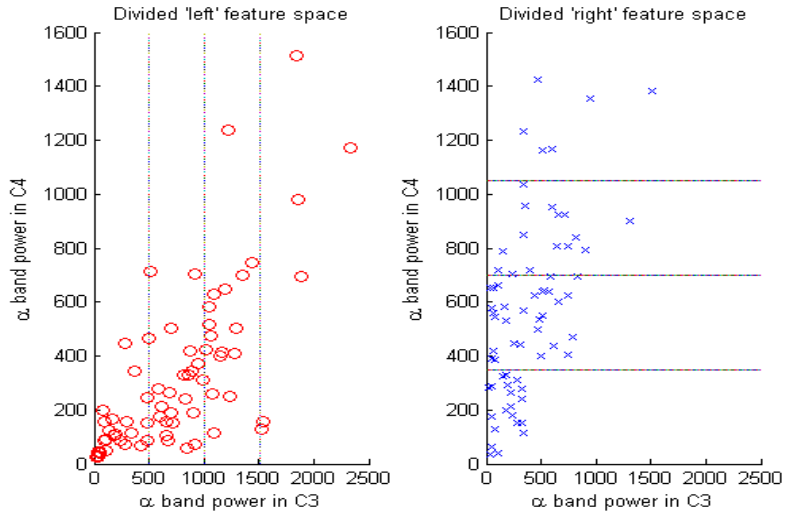

Figure 3.5. Evenly divided feature space. The two sub-figures are alpha band powers for imagined left hand and right hand movements respectively. The feature space is divided by the lines and all the parameters are estimated in a sub-region.

3.4.2 Gaussian models from EM algorithm

The second method, Gaussian mixture models (GMM) uses a different *a priori* weight $P(c_i)$ for each prototype. Gaussian mixture models consist of a set of local Gaussian prototype density functions, and an integrating network. There may be some overlaps amongst the M local Gaussian prototypes and

the joint probability of $\boldsymbol{x}$ with each class $P(\boldsymbol{x},y)$ is a weighted sum over all the M underlying Gaussian prototypes in class y. It is given by following equations:

$$P(\boldsymbol{x},y)=\sum_{i=1}^{M}P(\boldsymbol{x},c_i)=\sum_{i=1}^{M}w_iP(\boldsymbol{x}\mid c_i), \tag{3-21}$$

where

$$P(\boldsymbol{x}|c_i)=\frac{|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}, \tag{3-22}$$

and

$$\sum_{i=1}^{M}w_i=1, \tag{3-23}$$

where the quantities $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and w_i are the individual means, the covariance matrices and the *a priori* probability of i-th prototype in class y respectively.

All the parameters of each prototype are estimated from training samples in each cluster and clusters can be readily formed using a method with computational simplicity, for example, the well-known k-means algorithm gives compact clusters through minimizing the sum of point-to-centroid (representative) distances, summed over all k clusters [20]. An example of the initial clusters divided by the k-means algorithm is shown in Figure 3.6.

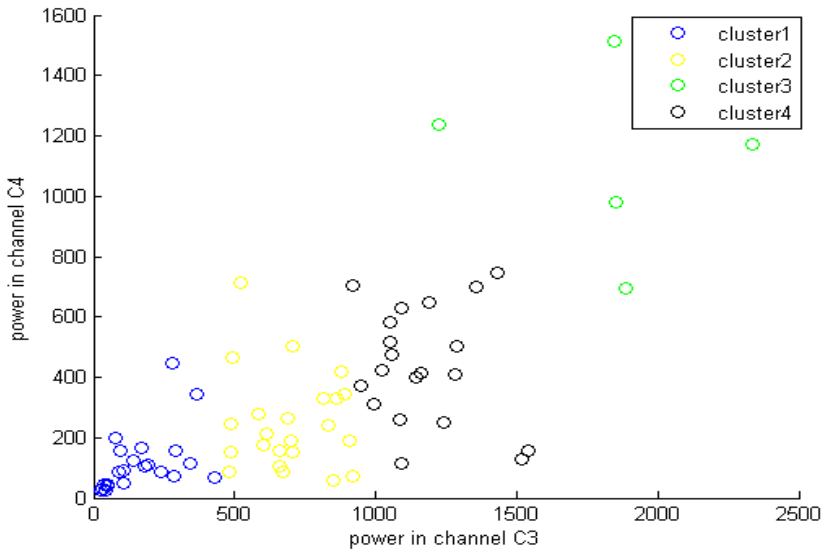

Figure 3.6. Feature space is divided by k-means algorithm. Alpha band powers of imagined left hand movements are divided into 4 subgroups.

Then, the expectation maximization (EM) algorithm [57, 65], which maximizes the likelihood of the training set generated by the estimated pdf, can be used to estimate the parameters of each Gaussian model. Each

iteration of the EM algorithm consists of two steps: E step and M step. In the *t*-th iteration, the E step computes an expectation of the likelihood function with the estimated parameters $\theta^{(t)}$ in the last M step, whilst the M step calculates the new parameters $\theta^{(t+1)}$ to maximize the likelihood function.

Assuming there are *N* training samples in a class *y* and N_i of them are used to estimate the *i*-th Gaussian prototype, the process can be described by following equations:

The initial weight is set as $w_i(0) = \frac{N_i}{N}$, in the *t*-th iteration,

$$\tau_{ip}(t) = \frac{P(\boldsymbol{x}_p, c_i)}{P(\boldsymbol{x}_p, y)} = \frac{w_i(t)P(\boldsymbol{x}_p \mid c_i)}{\sum_{j=1}^{M} w_j(t)P(\boldsymbol{x}_p \mid c_j)} \tag{3-24}$$

is the probability of the *i*-th Gaussian prototype gives to $\boldsymbol{x}_p$ of being labelled as in class y, where $\boldsymbol{x}_p$ is the *p*-th sample in the training set. Then the $\tau_{ip}(t)$ is used to estimate the new weights $w_i(t+1)$, means $\boldsymbol{\mu}_i(t+1)$ and covariance $\boldsymbol{\Sigma}_i(t+1)$ for the *i*-th Gaussian prototype according to following formulas:

$$w_i(t+1) = \frac{1}{N_i}\sum_{p=1}^{N} \tau_{ip}(t), \tag{3-25}$$

$$\mu_i(t+1) = \frac{1}{N_i w_i(t)}\sum_{p=1}^{N_i} \tau_{ip}(t)\boldsymbol{x}_p, \tag{3-26}$$

$$\boldsymbol{\Sigma}_i(t+1) = \frac{1}{N_i\, w_i(t)}\sum_{p=1}^{N_c} \tau_{ip}(t)(\boldsymbol{x}_p - \mu_i(t))(\boldsymbol{x}_p - \mu_i(t))^T. \tag{3-27}$$

The new weights, means, and covariance matrices are then used in (3-22) and (3-24) to estimate the new τ_{ip}. The likelihood function is updated in the following way:

$$L(t) = \sum_{p=1}^{N}\sum_{i=1}^{M} \tau_{ip}(t)\{\log w_i(t) - \frac{1}{2}\log[(2\pi)^n|\boldsymbol{\Sigma}_i(t)|] - \frac{1}{2}[\boldsymbol{x}_p - \boldsymbol{\mu}_i(t)]^T \boldsymbol{\Sigma}_i^{-1}[\boldsymbol{x}_p - \boldsymbol{\mu}_i(t)]\}. \tag{3-28}$$

The iteration will not stop until $\Delta L = L(t+1) - L(t) < \omega_0$, where the ω_0 is a given threshold or the number of iteration reaches a specified value. Figure 3.7 shows an example of Gaussian mixture models estimated by the EM algorithm.

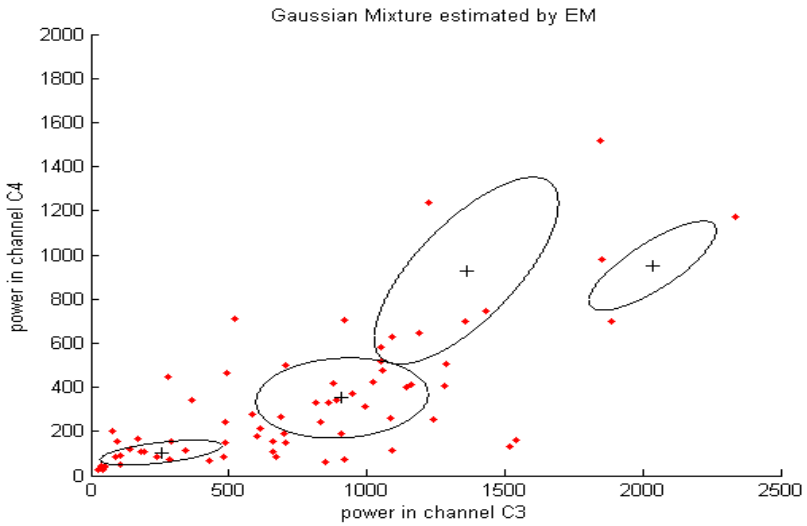

Figure 3.7. Estimated clusters of the alpha band power feature using the EM algorithm. The red dots stand for alpha band powers for imagined left hand movements, the '+' denotes the centre and the ellipse the variance of each cluster.

3.5 Fuzzy Logic classifier

As discussed in Chapter 1, the EEG signals in channels C3 and C4 are similar to each other when the subject is relaxed, but imaginary hand movements cause contralateral attenuations, which give rise to significant differences between the signals in channels C3 and C4. With this important piece of physiological knowledge, the classification can sometimes be achieved by simple intuitive measures. For example, intuitively, if the signal in C3 is quite strong but the signal in C4 is weak, it is probable that the subject is imagining a left hand movement. In this example, instead of a precise number, the concepts 'strong' and 'weak' are used, where the description is imprecise but more general and simple.

In the above example, the description of a certain signal, such as 'strong', leaves an uncertainty of signal intensity. Fuzzy logic classifier is based on the idea of achieving classification by an approach using intuitive natural language descriptions rather than precise numerical values. A fuzzy logic classifier can be considered as a system with several inputs and one output and fuzzy logic is used to formulate the mapping from given inputs to output. This process is shown in Figure 3.8.

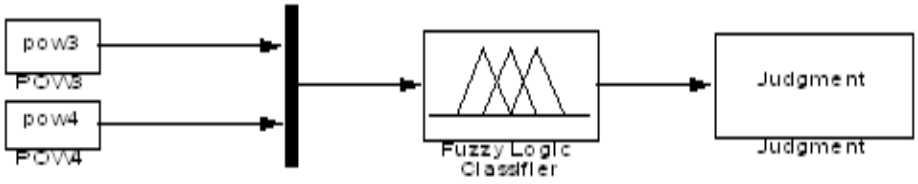

Figure 3.8. A fuzzy logic classification system. Alpha band power in channels C3 and C4 are two inputs and the output is a classification decision.

3.5.1 Foundations of Fuzzy Logic

Fuzzy logic is a multi-valued logic modeled by fuzzy sets. Fuzzy sets are sets whose members have degrees of membership [58]. Denoting feature $\boldsymbol{x}$ extracted from a trial and a set $\boldsymbol{S}$, a mapping can be defined as:

$$\boldsymbol{x}_{\boldsymbol{S}} = \begin{cases} 1, if\ \boldsymbol{x} \in \boldsymbol{S} \\ 0, if\ \boldsymbol{x} \notin \boldsymbol{S} \end{cases}, \tag{3-29}$$

which only shows whether or not $\boldsymbol{x}$ belongs to set $\boldsymbol{S}$. But sometimes, $\boldsymbol{x}$ may be located very near to the border between two sets and it is risky to assign it wholly to one set. In order to avoid the dilemma choice, the mapping can be generalized to a function which allows $\boldsymbol{x}$ to belong to the set $\boldsymbol{S}$ to a certain degree $\mu_S \in [0, 1]$. A fuzzy set can be defined as $\boldsymbol{S}=\{\boldsymbol{x}, \mu_S(\boldsymbol{x}) | \ \boldsymbol{x}\ \boldsymbol{U}\}$, where $\boldsymbol{U}$ is the universe of discourse and the function $\mu_S(\boldsymbol{x})$ is called membership function of $\boldsymbol{x}$ in $\boldsymbol{S}$. The membership function maps each element of $\boldsymbol{U}$ to $\boldsymbol{S}$ to a degree of membership between 0 and 1.

3.5.2 Training and building membership functions

In the training process of a fuzzy logic classifier, extracted features in each class are divided into several clusters. Based on each cluster, a fuzzy set is set up with a linguistic description and a membership function is estimated from samples in this cluster. The methods of clustering and estimation of parameters are similar to the methods used in the Bayesian statistical classifier.

Taking the alpha band power feature in channels C3 and C4 as an example, in the same class, the extracted alpha band power in each channel is divided into 5 clusters. Then 5 fuzzy sets {***vs***, ***s***, ***m***, ***l***, ***vl***} are set up and with linguistic descriptions {'very small', 'small', 'medium', 'large', 'very large'}. Their membership functions $\{mf_1, mf_2, mf_3, mf_4, mf_5\}$ are estimated from samples in the corresponding clusters respectively and shown in Figure 3.9.

In this research, each membership function is estimated as a Gaussian function in the beginning and modified later. For example, for the alpha power in channel C3, to estimate the membership function mf_4, in order to assign a degree of memberships μ_S =1 to features in a given range, the centre of the Gaussian function curve is expanded and it becomes a generalized bell function (see Figure 3.9). Also, as a part of training, the classifier is tested using the training data. Then, experience extracted from observing wrongly classified samples can be used to modify membership functions.

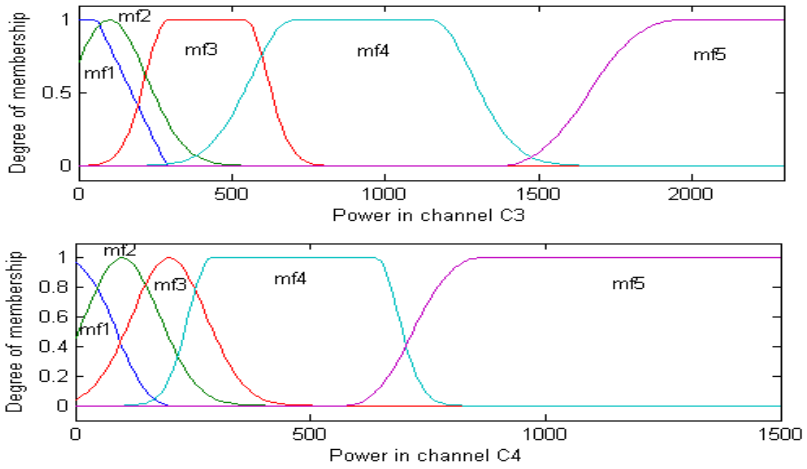

Figure 3.9. Membership functions for two inputs.

3.5.3 If-then rules

The "If-then" rule is key in mapping inputs to output. According to the label of samples in a cluster, some 'if-then' rules in the form: 'If x_1 is A and x_2 is B, then y is C' can be summarised and then used to build a classifier, where x_1, x_2 are input variables and y is the output variable; A, B and C are linguistic values defined by fuzzy sets on their universes of discourses. For example, a rule can be 'if the alpha band power in C3 is large and the alpha band power in C4 is small, then thinking is imaginary left hand movement'.

Each 'if-then' rule makes a conclusion and these conclusions are combined together to form a final decision, as shown in Figure 3.10. The process of combining will be discussed in section 3.5.5 and further examples are shown in section 3.5.6.

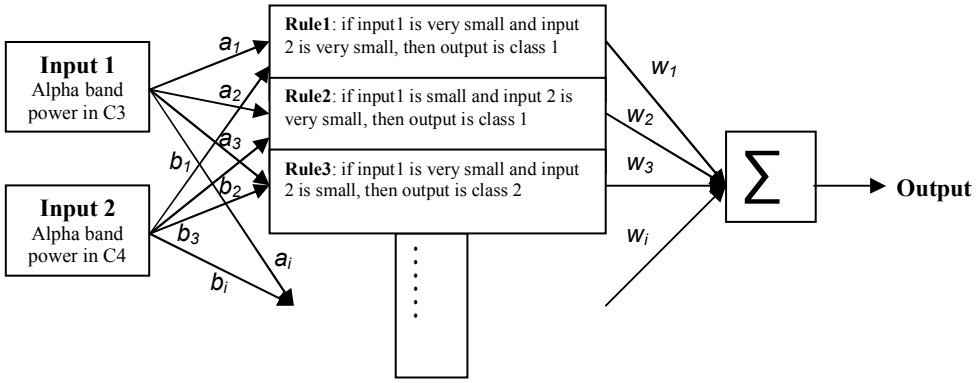

Figure 3.10. Different if-then rules are combined to make a final decision. The used feature is the alpha bandpowers in channels C3 and C4.

3.5.4 Define output membership functions

There are two types of fuzzy inference systems to define the output membership functions. One is Mamdani-type, whose output membership functions are fuzzy sets [59]. The other is Sugeno-type, whose output membership functions are either linear or constant [60]. In BCI research, the system has only one output as the classification decision. If the Mamdani-type is used, the number of output fuzzy sets is the same as the number of classes. In this research, it focuses solely on the 2-class problem and only a simple membership function, such as

$$\mu_{S1}(x) = \begin{Bmatrix} 0, if\ x \le 0 \\ x, if\ x > 0 \end{Bmatrix}, \tag{3-30}$$

is used for each output fuzzy set. In the Sugeno-type, the constants -1 and 1 are used as values for the output membership functions respectively. It was found that, in the BCI research carried out here, there is no significant difference between using two types of fuzzy logic classifiers, but the Sugeno-type system was more efficient computationally.

3.5.5 Fuzzy logic classification process

Here are 4 main steps to classify a test feature using the fuzzy logic classifier.

Step1: Fuzzify Inputs

The first step is to take the inputs and determine the degrees to which they are in each of the appropriate fuzzy sets via membership functions. It relies on matching the input variables to the membership functions. For example, when

the alpha band powers in channels C3 and C4 are used as inputs, the degrees of memberships can be looked up from the membership function curves in Figure 3.9.

Step2: Fuzzy operation

After the inputs are fuzzified, according to the fuzzy set used in a rule, a degree is assigned to each input in each rule. If a rule has more than one input, all inputs work together toward to an output according to the rule logic. Subsequently, a fuzzy operator is applied to obtain a number to represent a weight of this rule, which is also the degree of relevant output membership. This is decided by both the degrees of inputs and operation method. Usually, there are two methods, 'min' and 'prod', which are used in the fuzzy operation. For example, assuming a rule is that 'If x_1 is A and x_2 is B, then y is C' and the degrees of 'x_1 is A' and 'x_2 is B' are looked up from membership functions as a and b, respectively, with the operation method 'prod', the degree of output 'y is C' is $a \times b$.

Step3: Aggregate all outputs

Usually, several rules work together to make a decision, as shown in the example of Figure 3.10. In this case, each rule provides a result with a weight. The result is a fuzzy set for the Mamdani-type or a constant number for the Sugeno-type. Then, according to an aggregation method, these results are combined into a single fuzzy set or a number as an aggregate result. The aggregation method either uses the output with maximum degree or a sum of each rule's output.

Step4: Defuzzify

After aggregation, defuzzification process translates the aggregate result into a single number. Some methods [64], such as 'centroid', 'bisector' and 'maximum possibility', can be used. In this BCI research, the defuzzified single number can give expression to a classification decision. For example, for the 2-class problem of imagining left or right hand movement, the features are assigned to one class or the other depending on the sign of the number.

Each input membership function assigns a degree to any value in the universe of discourse, therefore, all the values in the feature space can be classified, and then, the feature space is divided by an output surface, which shows the mapping relationship from features to classification result. An

example of the fuzzy surface based on the alpha band power feature space is shown in Figure 3.11.

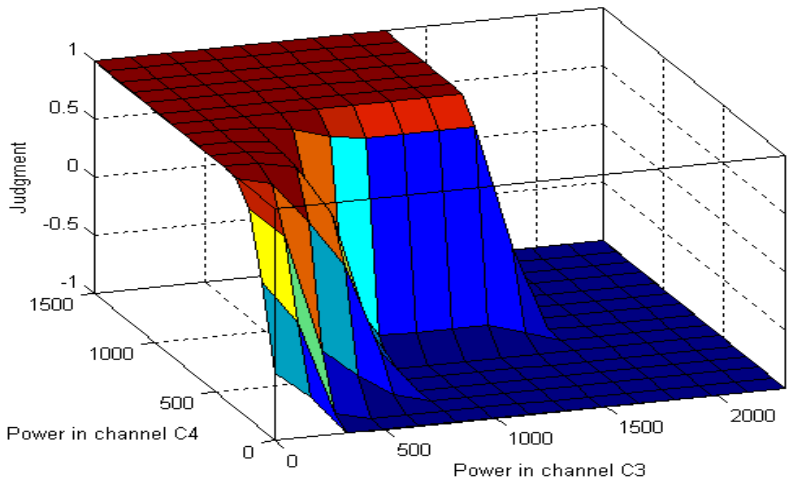

Figure 3.11. The alpha band power feature space is divided by an output surface. The classification result (imaginary left/right hand movement) depends on the sign of value matched on the output surface.

3.5.6 An example of the fuzzy logic classification process

In our research, different logic methods were tested in steps 2 to 4, and an example is used to show the classification process. Assuming the alpha band power in channel C3 is extracted from a test trial and denoted as X_{C3}; the degrees with which to assign X_{C3} to all fuzzy sets {***vs***, ***s***, ***m***, ***l***, ***vl***} are looked up from membership functions (see Figure 3.9) and they are {0,0.4,0.6,0,0} respectively. Similarly, the alpha band power in channel C4 is denoted as X_{C4} and the degrees of X_{C4} in the fuzzy sets {***vs***, ***s***, ***m***, ***l***, ***vl***} are {0.8, 0.2, 0, 0, 0} respectively. The output y is in the fuzzy set {***L***, ***R***}, which stand for imaginary left and right hand movement respectively. It is assumed the rules have been decided in advance. After ignoring rules whose input degree of memberships is zero, there remains a total of 4 rules which work together to make the decision and they are as follow:

If X_{C3} is small and X_{C4} is very small, then the thinking is left hand movement.

If X_{C3} is small and X_{C4} is small, then the thinking is right hand movement.

If X_{C3} is medium and X_{C4} is very small, then the thinking is left hand movement.

If X_{C3} is medium and X_{C4} is small, then the thinking is left hand movement.

If the ‘min’ method is used in step 2, the ‘max’ method in step 3 and the ‘maximum possibility’ method in step 4, the classification processing is as follows:

The weight of the first rule is 0.4∩0.8=0.4.

The weight of the second rule is 0.4∩0.2=0.2.

The weight of the third rule is 0.6∩0.8=0.6.

The weight of the fourth rule is 0.6∩0.2=0.2.

The degree, with which output y is in fuzzy set ***L***, is 0.4U0.6U0.2 =0.6.

The degree, with which output y is in fuzzy set ***R***, is 0.2.

Using the 'max' method, for each output variable, only the rule with the maximum weight is used in the aggregation, therefore the aggregated result is 'y is in {***L***, ***R***} with degree of {0.6, 0.2} respectively'. The membership functions in Figure 3.12 are assumed to be used for two output fuzzy sets. Then, the output membership mf_1 is chosen, since it has larger degree (0.6>0.2). The defuzzied output, obtained by matching the degree of output membership function $mf_1=0.6$, is equal to 0.6>0. So, the feature is assigned to class 1 (imaginary left hand movement).

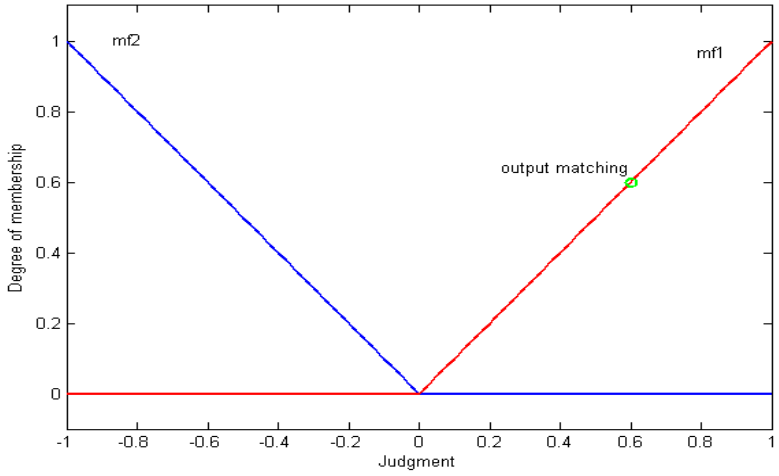

Figure 3.12. Output Membership functions. The two curves with different colours stand for the membership functions for class1 and 2 respectively. The green circle is the matched output in the example.

If the 'prod' method is used in step 2, the 'sum' method in step 3 and 'centroid' in step 4, the classification processing for the same example is as follows:

The weight of the first rule w_{11} is 0.4×0.8=0.32.

The weight of the second rule w_{12} is 0.4×0.2=0.08.

The weight of the third rule w_{21} is 10.6×0.8=0.48.

The weight of the fourth rule w_{22} is 0.6×0.2=0.12.

The degree, to which output y is in fuzzy set ***L***, $w_L=w_{11}+w_{21}+w_{22}=0.92$.

The degree, to which output y is in fuzzy set ***R***, $w_{R=}\ w_{12}=0.08$.

Using $mf_1=1$ and $mf_2=-1$ as output membership functions (Sugeno-type), the output $y=\frac{mf_1 \times w_L + mf_2 \times w_R}{\sum w_{ij}}=0.84>0$, so, the feature is assigned to class 1.

The fuzzy logic classifier models a flexible nonlinear classifier. We found that it gets high accuracy in this BCI application and shows good tolerance to recording errors, which is especially useful for the dynamic EEG signals. Additionally, it has the potential to give subjects more comprehensible feedback and help the training of a subject. However, it is hard to make the fuzzy logic classifier general to different subjects and experimental environments. The membership functions and the 'if-then' rules are core of the fuzzy logic classifier and they usually rely on a lot of training and very carefully tuning of parameters within the membership functions. Furthermore, the range of features usually changes during the feature updating process, which makes it awkward for the fuzzy logic classifier, since the membership functions have to be defined again. Therefore, the fuzzy logic classifier is more suitable to offline analysis, otherwise, it would require a lot of effort to determine in advance the membership functions and the set of 'if–then' rules for every sampling time.

Chapter 4 Application to Graz data

Following the discussions on feature extraction and classifiers, this chapter contains the results of their application to a publicly available dataset of BCI data. In order to assess and compare the different proposed features and classification algorithms discussed earlier, some typical classification results are presented in this chapter. Section 4.1 discusses the dataset used in this chapter, and in sections 4.2 and 4.3, the evaluation criteria for BCI performance and the validation method are discussed. According to the features used, the performance of most classifiers is listed and some related analysis is presented in section 4.4. Some discussions on suitable length of data for feature extraction, the optimal training period and computation times are presented in section 4.5. In section 4.6, the performance of various classifiers are discussed and compared with published results with the same dataset.

4.1 Data description

The data described in this chapter was provided by Department of Medical Informatics, Institute for Biomedical Engineering of the University of Technology Graz for the BCI competition 2003 [15]. During the experiment, a 25 year old female subject who had been trained before was asked to imagine left or right hand movements and feedback, in the form of a controlling bar on a screen, was shown to the subject. Three bipolar EEG signals were measured over the electrode positions of C3, C4 and Cz. The EEG was band-pass filtered between 0.5 and 30 Hz and then sampled at 128 Hz.

In total, 280 trials were contained in this dataset and these were originally divided into two equal parts for training and testing (70 left and 70 right imaginary hand movements are in each part). Every trial lasted for 9 seconds and the signed amplitudes of three EEG channels were recorded as digital time series. In each trial, the first two seconds was quiet and an acoustic stimulus indicated the beginning of the trial at t= 2 s. An arrow (left or right) as cue was presented on the screen from t = 3 s to 9 s. At the same time, the

subject was asked to move a bar in the direction of the cue through imagining hand movements and the classified bar movement was presented on the screen, which acts as a form of feedback to the subject [15].

4.2 Evaluation criteria of BCI performance

In order to measure and judge the performance of a BCI system, it is necessary to set some criteria for evaluation. Specifically, the criterion used is to examine how accurately the classifiers can identify the subject's intentions. Different criteria have been used by different researchers, which is also one of the reasons why it is difficult to compare different BCI systems. For example, a study [7] adds a rejection criterion in the evaluation, which sets some trials as 'unknown' thus reducing the number of errors. In this way, it is hard to compare with the other methods without the rejection strategy, since the 'unknown' trials were classified by them as well.

4.2.1 Classification accuracy

Classification accuracy is a common criterion to evaluate the performance of any classification algorithm, and is usually defined as the ratio of the number of correctly classified trials to the number of all tested trials. If the aim of the BCI application is to achieve some goals in a real or virtual environment, the classification accuracy can be measured by the ratio of the number of successes a subject manages to achieve against the goal to the numbers of attempts made. However, this ratio, termed the achievement ratio, is not the ideal way to measure the accuracy of the classifier, since the achievement ratio just describes an aggregated classification result which may require a long time to achieve. In the research carried out in this thesis, the EEG signal was classified continuously at every time sample and consequently the classification accuracy achieved was dependent on the length of the time interval over which the data was observed. The key idea behind this was to investigate the dependence of the classification results on the observation interval. The highest and average accuracy during varying observation intervals were selected as criteria to represent the best recognition rate and the average ability to identify a subject's intentions for a given feature and

classifier. Usually, a classification process with a high maximum accuracy was accompanied by a high average accuracy, but this was not always the case.

4.2.2 Other criteria

There are other criteria to measure the reliability of decision, such as confusion matrix [20], Cohen's Kappa [5] and information rate [19]. These are briefly discussed below.

The confusion matrix ***Co***, is that matrix whose (i,j) element, ***Co***(i,j), is the number of trials that belong to the i-th class and are assigned to the j-th class. An example of confusion matrix is shown in Table 4.1. The confusion matrix is a wonderful way to describe the classification result. Not only can the accuracy be derived from the diagonal elements of the confusion matrix, but also the off-diagonal elements reflect the wrongly classified and show if the classification is biased. However, it is inappropriate to compare the numbers in two confusion matrices directly if they were obtained with different numbers of test trials.

Class	1	2	Total
1	64	6	70
2	8	62	70
Total	72	68	140

Table 4.1. An example of confusion matrix. The result is from using the LDA classifier with both alpha and beta band powers feature at a time t=6.08, when it achieved accuracy of 90%. Class 1 and 2 are imagine left and right hand respectively.

For the 2-class problem, the observed agreement

$$P_0 = P_{11} + P_{22}\,, \tag{4-1}$$

is equal to the total accuracy, where P_{11}, P_{22} are the rates of correct classification in each category. The chance of random agreement is given by

$$P_e = P_1^2 + P_2^2\,, \tag{4-2}$$

where P_1, P_2 are the prior probabilities in each category and P_e is the hypothetical accuracy. The value of Kappa is defined as:

$$K = \frac{P_0 - P_e}{1 - P_e}\,. \tag{4-3}$$

In this research, it is believed that the different classes have equal prior

probabilities, specifically, $P_1=P_2= 0.5$ and $P_e= 0.5$, therefore, for the 2-class problem,

$$K = 2(P_0\text{-}0.5)\,. \tag{4-4}$$

The information transfer rate, *ITR*, is defined as

$$ITR = \log_2 n + P_0 \log_2 P_0 + (1\text{-}P_0)\log_2 \frac{1-P_0}{n-1}\,, \tag{4-5}$$

where n is the number of classes and the above equation is for the information transfer in bits/trials. When the time of trials is considered, it can be described as information transfer in bits/s. For the 2-class problem, $n=2$, therefore,

$$ITR = 1 + P_0 \log_2 P_0 + (1- P_0)\log_2 (1-P_0)\,. \tag{4-6}$$

The relationships of the Kappa, *ITR* with accuracy are shown in Figure 4.1, whilst the *ITR* is usually meaningful only for accuracies higher than the chance of random (e.g., in 2-class problem, for equal prior probabilities of classes, accuracy of 0.5 can be achieved by chance). From the above formulae and Figure 4.1, it can be seen that both the Kappa and information transfer rate are functions of the classification accuracy, whilst, the kappa enlarges changes of the accuracy linearly and *ITR* display high sensitivity to small changes in the range of high accuracy. However, for the 2-class problem, they are only mapping the accuracy to different numbers but requiring more computations, also they are not as easily interpretable as simple accuracy. Therefore, their advantages are not very significant in the 2- class problem and so these are not used in this thesis.

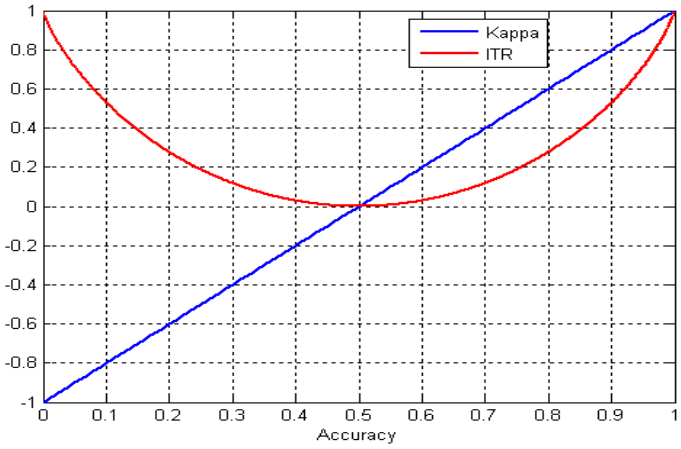

Figure 4.1. Relationships of the Kappa and *ITR* with accuracy for the 2-class problem.

4.3 Cross validation

In order to test the ability of the classifiers to generalise, a randomly chosen subset of the trials is chosen as the validation data and whilst the other parts are used for training. *K*-fold cross validation [66] was used as the strategy, in this approach the dataset is divided as *k* subsets. Of the *k* subsets, a single subset is used for validation and the remaining *k-1* subsets are used to train various classifiers which are then evaluated using the single validation subset. Two typical examples in this thesis are dividing the data into two even parts or the leave one out (LOO) method [66], which considers every trial as the validation subset and *k* is equal to the number of trials. During this process, all the trials are used for both training and validation, and each trial is used for validation once.

4.4 Performance of feature and classifier pairs

The classification results achieved are shown below as a function of the length of the observation window for each of the features discussed earlier. During the classification process, analysis showed that the classification before 4 s was almost by chance, since the idle signals and the first one second of task driven EEG signal could not provide enough information for effective discrimination. It was also observed that the amount of discriminatory information contained in the signal diminishes after 8 s, perhaps due to subject fatigue. This can be verified by inspection of the classification accuracy over whole 9 seconds period. For example, the classification accuracy when using both the alpha and beta band powers as features with the LDA classifier in the whole 9 seconds period is shown is Figure 4.2 and it indicates that the main period of interest is from 4 to 8 seconds. In the following, only the classification accuracies in this period are discussed.

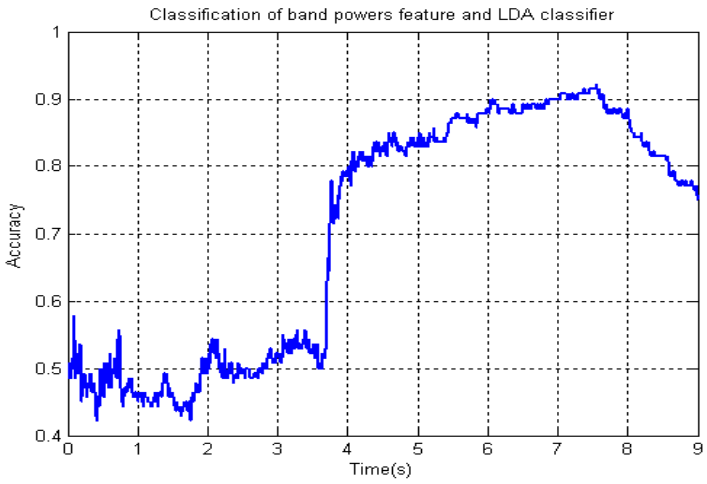

Figure 4.2. Classification accuracy versus time for the band power feature with the LDA classifer.

Features were updated and classifiers were retrained at every sample within the 4 to 8 second period. During this process, all the previous samples are utilised to estimate features and the classifications were made at every sample. This enables continuously varying values of classification accuracy to be obtained. In order to compare with results in the literature, listed accuracies in the tables in this chapter were obtained from the arrangement of training and test data used in the BCI competition 2003.

4.4.1 Classification results using time series

As discussed in Chapter 3, three methods were used to build templates of the time series waveform. Figure 4.3 illustrates the classification accuracies achieved by the different template building methods as functions of time.

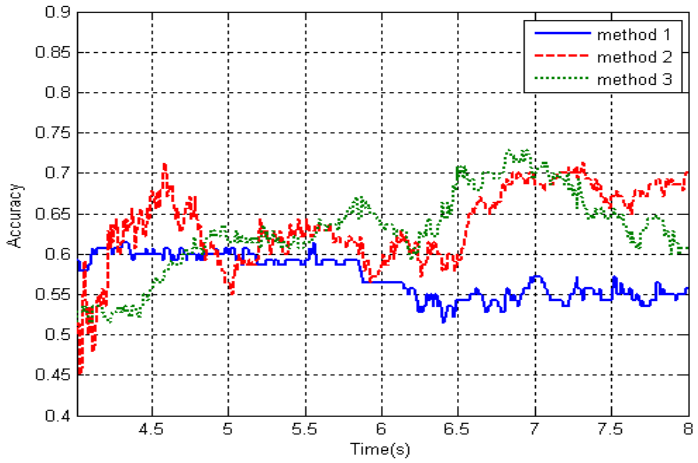

Figure 4.3. Classification accuracies versus time for 3 template building methods

The first method, which averaged all time series in the same class, could not capture sufficient discriminative information, especially in the period after 6 seconds. The second method of adding the nearest neighbour method to template matching and the third method of dividing the time series into subgroups gave similar improvements over the first method, particularly for longer length data blocks. However, the combination of the nearest neighbour and template matching method, i.e., method 2, entailed a much longer computation time. The results of using method 3 show that the idea of dividing the EEG signal into subgroups according to the signal intensity in idle period is reasonable, especially when a huge amount of EEG data needs to be processed but may be not be suitable for the dataset including just a few trials (such as the dataset in Chapter 5). In this dataset, the EEG signals were divided into 4 subgroups, but limited by the amount of data, the appropriate number of subgroups and its variability for different subjects still need to be further investigated.

The result of template matching using the cross correlation sequence method discussed in section 3.1.2 is shown in Figure 4.4. In this way, a better classification result was obtained, but the long computation time for this method makes it impractical.

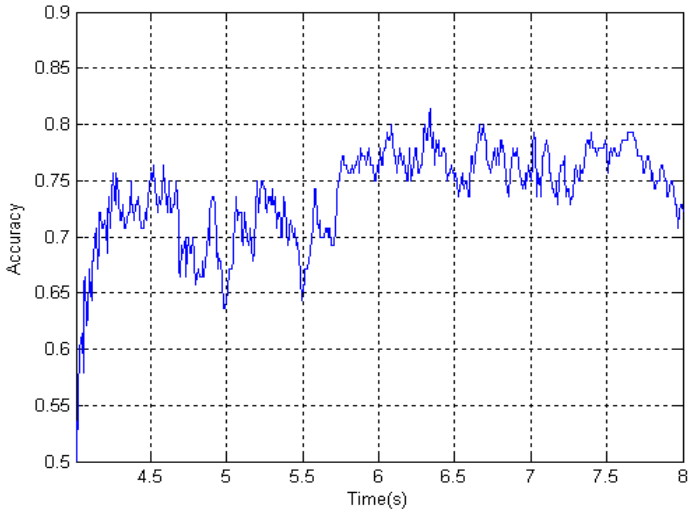

Figure 4.4. Classification accuracy versus time for the correlation sequence method

The best and average classification accuracies, and computation times of processing 140 training trials and 1 random test trial, for all the four methods discussed above are shown in Table 4.2.

Feature	Classifier(template)	Classification accuracy		Computation time
		Best accuracy	Average accuracy	
Zero lag correlation	Simple average template	61%	57%	2.1 s
	Nearest neighbour template	71%	63%	65.3 s
	Subgroup template	73%	63%	5.1 s
Correlation sequence	Nearest neighbour template	81%	74%	91.5 s

Table 4.2. Classification results using template matching

4.4.2 Classification results using AR components

The EEG signal can be modelled by an autoregressive (AR) model. As discussed in Chapter 2, the AR coefficients and AR poles were considered as features separately. Figure 4.5 illustrates the classification results of using AR coefficients as a feature vector with different classifiers.

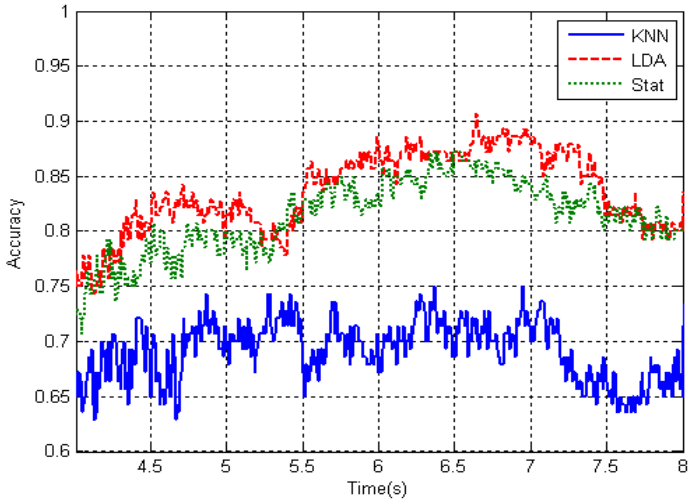

Figure 4.5. Classification accuracies versus time for the AR coefficients feature with 3 classifiers. The AR order was 4 and the AR coefficients were estimated from all the previous samples. *k* nearest neighbour, LDA, Bayesian statistical classifiers were used and denoted as 'KNN', 'LDA'and 'Stat' respectively.

The best and average classification accuracies, and computation times of processing 140 training trials and 1 random testing trial, for all the AR features and different classifiers are shown in Table 4.3.

Feature	Classifier	Classification accuracy		Computation time
		Best accuracy	Average accuracy	
AR coefficients	KNN	75%	69%	45.0 s
	LDA	91%	83%	51.7 s
	Stat	87%	81%	47.3 s
AR poles	KNN	86%	77%	85.4 s
	LDA	86%	78%	85.9 s
	Stat	89%	77%	85.5 s

Table 4.3. Classification results using AR components

The AR model expresses the signal characteristics through AR coefficients, which contain information from the entire frequency band. AR poles are roots of the AR polynomial and are related to AR spectral peaks, so, they focus on the characteristics of signals in frequency bins where spectral peaks appear. Therefore, compared with the AR coefficients feature, the AR poles feature contains less but more specific information. For the LDA and Bayesian statistical classifiers, using the AR coefficients as features provided better classification accuracies, whilst the *k*-nearest neighbour classifier obtained better results from the AR poles feature. The AR poles are derived from the AR coefficients and more calculations are required during this process, which makes the computation time longer.

4.4.3 Classification results using spectral components

Both the total power and spectral peaks in the two most prominent frequency bands: alpha (7~13 Hz) and beta (14~26 Hz) bands were tested as features. Figure 4.6 illustrates the classification accuracies for the band powers feature with different classifiers.

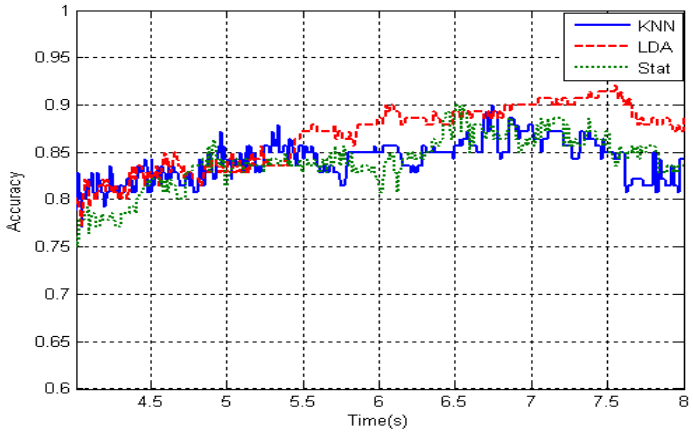

Figure 4.6. Classification accuracies versus time for the band power feature with 3 classifiers. Both of the alpha and beta band powers were used for the LDA classifier, while only the alpha band power was used for the *k*-nearest neighbour and Bayesian statistical classifier.

The best and average classification accuracies, and computation times, including a processing of 140 training trials and 1 random test trial, for both of the spectral features and 3 classifiers are shown in Table 4.4.

Feature	Classifier	Classification accuracy		Computation time
		Best accuracy	Average accuracy	
Band powers	KNN	90%	83%	16.7 s
	LDA	92%	87%	7.2 s
	Stat	91%	84%	11.6 s
Spectral peaks	KNN	89%	84%	17.5 s
	LDA	89%	84%	8.0 s
	Stat	89%	82%	11.7 s

Table 4.4. Classification results using spectral components

In most trials from this dataset, both the mu and central beta rhythms were active when movements were imagined, but the mu rhythm produced more discriminative information. If only the alpha band power in two channels were combined and used as a feature vector, the classification results of all the classifiers were satisfactory. If both the alpha and beta band powers were used as features, it was found that the classification result of the LDA classifier was improved but not for the *k*-nearest neighbour and Bayesian statistical classifiers.

To calculate a distance for the nearest neighbour classifier, the absolute differences in each channel between two trials are accumulated according to the selected metric. For example, when both alpha and beta powers in channels C3 and C4 are used as features, where the test feature is denoted as $\boldsymbol{X}= (P_{\alpha C3}, P_{\alpha C4}, P_{\beta C3}, P_{\beta C4})^T$ and the training feature is $\boldsymbol{Q}=(q_1,q_2,q_3,q_4)^T$, the Manhattan distance is

$$D_m = \sum_{1}^{4} | \boldsymbol{X}_i - \boldsymbol{Q}_i | . \qquad (4\text{-}7)$$

However, the power variation in the beta band is not significant in some trials. So, two trials in the same class may have very close measures of the alpha band power but quite different measures of the beta power at the same time. As shown in (4-7), a distance is obtained by summing all the differences from 2 bands; so, a distance between features in different classes may be shorter than a distance between features in the same class. This makes the *k*-nearest neighbour classification less able to discriminate between classes.

For the Bayesian statistical classifier, using both alpha and beta band powers as a feature vector results in a higher dimensional feature space, which is harder to divide into different clusters corresponding to different Gaussian distributions. Therefore, except for the LDA classifier, the other classifiers obtained better accuracies with only the power or spectral peak in the alpha band.

The alpha band power in channels C3 and C4 were used as input features to the fuzzy logic decision classifier, but the classification was just tested using a fixed number of samples of the signal waveform, this is because it requires a lot of training and modifications of membership functions prior to classification, which makes it not suitable for a continuous classification process. For the given period, the fuzzy logic decision classifier achieved an accuracy of 90.71%. With modifications of membership functions, the fuzzy logic decision classifier could achieve very high accuracy any time in the period of 4-8 s, but it would be at a high computation cost and the classifier would likely not be general and applicable to other datasets. So, the fuzzy logic classifier is more suitable to make a judgment for a whole trial and it was not applied to other features.

When only the alpha band power was used as feature, the piecewise LDA classifier (as discussed in Chapter 3) could provide a little improvement. The best accuracy of the piecewise LDA classifier was 90.71%, whilst it was 89.29% for the original LDA classifier when using the alpha band power as the feature. But it is not particularly helpful for the mixture of the alpha and beta band powers, since it is difficult to divide the mixed feature into subsections, which is similar to the problem encountered in the Bayesian statistical classifier.

4.4.4 Classification results using eigenvector components

During the analysis of the covariance matrix of multi-channel EEG signal, the method of principal eigenvector obtains the dominant basis vector associated with the covariance matrix, whilst the method of common spatial pattern (CSP) projects the covariance matrix to a new space through the K-L transform. Figures 4.7 and 4.8 illustrate the classification processes for the

principal eigenvector and common spatial pattern features with different classifiers respectively.

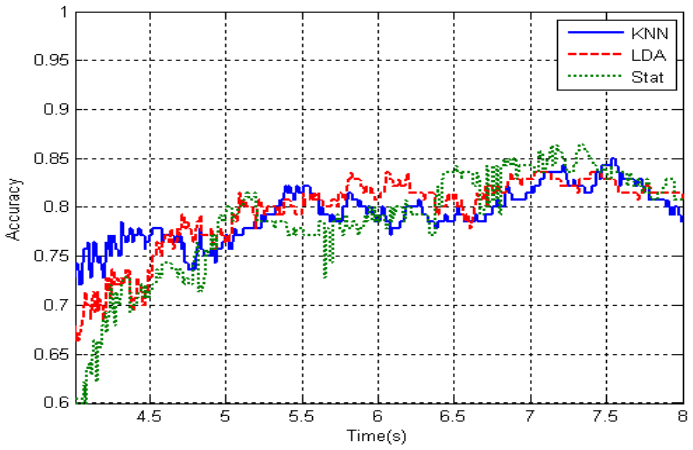

Figure 4.7. Classification accuracies versus time for the principal eigenvector feature with 3 classifiers.

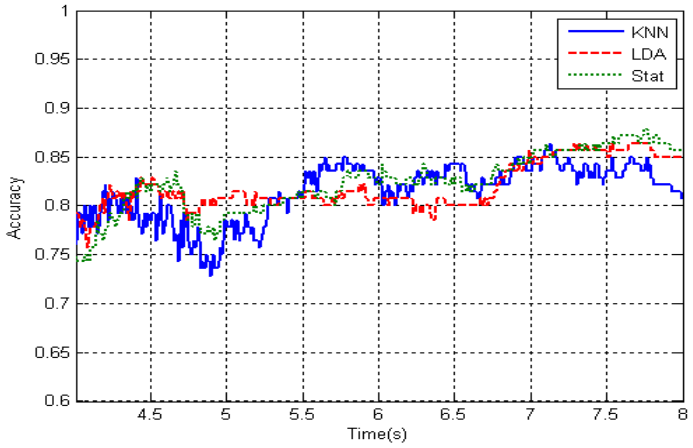

Figure 4.8. Classification accuracies versus time for the common spatial pattern feature with 3 classifiers.

The best and average classification accuracies, and the computation times of processing 140 training trials and 1 random testing trial, for both the principal eigenvector and common spatial pattern features with the different classifiers are shown in Table 4.5.

Feature	**Classifier**	**Classification accuracy**		**Computation time**
		Best accuracy	**Average accuracy**	
Principal eigenvector	KNN	85%	79%	9.1 s
	LDA	84%	80%	11.2 s
	Stat	86%	79%	11.5 s
CSP	KNN	86%	81%	18.2 s
	LDA	86%	82%	18.1 s
	Stat	88%	82%	18.0 s

Table 4.5. Classification results using eigenvector components

The principal eigenvector reflects the direction of the highest energy concentration. The distribution of the principal eigenvectors of all trials in the same class was found to be tightly clustered in a given region, therefore, it is not necessary to divide the feature space into several clusters and only one Gaussian model was used in the Bayesian statistical classifier.

As discussed in section 4.4.3, the mu rhythm is the main brain wave motivated by imaginary hand movements. If the alpha band filter was applied before the analysis of the eigenvector components, the highest accuracy was better but the average accuracy was lower. For example, when the Bayesian statistical classifier was applied to the CSP feature extracted from the alpha band filtered EEG signal, an accuracy of 89.3% was achieved at several samples during 7.5 to 8 seconds intervals, which is better than the highest accuracy of 87.86% without filtering; but the average accuracy was 80.97%, which is lower than the average accuracy of 82.34% without filtering. In particular, the classification accuracy at the beginning of observation period (4 seconds after trials started) was very low. The signal in this dataset was filtered with a pass band from 0.5 to 30 Hz in advance, and thus contained information from both the alpha and beta bands. Therefore, for the eigenvector features, using an alpha band filter before a continuous classification can provide higher accuracy at some samples, but it is not as stable as the classification without filtering. The results outlined in Table 4.4 are from the classification without filtering.

4.5 Analysis of data length of feature extraction, optimal training time and computation time

It was assumed that updating features in synchrony with the thinking processes, the features would be robust and helpful for classification. However, this is a question about how many previous samples should be used for the feature extraction and if a fixed length of data can provide more robust features. This was tested by comparing the classification performances achieved by using a variable data block as discussed above with those obtained using a fixed block that is slid over the data. An example is discussed here, which uses the LDA classifier and features of either the AR coefficients or band powers.

The length of the fixed block is 200 samples. The best accuracy and average classification accuracies from 4 to 8 seconds are shown in Table 4.6 for these two different feature extraction methods. Compared with the variable sample length method that incorporates all previous data, the fixed length signal did not perform as well, since it contained less information.

Features	Length of signals	Classification accuracy	
		Best accuracy	Average accuracy
AR coefficients	Fixed	87%	77%
	Variable	91%	83%
Band powers	Fixed	87%	74%
	Variable	92%	86%

Table 4.6. Classification results using features extracted in two different ways

The sample by sample classification approach is an online implementation, which means that the classifiers can track the dynamic EEG processes well. However, there was still a question about the optimal training time to obtain the best classification accuracy. Applying the LDA classifier to both the alpha and beta band powers features provided very encouraging results in previous tests, so it was selected for a test to determine the best training and testing times. With the variable sample length method, the classifier was retained at every training sample and then applied to all the test data over varying time intervals separately. The best classification accuracies in every half a second period are shown in Figure 4.9 and it shows that the classification results tend to be better if the training and testing time periods are the same and that a longer time provides more information and achieves better results.

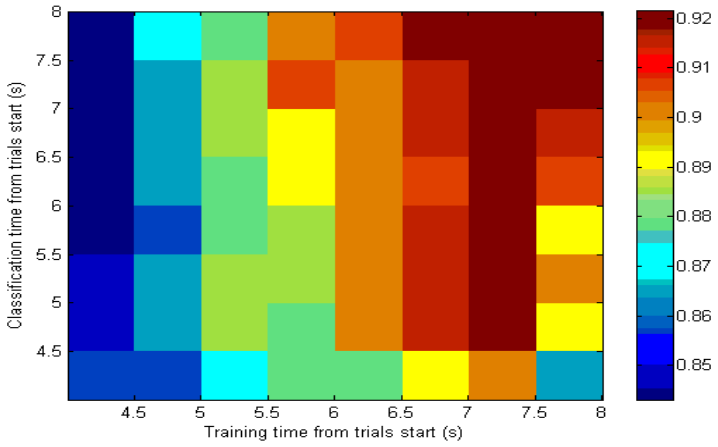

Figure 4.9. Classification accuracies of all possible combinations of training and testing periods runs. The colorbar corresponds to classification accuracy.

In this research, besides the classification of the test trials, the BCI system needed to be retrained at every sample, therefore, the computer processing time includes both the training and classification times. The computation times shown in the Tables 4.2-4.5 are for processing 140 training trials and only one test trial. In the offline classification process, the LDA classifier could finish the classification of all test trials together in a matrix calculation, but the nearest neighbour and Bayesian statistical classifiers have to analyse the test trials one by one. So, for the LDA classifier, the total computation time for many test trials is much shorter, but in an online application, one trial needs to be deal with at any time. That is the reason why only one test trial was chosen to compare the computation times.

4.6 Comparisons and discussion

During the continuous classification for the EEG signals in period 4-8 s, most of the features gave satisfactory results with all classifiers tested except the time series waveform. In particular, the band power features provided the highest accuracy for every classifier. For each kind of feature, the performance achieved by different classifiers was also compared and the LDA and Bayesian statistical classifier achieved the highest accuracies for 3 kinds of features separately. Specifically, the best accuracies achieved by both the LDA and the Bayesian statistical classifiers are about 90 percent for many features, which is on a par with the reported best accuracy in the BCI competition 2003 for the same dataset as 89.7 percent [34]. Figure 4.10 shows the best classification accuracies obtained during the 4-8 second period for each different feature and classifier combination – most of these results are similar or even better than reported results in the literature. Additionally, applying the LDA classifier to band powers feature achieved an accuracy of 95 percent in cross validation (its best accuracy is $90.48 \pm 2.15\%$ and average accuracy is $84.10 \pm 1.85\%$, where the error ranges correspond to a 95% confidence interval.).

Since the template matching algorithms were applied to the time series waveform only and the fuzzy logic decision classifier was applied to alpha band power feature only they are not listed in Figure 4.10.

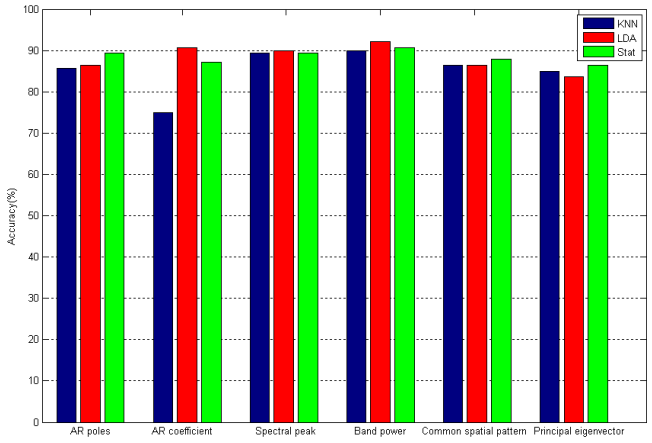

Figure 4.10. The best classification accuracies for different features and classifiers. The order of AR model was 4 and only 1 Gaussian model was used in Bayesian statistical classifier for the principal eigenvector feature.

The *k*-nearest neighbour classifier is more suitable for low dimensional features, such as the AR poles and principal eigenvector but not for the high dimensional feature like AR coefficients, and mixed alpha and beta band powers. The Bayesian statistical classifier is effective and stable for all kinds of features. It may have more potential for the multi-class problems, but preliminary results indicate it does not perform well for a small quantity of training data, since insufficient samples in a cluster can give quite inaccurate estimates of parameters of the model required for the cluster.

Computation time is also an important issue if the BCI system is to respond to the subjects' intention rapidly. Using 140 training trials and only one test trial, the computer processing time of different algorithms is shown in Figure 4.11. In this process, no SIMD (single instruction, multiple data) technique was used, but the training process was achieved by using vector operations in Matlab. The results show that all the classifiers could finish the classification quickly except when they were applied to the AR features. Compared to other features, the band power and the principal eigenvector features require shorter computation times. The LDA classifier finished classification quicker for the spectral components features, while when using AR components features the *k*-nearest neighbour classifier was the fastest.

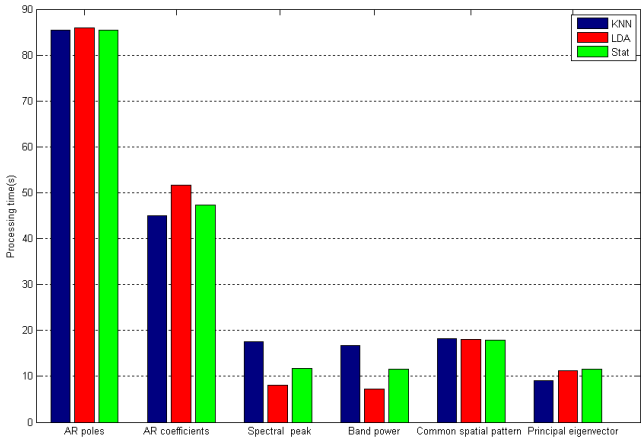

Figure 4.11. Computation times for different features and classifiers. The computing includes extracting feature from 140 training trials and classification for a random test trial.

In addition to optimizing the code, the computation time can be further improved, since for practical applications, the training can be finished in advance if it is not updated online. In fact, training is the major computational burden, since a lot of data is needed to populate the feature space with good coverage and once the classifier has been determined, it can be re-used for all the test trials. This can be validated by observing the computation time of the classification of many trials together. An example of applying the LDA classifier to the band powers feature is discussed here. Using 140 training trials and other 140 trials for test, the processing time is approximately 16 s. Comparing with the computation time of classifying only one test trial, it can be seen that training is the major computational burden.

In this experiment, each trial lasts 9 s and the sampling rate is 128 Hz. From results shown above, the computation time of classification of one test trial cost is much shorter than the trial's duration. Furthermore, the processing time can be made much shorter if the training can be carried out offline in advance. Therefore, we can conclude that the algorithms presented here have potential for online applications, where rapid response to trials is essential.

Chapter 5 Application to Adelaide data

The previous chapter presented the classification results from the Graz dataset and it was shown that most features and classification algorithms achieved satisfactory classification performance. In order to verify the generality of the tested methods, the same operations were repeated on a dataset recorded at the University of Adelaide. Based on the experience and conclusions in the last chapter, some typical features, including the time series waveform, AR coefficients, band powers and common spatial pattern, and classifiers, including the template matching, LDA and Bayesian statistical classifiers are investigated using this dataset. Section 5.1 introduces details of the measurements and data recording methodology. In section 5.2, the performance results when using different features and classifiers are presented. Some comparisons and analysis are discussed in section 5.3.

5.1 Experiment and Data Acquisition

In order to obtain more data for our BCI research and to verify the previous research on the Graz dataset, experiments were carried out in the Human Sensorimotor Plasticity Laboratory at the University of Adelaide. Some descriptions of the experiment process are briefly reported below.

5.1.1 Experiment procedure

A subject (25 years old, male, no colour-blindness, never trained before) was seated in a comfortable chair and was asked to look at a custom-made 3 light indicator (yellow for ready, red for right and green for left, 3 lights are in different positions) placed at eye level approximately 1 meter in front of him. The equipment and the indicator used in the experiment are shown in Figure 5.1. During the experiment, the subject was asked to keep his arms and hands relaxed and motionless, and a number of separate trials took place. Each trial started with 2 seconds of relaxation time, during which all lights were off. At t=2 s, the yellow light lit up for one second to warn the subject to prepare for the commencement of the trial. At t=3 s, the green or red light were lit as an instruction to the subject to imagine vertical movements in the

left or right hand, respectively. This light continued for 7 seconds and there was a further two seconds for the subject to relax again. This timeline of process is shown in Figure 5.2. The labels of trials were produced randomly by the EEG capture setup; a rule was set up to forbid sequences longer than four successive trials with the same label. The experiment consisted two runs (reference electrode placed at Fz and bipolar), each of 40 trials (20 left and 20 right hand imagined movements).

Figure 5.1. The Equipment and indicator used in the experiment

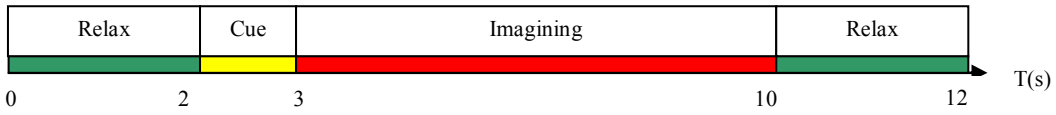

Figure 5.2.Sequence of experimental events

5.1.2 Recording methodology

Both reference and bipolar recording methods were used in the experiment. For the reference recording, 4 Ag-AgCl electrodes were used and they were placed at C3, Cz, C4 and Fz positions. The signal in the Fz electrode was used as the reference for the signals in the other three electrodes. This yielded three channels of signals referenced to a common node. For the bipolar recording, 3 pairs of Ag-AgCl electrodes are placed approximately 3 cm on either side (anterior and behind) of C3, Cz and C4. All the placements of electrodes were made by a highly skilled technician with many years of experience in EEG measurements. The EEG signal was sampled at 500 Hz and amplified by a low noise specifically for EEG applications amplifier CED 1902 (Cambridge Electronic Design, UK), then the signal was fed into a computer via a CED 1401 interface and recorded in a

software called CED 'signal for windows version 3', while the dc signal was removed by the software (more details about the amplifier and software can be found in reference [36]). Care was taken to ensure electrode impedances do not exceed the recommended maximum of 5kohm. If the impedance exceeded this limit, then sufficient conducting gel was applied at the interface of the electrodes and the scalp until the impedance was below 5kohm.

5.2. Classification and performance

With the previous experience of investigations using the public dataset and in order to limit scope to some typical features, only the best features in each category were selected for the processing of the data collected. These are, namely, the time series waveform, AR coefficients, band powers and common spatial pattern. The LDA and Bayesian statistical classifiers, which provided accurate and stable classifications in the last chapter, were applied to most features in this dataset, except the time series waveform as the classification process is simply template matching.

Only 40 trials were recorded in each dataset and in order to use as many trials as possible for training, the strategy of leave one out (LOO) was used for validation. Since the aim of this research was to compare the features and classifiers rather than to obtain very high classification accuracy, a simple method of setting test sets was used, that is, a pair of trials in different classes but the same order (e.g. the first trial in each class) were chosen as the test set, while the remaining trials were used as the training set. In this manner, every trial was tested only once at each sampling time and the overall classification accuracies were obtained from cycling through all the 20 pairs of the test trials and averaging the classification accuracies.

The computation times for different features and classifiers have already been discussed and compared in detail in the previous chapter. With the higher sampling rate, the number of samples in each trial is larger in this dataset than the Graz dataset. However, the findings for the relative computation times of the different feature extraction and classification schemes remain valid for this dataset. To avoid repetition, the computation time issue is not discussed again in this chapter.

5.2.1 Classification results using time series waveform

The template matching method was used for the time series waveform feature. The data was recorded with the DC value removed and so some AC noise still existed. In order to remove some of this noise, the alpha band pass filter was applied to the trial time series before it was used to build a template or for classification.

As only a few trials were carried out, it is difficult to divide the trials into subgroups. Therefore, only the method of using an average template (method 1) and the method of combining template matching with the nearest neighbour (method 2), as discussed in Chapter 3, were used. Except for the time series waveform of the trial under test, the remaining time series waveforms from other trials in the same class were averaged to build the template in the first method. In the second method, the time series waveform from each trial in the training set was used as a template. Then, correlations between the test data and the template(s) in each class were calculated. In exactly the same manner as for the Graz dataset, the classification was performed as a continuous process in time and the classification accuracy time courses achieved by two methods over the 4 to 10 seconds period are shown in Figure 5.3. Surprisingly, both template matching schemes, i.e., methods 1 and 2, achieved very stable classification results, even better than those for the Graz dataset. Additionally, for the reference dataset, method 2 did not provide obvious better results than method 1, which is not consistent with the conclusion obtained from the application to the Graz dataset.

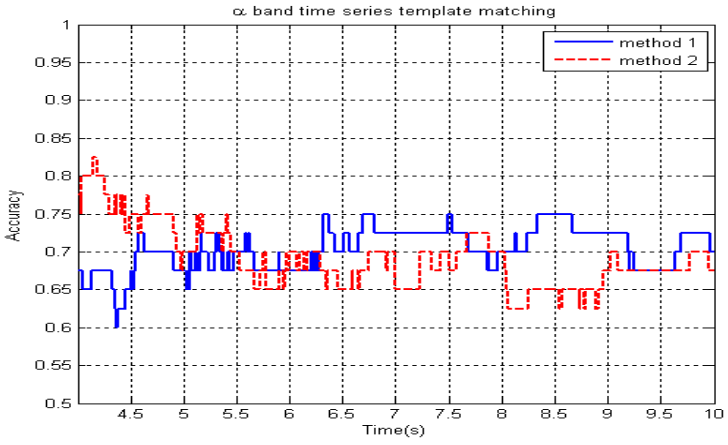

Figure 5.3. Classification accuracies versus time for time series waveform template(s) for the reference recording dataset.

5.2.2 Classification results using AR coefficients

A 4-th order AR model was applied to the EEG signals recorded from the experiment. As discussed in the last chapter, whilst using the AR poles as features provided a low dimensional feature vector, which is welcomed by the *k*-nearest neighbour classifier, but for the LDA and Bayesian statistical classifiers, using the AR coefficients as features can provide more discriminative information. Thus the AR coefficients were chosen as the feature vectors. Continuous classification accuracy time courses with the LDA and Bayesian statistical classifiers are shown in Figure 5.4. When compared with the time series waveform feature, the best classification accuracies were higher, but as shown in the figure, the classification accuracies were more volatile. This is especially true in the case of the Bayesian statistical classifier, which produced some very low classification accuracies at several points in time.

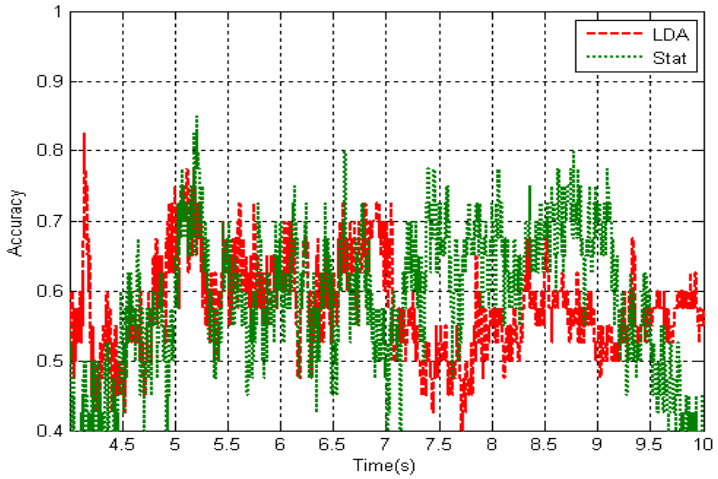

Figure 5.4. Classification accuracies versus time for the AR coefficients feature with the LDA and Bayesian statistical classifiers for the reference recording dataset.

5.2.3 Classification results using band powers

In Chapter 4, it was shown that using the band powers as features always provides better classification results than using the spectral peaks, regardless of the choice of classifier. Therefore, the band powers were selected as the features in this spectral approach. The LDA and Bayesian statistical classifiers were used and based on the conclusions from Chapter 4, both the alpha and beta band powers were used as features for the LDA classifier but only the

alpha band power was chosen as a feature for the Bayesian statistical classifier. The classification accuracy time courses of these two approaches are shown in Figure 5.5. The classification accuracies over time achieved by the two classifiers have similar trends, with the Bayesian statistical classifier performing demonstrably worse than the LDA classifier for t<7 s, but managing to achieve comparable performance to the LDA classifier in the period from 7 to 9 s.

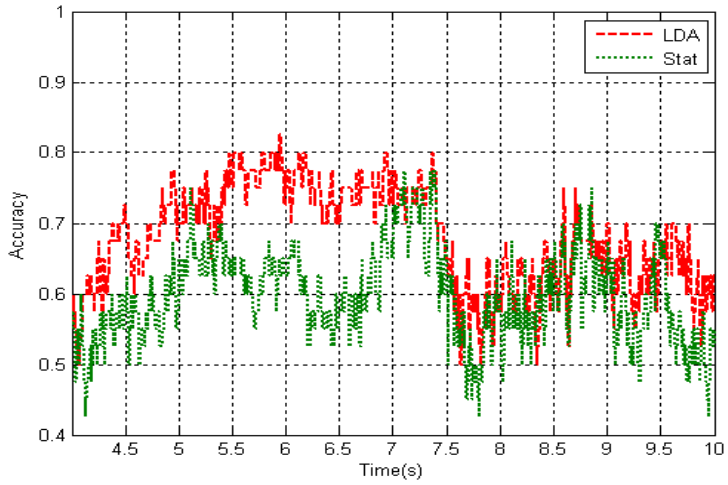

Figure 5.5. Classification accuracies versus time for the band power feature with the LDA and Bayesian statistical classifiers for the reference recording dataset.

5.2.4 Classification results using common spatial pattern

In the category of eigenvector features, the common spatial pattern was chosen. In order to utilise the information from both the alpha and beta bands, the EEG data was processed by a band pass filter (7-26 Hz). The covariance matrix computed from every trial was updated at every sample and then projected onto a new feature space through the K-L transform. Both the LDA and the Bayesian statistical classifiers were used. Figure 5.6 shows the classification accuracy time course achieved by the two classifiers with the LDA algorithm again providing the best results.

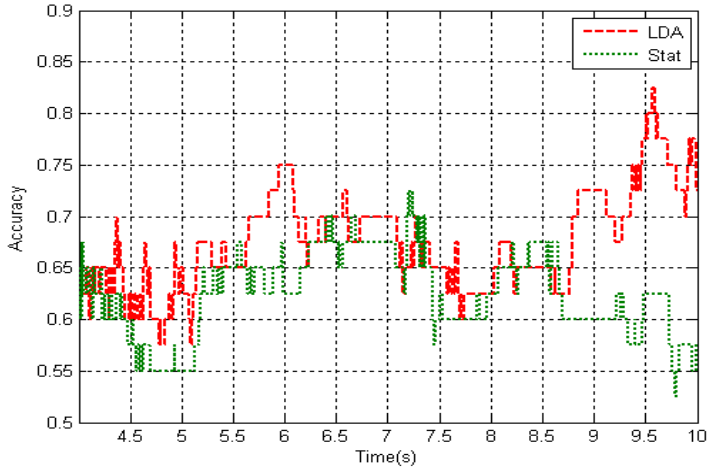

Figure 5.6. Classification accuracies versus time for the common spatial pattern feature with the LDA and Bayesian statistical classifiers for the reference recording dataset.

All the classification results for the reference and bipolar recording datasets, in the period of 4-10 s are summarised in Tables 5.1 and 5.2. Both average and the highest accuracy scores are presented.

Feature	**Classifier**	**Classification accuracy**	
		Best accuracy	**Average accuracy**
Time series waveform	Average template	75%	70%
	Nearest neighbour template	83%	69%
AR coefficients	LDA	83%	59%
	Stat	85%	58%
Band powers	LDA	83%	68%
	Stat	78%	60%
CSP	LDA	83%	67%
	Stat	73%	63%

Table 5.1. Classification results for the reference recording dataset.

Feature	**Classifier**	**Classification accuracy**	
		Best accuracy	**Average accuracy**
Time series waveform	Average template	73%	60%
	Nearest neighbour template	78%	71%
AR coefficients	LDA	73%	53%
	Stat	70%	52%
Band powers	LDA	70%	52%
	Stat	68%	52%
CSP	LDA	70%	57%
	Stat	65%	53%

Table 5.2. Classification results for the bipolar recording dataset.

In order to further verify the classifications, a strategy was applied that any possible pair of test trials from different classes were used as a test set, and each trial was tested 20 times. During the process, large variances for the

average classification accuracies were found. For example, to apply the LDA classifier to both the alpha and beta band powers feature extracted from the reference recording dataset, 20 iterations were carried out. Every iteration included 20 times classifications and the test set in each classification contained two trials from different classes. In the test set, the trial from class 1 was fixed and the trial in class 2 was varied over the whole remaining (20) trials. The classification accuracy in each iteration was obtained by averaging the 20 times classifications and the variance of accuracy in the 20 iterations were calculated. The average accuracy in the time course was 67.92±14.06% and the best accuracy was 91.63±3.17%, where the error ranges correspond to a 95% confidence interval. This result showed the classification was not stable and there was high variability between the recordings even in the same dataset but the high peak accuracy shows that the subject's intentions can be identified very accurately for some periods.

5.3 Comparisons and analysis

During the continuous classification for the EEG signals in period 4-10 s, the classification results when using different features were similar. The peak classification accuracies appeared during the period of 5.5-7.5 s and would decrease in the period of 7.5-8.5 s. As in the case of the Graz dataset, the band powers feature provided the best classification accuracy, but the differences between different features are less pronounced for this dataset. Surprisingly, the time series waveforms with the template matching classifier achieved very stable classifications and the result was comparable to other features. It was found that the Bayesian statistical classifier could not match the performance of the LDA classifier. It is conjectured that this is mainly due to the small number of trials, which meant that the parameters of the Gaussian prototypes could not be estimated accurately.

By comparing the results in the Tables 5.1 and 5.2, it can be seen that the reference recording dataset produced more reliable BCI classification results than the bipolar recording datasets. The only exception was the average results obtained using the method of template matching with the nearest neighbour for the time series waveform feature. The difference between the

results using method 1 and method 2 of template matching is very trivial for the reference recording dataset, whilst method 2 obtained much better accuracies than method 1 for the bipolar recording dataset, even better than the results for the reference dataset. A possible improvement in the data collection setup is to use more electrodes near to C3 and C4, since the peak EEG signal for left and right imagery may not be located exactly at C3 and C4.

It was found that our dataset was not able to provide as high classification accuracies as the Graz dataset in the Chapter 4. In particularly, some very low values of accuracy were obtained from several trials. There could be several reasons for this. Firstly, the subject had not received any training prior to the experiments and secondly, artifact removal processing techniques were not used to eliminate the very noisy sections of EEG signals typically arising from the subject blinking. Thirdly, due to limited access to the data acquisition equipment, the Adelaide dataset contained a much smaller number of trials and also had few channels. Finally, the subject could recall several eye blinks and real arm movements during the experiment, and imagining the wrong movements in some short sections in several of the trials. In the future, it would be useful to record EMG (muscle activity) and EOG (eye movements) for identifying and rejecting invalid or contaminated data. Such steps will likely lead to improvements in the reliability of the BCI classification.

Despite the issues discussed above, the classification results clearly indicated that most of the features studied did contain some discriminative information with respect to the subject's intended imaginary movements. Based on this information, the classifiers were able to identify the thinking activities with reasonable accuracy for many trials. In further experiments, more trials with more channels and more subjects are required.

Chapter 6 Conclusion

This thesis has presented a Brain Computer Interface study based on EEG signals of imaginary hand movements. The BCI is considered as a pattern recognition system and two main parts: feature extraction and classification were investigated and applied to a public dataset and another dataset from experiments conducted at the University of Adelaide.

Several features, specifically, time series waveform, autoregressive (AR) components and spectral components, eigenvector components have been studied in Chapter 2. Several classifiers, such as, template matching, *k*-nearest neighbor, linear discriminant analysis, Bayesian statistical and fuzzy logic classifiers have been studied in Chapter 3. All features and classifiers were applied to the public dataset from an international BCI competition and the results are reported in Chapter 4. Some selected features and classifiers were applied to the dataset recorded from the Adelaide experiment and results are reported in Chapter 5. These features and classifiers were compared in the result analysis sections in Chapter 4 and Chapter 5.

The classification was done in a continuous fashion, to match a real time application. In this process, the average and best accuracy, as well as the computation time were analysed. In the application to the public dataset, most classifiers achieved very high accuracies and short computation times for most features. In particular, the band powers feature provided the highest accuracy for each classifier. The LDA and Bayesian statistical classifiers, which achieved the highest accuracies for 3 kinds of features separately, were found to be the most reliable classifiers in this application. So, they were selected together with the template matching and used for the Adelaide dataset. The results showed that the selected classifiers can work well with this new dataset without much additional preprocessing or modifications, but, for a variety of reasons the classifiers, did not achieve as high accuracies as was obtained using the Graz dataset. Further, the Bayesian statistical classifier performed significantly poorer than in the first dataset, due to the inaccurate parameters estimation limited by small quantity of training data.

Some further work needs to be done in the future. Firstly, more experiments will be required, since the data processing was still limited by size of data. In order to make the research more universal, more subjects should be involved, since there are significant differences between EEG signals from different subjects, such as the pre-stimulus signal amplitude.

Secondly, the quality of the data from the experiment should be improved. The data from the experiment carried out at the University of Adelaide had not been preprocessed in advance, such as for artifacts removal, therefore, more work is needed to determine how additional signal preprocessing can improve the classification accuracy. One important thing would be to record the EMG (muscle activity) and EOG (eye movements) for identifying and rejecting invalid or contaminated data. Also, it may be beneficial to use more electrodes for the data collection, which may provide more spatial information for further processing. This would be especially useful for the study of common spatial pattern feature.

Additionally, the benefit of some short time averaging needs to be investigated. Providing classification information at the data sample rate is obviously at too high a rate for practical applications and the effect of averaging classification results over short period needs be investigated. Unfortunately time did not allow further investigation of this aspect.

Finally, this thesis only studied offline classification and analysis processes, but it is desirable in practical applications for such processing to occur in real time, e.g., control a robot. This requires more work to train subjects and build a closed loop and real time processing system.

Brain Computer Interface can not only provide an important prosthesis to disabled people, but also has the potential to improve human life greatly. With this new communication pathway, many laborious or tedious tasks could become convenient. It is anticipated that more and more human thinking activities will be identified via BCI systems and novel applications will arrive. However, there are still many challenges in BCI research, and further work should be carried out to make this technology really progress beyond laboratory demonstrations.

Bibliography

[1] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T.M.Vaughan, “Brain-computer interface technology: A review of the first international meeting,” IEEE Trans. Rehab. Eng., vol. 8, pp. 164–173, June 2000.

[2] G. Blanchard, B. Blankertz, "BCI Competition 2003—Data Set II a: Spatial Patterns of Self-Controlled Brain Rhythm Modulations," IEEE Trans.biomed.Eng., vol 51, No. 6, June 2004.

[3] R.S. Clemens Brunner, Bernhard Graimann, Gernot Supp,Gert Pfurtscheller, “Online Control of a Brain-Computer Interface Using Phase Synchronization IEEE Trans.biomed. Eng., vol 53, No. 12, December 2006.

[4] G. S. Dean J. Krusienski, Dennis J. McFarland, Jonathan R. Wolpaw, "A μ-Rhythm Matched Filter for Continuous Control of a Brain-Computer Interface," IEEE Trans.biomed. Eng., vol.54, No.2, February 2007.

[5] H. C. Kraemer, Encyclopedia of Statistical Sciences. New York: Wiley, 1982.

[6] J. M.-G. Herbert Ramoser, Gert Pfurtscheller, "Optimal Spatial Filtering of Single Trial EEG during Imagined Hand Movement”, IEEE Trans. Rehab. Eng.,vol. 8, No. 4 December 2000

[7] F. R. José del R. Millán, Josep Mouriño, and Wulfram Gerstner, "Noninvasive Brain-Actuated Control of a Mobile Robot by Human EEG," IEEE Trans.biomed.Eng., vol 51, No. 6, June 2004

[8] N. G. H. M. D. Serruya, L. Paninski, M. R, and J.Donoghue, "Instant neural control of a movement signal," Nature, vol.416, pp. 141–142, 2002.

[9] G. R. M. Reinhold Scherer, Christa Neuper, Bernhard Graimann, and and M. Gert Pfurtscheller, "An asynchronously controlled EEG-Based Virtual Keyboard: Improvement of the Spelling Rate," IEEE Trans.biomed.Eng., vol 51, No. 6, June 2004.

[10] P. B. Y.Wang, and M. Scherg, "Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions A simulation study," Clin. Neurophysiol, vol.110, pp.604–614,1999

[11] Z. Z. Yijun Wang, Yong Li, Xiaorong Gao, Shangkai Gao and Fusheng Yang, "BCI Competition 2003—Data Set IV: An Algorithm Based on CSSD and FDA for Classifying Single-Trial EEG," IEEE Trans.biomed.Eng., vol 51, No. 6, June 2004.

[12] K. M. Nikolaus Weiskopf, Simon W. Bock, Frank Scharnowski, Ralf Veit, Wolfgang Grodd, Rainer Goebel, and Niels Birbaumer, "Principles of a Brain-Computer Interface (BCI) Based on Real-Time Functional Magnetic Resonance Imaging (fMRI)," IEEE Trans.biomed. Eng., vol 51, No. 6, June 2004

[13] N. W. Thilo Hinterberger, Ralf Veit, Barbara Wilhelm, Elena Betta, and Niels Birbaumer, "An EEG-Driven Brain-Computer Interface Combined With Functional Magnetic Resonance Imaging (fMRI)” IEEE Trans.biomed.Eng., vol 51, No. 6, June 2004

[14] Anderson C W and Sijercic Z, “Classification of EEG signals from four subjects during five mental tasks” Solving Engineering Problems with Neural Networks: Proc. Int. Conf. on Engineering Applications of Neural Networks (EANN’96), 1996

[15] B. Blankertz, K.-R. Müller, T. V. G. Curio, G. Schalk, J. Wolpaw, A.Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, and M. S. N. Birbaumer,“The BCI competition 2003: Progress and perspectives in detectionand discrimination of EEG single trials,” IEEE Trans. Biomed.Eng., vol. 51, pp. 1044–1051, June 2004.

[16] M. K. Pradeep Shenoy, Benjamin Blankertz, Rajesh P. N. Rao, Klaus-Robert Müller, "Towards adaptive classification for BCI." Neural Eng., 3:R13-R23, 2006.

[17] P. W. F. Anna Buttfield, and José del R. Millán, "Towards a Robust BCI: Error Potentials and Online Learning," IEEE Trans. Neural systems and Rehabilitation Engnieeing, vol. 14, No. 2, June 2006.

[18] F. H. L. d. S. G. Pfurtscheller, "Event-related EEG/MEG synchronization and desynchronization: basic principles," Clinical Neurophysiology, vol. 110, pp. 1842-1857, 1999.

[19] G. Krausz, R. Scherer, G. Korisek and G. Pfurtscheller, Critical decision-speed and information transfer in the "Graz brain–computer" Applied Psychophysiology and Biofeedback, Vol. 28, No. 3, September 2003

[20] Sergios Theodoridis, Konstantinos Koutroumbas, Pattern Recognition, Academic press, 1999

[21] Blankertz B, Curio G and Muller K R, Classifying single trial EEG: towards brain computer interfacing Adv. Neural Inf. Process. Syst. (NIPS 01) 14 157–64,2002

[22] Borisoff J F, Mason S G, Bashashati A and Birch G E, Brain–computer interface design for asynchronous controlapplications: improvements to the lf-asd asynchronous brain switch, IEEE Trans. Biomed. Eng. 51, 985–92, 2004

[23] M. Loève, Probability theory. Vol. II, 4th ed., Graduate Texts in Mathematics, Vol. 46, Springer-Verlag, 1978

[24] Babiloni, F.; Bianchi, L.; Semeraro, F.; del R Millan, J.; Mourino, J.; Cattini, A.; Salinari, S.; Marciani, M.G.; Cincotti, F.Mahalanobis distance-based classifiers are able to recognize EEG patterns by using few EEG electrodes,Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE

[25] P. Sykacek, S. J. Roberts and M. Stokes. Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation. In IEEE Trans. Biomedical Engineering, 51(5):719--727, 2004.

[26] M. Peltoranta and G. Pfurtscheller, "Neural network based classification of nonaveraged event-related EEG responses," Med. Biol. Eng. Comput.,vol. 32, pp. 189–196, 1994.

[27] A. Schlögl, K. Lugger and G. Pfurtscheller ,Using Adaptive Autoregressive Parameters for a Brain-Computer-Interface Experiment, Proceedings of the 19th Annual International Conference if the IEEE Engineering in Medicine and Biology Society ,vol 19 , pp.1533-1535.1997

[28] Pfurtscheller G, Neuper C, Schlögl A, Lugger K. Separability of EEG signals recorded uring right and left motor imagery using adaptive autoregressive parameters. IEEE Trans Rehabil Eng. 6(3):316-25. 1998

[29] Tanaka, K. Matsunaga, K. Wang, H.O. Electroencephalogram-Based Control of an Electric Wheelchair IEEE Trans Robotics, VOL. 21, NO. 4, August 2005

[30] Garrett D, Peterson D A, Anderson C W and Thaut M H,Comparison of linear, nonlinear, and feature selection methods for EEG signal classification IEEE Trans. Neural Syst. Rehabil. Eng. 11: 141–4, 2003

[31] Guger C, Edlinger G, Harkam W, Niedermayer I, Pfurtscheller G. How many people are able to operate an EEG-based brain-computer interface (BCI)? IEEE Trans Neural Syst Rehabil Eng. 2003;11:145–147.

[32] The 10-20 electrode system of the International Federation HH Jasper - Electroencephalogr Clin Neurophysiol, 1958

[33] Blankertz, B. Tomioka, R. Lemm, S. Kawanabe, M. Muller, K.-R Optimizing spatial filters for robust EEG single-trial analysis, IEEE. signal Proc. Magzine, vol. 25, pp. 41–56, Jan. 2008

[34] Lemm S, Schafer C and Curio G 2004 BCI competition 2003–data set iii: probabilistic modeling of sensorimotor mu rhythms for classification of imaginary hand movements IEEE Trans. Biomed. Eng. 51 1077–80

[35] B Obermaier, C Guger, C Neuper, G Pfurtscheller, Hidden Markov Models for online classification of single trial EEG Data, Recognition Letters archive Volume 22 , Issue 12 2001

[36] http://www.ced.co.uk/pru.shtml

[37] www.brainm.com

[38] Pfurtscheller, G., Stancák, A., Neuper, C., 1996. Event-related synchronization (ERS) in the alpha band — an electrophysiological correlate of cortical idling: a review.Int. J. Psychophysiol. 24, 39–46.

[39] http://pharyngula.org/~pzmyers/neuro/chap9

[40] http://www.incrediblehorizons.com/

[41] J MALMIVUO, R PLONSEY-Bioelectromagnetism - Press: New York, USA, 1995

[42] Leuthardt EC, Miller KJ, Schalk G, Rao RP, Ojemann JG. Electrocorticography-based brain computer interface--the Seattle experience IEEE Trans Neural Syst Rehabil Eng. 2006 Jun;14(2):194-8.

[43] Margalit E et al 2003 Visual and electrical evoked response recorded from subdural electrodes implanted above the visual cortex in normal dogs under two methods of anesthesia J. Neurosci. Methods 123 129–37

[44] A. Kübler , F. Nijboer, J. Mellinger, T. M. Vaughan, H. Pawelzik, G. Schalk, D. J. McFarland, N. Birbaumer, and J. R. Wolpaw, “Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface,” Neurol., vol. 64, pp. 1775–1777, 2005.

[45] E. E. Sutter, "The brain response interface: communication through visually-induced electrical brain responses ‘’J. Microcomput. Appl. , vol. 15 pp. 31-45 1992

[46] G. Pfurtscheller and F. H. L. da Silva, “Event-related EEG/MEG synchronization and desynchronization: Basic principles,” Clin. Neurophysiol vol. 110, pp. 1842–1857, 1999.

[47] S. Roberts and W. Penny, Real-time brain computer interfacing: a preliminary study using Bayesian learning. Medical and Biological Engineering and Computing, 38(1):56–61, 2000.

[48] Croft RJ, Barry RJ. Issues relating to the subtraction phase in EOG artefact correction of the EEG. Int J Psychophysiol 2002;44:187–95.

[49] A. Schlogl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, G. Pfurtscheller. A fully automated correction method of EOG artifacts in EEG recordings. Clin.Neurophys. 2007 Jan;98-104.

[50] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic, 1990.

[51]. K. Fukunaga and WLG Koontz, Application of the Karhunen-Loève expansion to feature selection and ordering. IEEE Trans. Computers C-19 No. 4, 1970

[52] Zachary A. Keirn and Jorge I. Aunon, Man-Machine Communications Through Brain-Wave Processing, IEEE Engineering in Medicne and Biology Magazine. Vol. 3, No. 10, March 1990, 55-57

[53] Townsend, G. Graimann, B. Pfurtscheller, G. A comparison of common spatial patterns with complex band power features in a four-class BCI experiment, Biomedical Engineering, IEEE Transactions on, vol 53, NO. 4, APRIL 2006

[54] Kay, S.M. and S.L. Marple, Spectrum analysis-a modern perspective. Proceedings of the IEEE, 1981. 69(11): p. 1380-1419.

[55] S.J. Johnsen and N. Andersen, On power estimation in maximum entropy spectral analysis, Geophysics 43 (4) (1978), pp. 681–690.

[56] Gene H. Golub and Charles F. Van Loan. Matrix Computations. Johns Hopkins University Press, Baltimore, MD, 2 edition, 1989.

[57]. K. V. Mardia, "Multivariate Analysis", Academic Press, 1979.

[58] Zadeh, L.A. (1965). "Fuzzy sets", Information and Control 8 (3): 338-353

[59] Mamdani, E.H., “Applications of fuzzy logic to approximate reasoning using linguistic synthesis,” IEEE Transactions on Computers, Vol. 26, No. 12,pp. 1182-1191, 1977.

[60] Sugeno, M., Industrial applications of fuzzy control, Elsevier Science Pub. Co., 1985.

[61] Rahn,E. and Basar, E., Pre-stimulus EEG-activity strongly influences the auditory evoked vertex response: a new method for selective averaging. Int J Neurosci 69 1–4, pp. 207–220. 1993.

[62] P. E. Hart, The condensed nearest neighbor rule, Trans. IEEE Inform. Theory, IT-14, pp. 515-516, 1968.

[63] K. Fukunaga and P. M. Narendra, A branch and bound algorithm for computing k-nearest neighbors, Trans. IEEE Computers, C-24, pp. 750- 753, 1975.

[64] Berkan, R. C. Trubatch, S. L. Fuzzy Systems Design Principles : Building Fuzzy If-Then Rule Bases, Wiley-IEEE Press, May 1997

[65] http://www.mathworks.com/matlabcentral/fileexchange/8636

[66] Hastie T, Tibshirani R, Friedman JH, The elements of statistical learning: data mining, inference, and prediction, 2nd ed. New York: Springer, Feb 2009.

[67] Toga, A. W. and J. C. Mazziotta, Brain Mapping: The Methods Academic Press 2002.

[68] Drachman, D. A. “DO we have brain to spare?” Neurology 64(12): 2004- 2005.

[69] Nunez, P. L. and R. Srinivasan, Electric Fields of the Brain: The Neuro physics of EEG Oxford University Press, 2005.