Check for updates

Filtering techniques for channel selection in motor imagery EEG applications: a survey

Muhammad Zeeshan Baig[1] · Nauman Aslam[1] · Hubert P. H. Shum[1] ⓘ

Abstract

Brain computer interface (BCI) systems are used in a wide range of applications such as communication, neuro-prosthetic and environmental control for disabled persons using robots and manipulators. A typical BCI system uses different types of inputs; however, Electroencephalography (EEG) signals are most widely used due to their non-invasive EEG electrodes, portability, and cost efficiency. The signals generated by the brain while performing or imagining a motor related task [motor imagery (MI)] signals are one of the important inputs for BCI applications. EEG data is usually recorded from more than 100 locations across the brain, so efficient channel selection algorithms are of great importance to identify optimal channels related to a particular application. The main purpose of applying channel selection is to reduce computational complexity while analysing EEG signals, improve classification accuracy by reducing over-fitting, and decrease setup time. Different channel selection evaluation algorithms such as filtering, wrapper, and hybrid methods have been used for extracting optimal channel subsets by using predefined criteria. After extensively reviewing the literature in the field of EEG channel selection, we can conclude that channel selection algorithms provide a possibility to work with fewer channels without affecting the classification accuracy. In some cases, channel selection increases the system performance by removing the noisy channels. The research in the literature shows that the same performance can be achieved using a smaller channel set, with 10–30 channels in most cases. In this paper, we present a survey of recent development in filtering channel selection techniques along with their feature extraction and classification methods for MI-based EEG applications.

Keywords Channel selection · EEG · Filter method · BCI · Motor imagery

✉ Hubert P. H. Shum
   hubert.shum@northumbria.ac.uk

   Muhammad Zeeshan Baig
   baig.mzeeshan@gmail.com

   Nauman Aslam
   nauman.aslam@northumbria.ac.uk

[1] Faculty of Engineering and Environment, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

1 Introduction

The EEG signals provide information about the electrical activity in the brain which plays a vital role in many useful applications/systems designed to improve quality of life for the disabled people (Wolpaw et al. 2000). Examples of such applications are communication, neuro-prosthetic, seizure detection, sleep state classification and environmental control for disabled persons using robots and manipulators. Processing and analysis of EEG signals generally consist of a signal acquisition part followed by feature extraction and classification techniques as shown in Fig. 1.

The signal acquisition part as shown Fig. 1 can be carried out on the scalp surface or interior of the brain. In a typical EEG signal acquisition system, a number of electrodes are used as sensors to record the voltage level. These electrodes can be invasive, mounted into the skull, or non-invasive, mounted on the surface of the skull. The placement of these electrodes over a number of points facilitate recording the EEG data from more than one location on the brain. Once the data is recorded it is processed to remove noise and artefacts resulting from body movement of the subjects, outside electrical interference and electrodes pop, contact and movement (Shao et al. 2009).

To differentiate between the brain's states the signals are transformed or filtered to find the variables that effectively separate out different states of the brain, this process is known as feature extraction. The purpose of EEG feature extraction is to generate discriminative features from channels data that can increase the variance difference between classes. The last part is to efficiently classify the EEG signals and generate decision signals. The device that uses brain signals to control and operate the environment is called the Brain Computer Interface (BCI) system (Wolpaw et al. 2002). In the last two decades, the BCI, due to its numerous benefits, has been elevated to great significance in industry and scientific institutes (Ortiz-Rosario and Adeli 2013). The main advantage of using the BCI is its ability to reduce the risk to human life and increase the quality of life (improve daily living activities) for disabled persons.

There are many sources to power a BCI system that belongs to the category of invasive and non-invasive inputs. For invasive input, where electrodes are mounted into the scalp, electro-corticograpy (ECoG) (Hill et al. 2006b), single micro-electrode (ME), micro-electrode array (MEA) (Van Gerven et al. 2009) and local field potentials (LFPs) (Moran 2010) can be used to acquire signals. Electroencephalography (EEG) (Schalk et al. 2004), magnetoencephalography (MEG) (Lal et al. 2005), Functional Magnetic Resonance Imaging (fMRI) (Weiskopf et al. 2004) and Near Infra-red Spectroscopy (NIRS) (Coyle et al. 2007) are examples of non-invasive inputs . EEG powered BCI are the most reliable and are frequently used because of their easy availability, non-invasiveness, low cost and high temporal resolution (Coyle et al. 2007; Hill et al. 2006a). A generic graph between temporal resolution and spatial resolution of different signal acquisition techniques is presented in Fig. 2.

Acquiring of EEG signals for BCI can be done in different ways. Some of the methods require an event to generate the EEG signals and some are event independent. Motor Imagery (MI) is one of the methods used to generate EEG signals that are related to motor movements.

Fig. 1 Steps of EEG signal processing and analysis

Fig. 2 Comparison of spatial and temporal resolution (Hill et al. 2006b)
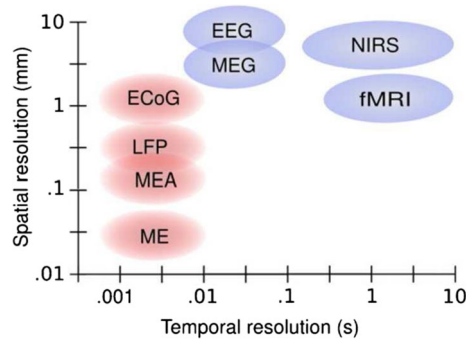
MI-based EEG signals have been applied to many BCIs applications where these signals have been controlled to open an interface with the external environment (Pineda 2005). These signals can be extracted from different positions over the brain. The most widely used system for electrode placement is the International 10–20 system, recommended by The International Federation of Societies for Electroencephalography and Clinical Neurophysiology (IFSECN). The placement of EEG electrodes is also related to the brain activities and is application dependent (Daly and Pedley 1990).

Most of the useful brain state information lies in low-frequency bands; therefore the frequency of EEG signals was divided into 6 bands. Each frequency band corresponds to different functions of the brain. The frequency range of 0–4 Hz is called the delta band and it's usually related to the deep sleep states that require continuous attention (Sharbrough 1991). The theta band ranges from 4 to 7 Hz and normally corresponds to drowsiness or a state in which the body is asleep and the brain is awakened (Kirmizi-Alsan et al. 2006). Alpha waves (8–15 Hz) are normally found in relaxing and closed eyes state (Niedermeyer 1997). The beta band is from 16 to 31 Hz and it contains the information of active concentration and working (Pfurtscheller and Da Silva 1999). 30 Hz or more is called the gamma band which relates to sensory processing and short term memory matching (Kanayama et al. 2007). The last band, mu ranges from 8 to 12 Hz; same as the alpha band but is recorded over the sensorimotor cortex and contains information of rest state motor neurons (Gastaut 1952). To analyse certain states or illness, brain waves are processed for the proper diagnosis of a particular state.

Development of on-line BCI faces many challenges including computational cost, equipment cost and classification accuracy. To successfully address these challenges, researchers have suggested different algorithms. For example, signal pre-processing, feature extraction and selecting an appropriate classifier helps in increasing classification accuracy of a BCI. Multi-channel EEG gives a variety of possibilities to applications but specific channel selection is better suited for outcomes. Researchers usually ignore the channel selection part while developing a real-time system. The problem with not using channel selection algorithm results in noisy data and redundant channels increase the computational and equipment cost of a BCI system. Channel selection also improves or stabilizes the classification results (Faller et al. 2014). The same problem also appears while doing feature extraction. Therefore, it is vital to opt for an effective solution to select the optimal number of channels rather than using all channels for processing and classification. Some researchers have used feature selection algorithms after applying the channel selection algorithms to further improve the system performance (Handiru and Prasad 2016). Subset channels are selected based on certain criteria

that usually incorporates all aspects including channel location, dependency and redundancy (Garrett et al. 2003).

It is important to note that a viable BCI system for medical applications relies heavily on efficient algorithms for accurate predictions of any disease or abnormal physiological conditions. Hence, the channel selection part plays a vital role in designing efficient algorithms. Selecting a minimum number of channels decreases the computation complexity, cost and results in low power hardware design. Considering the importance of the channel selection process in BCI systems, we studied recently developed channel selection schemes based on filtering techniques for MI-based EEG signals. The survey focuses on applications involving motor imagery applications. The explanation and discussion are supported by flowcharts and tables to present a clear understanding to the reader. Clear and informative comparison of different channel selection algorithms is given based on classification algorithm, channel selection strategy, and dataset. In addition, performance of the discussed method is also given, in terms of classification accuracy and number of channels selected, to provide some assistance to BCI programmers and researchers to choose appropriate selection algorithms for different applications.

The rest of the paper is structured as follow: Sect. 2 deals with channel selection techniques in general. Channel selection algorithms and specifically filtering techniques of channel selection for MI-based EEG are described in Sect. 3. Section 4 is devoted to discussion and conclusion.

2 Channel selection techniques

EEG-based BCI has become a hot field of research during the last few decades due to the availability of EEG signal acquisition systems. As the EEG equipment is cheap compared to other systems, it allows us to record data related to brain activity with a large number of channels. A large number of channels allows researchers to develop techniques for selecting optimal channels. The objective of these algorithms is to improve computation time, classification accuracy and identify channels best suited to a certain application or task.

The algorithms used for EEG channel selection are derived from feature selection algorithms available in the literature. Selecting the optimal subset of features is known as feature selection thus using these algorithms to select channel is known as channel selection. In channel selection, features are extracted for classification after selecting the optimal channel set. However, in feature selection, the optimal feature set supplied directly to the classification algorithm. In filtering channel selection techniques, the features extracted from the optimal channel subset may or may not produce good results. On the other hand, for wrapper and hybrid channel selection techniques, feature extraction and classification are part of the selection procedure, so they produce the best results, which, however, are classifier and subject specific.

The credibility of a feature subset is evaluated by a criterion. The number of features is directly proportional to the dimensionality of the signal, so higher dimension signals have more features. Finding an optimal subset is difficult and known to be an NP hard problem (Blum and Rivest 1989). Typical feature selection algorithms involve four stages as shown in Fig. 3. The subset generation stage is done using a heuristic search algorithm such as complete search, random search or sequential search to generate candidate subset of features for evaluation purposes. Each new candidate feature subset is evaluated and its results are compared with the previous best one according to evaluation criterion.
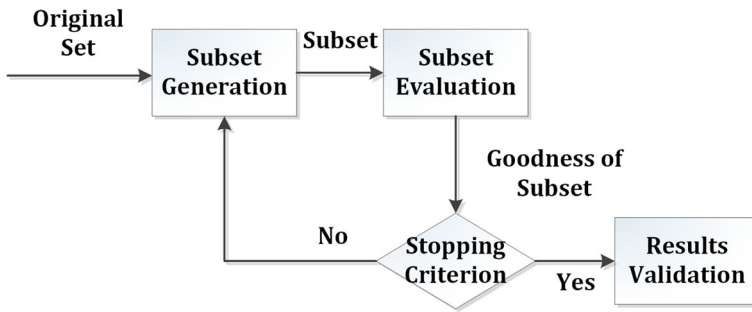
Fig. 3 Feature selection process

If the new subset has better results compared to the previous one, the new candidate will take the position of the previous subset. The subset generation and evaluation process will be repeated until a stopping criterion is fulfilled. Then the selected optimal subset is evaluated using some classifier or prior knowledge of the problem.

Feature selection processes can be found in many fields of machine learning and data mining. In statistics, feature selection is also called variable selection and these techniques are also used for samples and channel selection (Liu and Yu 2005). There are four main sorts of techniques available to evaluate features subset, namely, Filtering, Wrapping, Embedded and Hybrid techniques that are discussed in further details in the following subsections.

Algorithm 1: A Generalized Filter Algorithm

Data:
$D(f_0, f_1, \ldots, f_{n-1})$ `// Training dataset with n features`
S_0 `// Search starting subset`
δ `// Stopping Criterion`
Result: S_{best} `// Optimal subset`
Initialize $S_{best} = S_0$;
$\gamma_{best} = evaluate(S_0, D, M)$;
`// evaluate `S_0` by an independent measure M`
while δ *is achieved* **do**
$\quad$ $S = generateSubset(D)$;
$\quad$ $\gamma = evaluate(S, D, M)$;
$\quad$ **if** γ *is better than* γ_{best} **then**
$\quad\quad$ $\gamma_{best} = \gamma$;
$\quad\quad$ $S_{best} = S$;

return S_{best};

2.1 Filtering techniques

Filtering techniques use an autonomous assessment criterion such as distance measure, dependency measure and information measure to evaluate the subset generated by a search algorithm (Guyon and Elisseeff 2003; Dash and Liu 1997). Most of the filtering techniques are high-speed, stable, providing independence from the classifier, but of low accuracy (Chandrashekar and Sahin 2014). Algorithm 1 shows a generalized algorithm for filter based channel selection algorithms. Let D be the given data set and S_0 is the subset candidate generated by a search algorithm, S_0 can be an empty, full or randomly selected subset and it propagates to

the whole data set through a particular search strategy. An independent measure M has been used to assess the generated candidate subset S. After evaluation the subset is compared with the previous best subset and if it is better than the previous one, it is stated as the current best subset. This process continues in an iterative manner until a stopping criterion δ is achieved. The final output of the algorithm is the last best subset denoted by S_{best}. The commonly used methods to find relevancy and dependency are discussed.

2.1.1 Correlation criteria

The most simple and commonly used criteria to find the correlation between variable and target is Pearson coefficient (Guyon and Elisseeff 2003):

$$R(i) = \frac{cov(x_i, Y)}{sqrt(var(x_i) * var(Y))} \tag{1}$$

where x_i is the ith variable and Y is the output target class. Linear dependencies between variable and output can be detected with correlation criteria.

2.1.2 Mutual information (MI)

Mutual information is a measure to find the dependency between two variables (Lazar et al. 2012). The Mutual Information is calculated as:

$$I(X, Y) = H(Y) - H(Y|X) \tag{2}$$

where I is the mutual information, $H(Y)$ is the entropy of Y and $H(Y|X)$ is entropy of variable Y observing a variable X. Mutual information will be zero if X and Y are independent and greater than zero if they are dependent. Some researchers also use Kullback–Leibler divergence (Torkkola 2003) formula to measure mutual information between two densities. After measuring the mutual information, the next step is to rank the features through a threshold. The problem with mutual information is that it ignores the inter-feature mutual information. Another common variation of mutual information used in the literature is conditional mutual information (Fleuret 2004).

2.1.3 Chi-squared

The chi-squared statistic is a univariate filter that evaluates feature independently. The value of chi-squared is directly proportional to the relevance of the feature with respect to the class (Greenwood and Nikulin 1996). The chi-squared statistic for a feature is measured as:

$$X^2 = \sum_{i=1}^{V} \sum_{j=1}^{B} \frac{[A_i j R_i * B_j/N]^2}{R_i * B_j/N} \tag{3}$$

where V is the number of intervals, N is the total instances, B is the total number of classes, R_i is the number of instances in the range, B_j is the number of instances in class j_{th}, and A_{ij} is the number of instances in the range i and class j. Various other techniques are used in the literature to validate feature subset and classifier for data with unknown distribution including consistency based filter (Dash and Liu 2003), INTERACT algorithm (Zhao and Liu 2007), Relief and ReliefF algorithm (Kononenko 1994), minimum redundancy maximum relevance (mRMR) (Peng et al. 2005).

Different filtering techniques can be developed by changing the search strategy for subset generation and the evaluation function for assessing the independent measure of each subset (Liu and Yu 2005). This paper focuses on filtering techniques for channel selection because wrapper and hybrid selection techniques heavily rely on the selection of classifier and the subject. However, filter techniques are subject and classifier independent as these techniques use other signal information related parameters to select the best channel subset. Most filtering techniques are based on some statistical criteria for channel selection. Measures based on location (Faul and Marnane 2012), variance (Duun-Henriksen et al. 2012), mutual information (Atoufi et al. 2009), redundancy and relevancy (Yang et al. 2013; Shenoy and Vinod 2014) are more common for filtering techniques. In MI applications, Common Spatial Pattern (CSP) based measures are mostly used to rank the channels for selection (Alotaiby et al. 2015).

Algorithm 2: A Generalized Wrapper Algorithm

Data:
$D(f_0, f_1, \ldots, f_{n-1})$ // Training dataset with n features
S_0 // Search starting subset
δ // Stopping Criterion
Result: S_{best} // Optimal subset
Initialize $S_{best} = S_0$;
$\gamma_{best} = evaluate(S_0, D, A)$;
// evaluate S_0 by a mining technique A
while δ *is achieved* **do**
    $S = generateSubset(D)$;
    $\gamma = evaluate(S, D, A)$;
    if γ *is better than* γ_{best} **then**
        $\gamma_{best} = \gamma$;
        $S_{best} = S$;

return S_{best};

2.2 Wrapper techniques

The wrapper method uses a predictor and its output as an objective function to evaluate the subset. The advantage of using wrapper techniques is the accuracy as wrappers generally achieve better recognition rates than filters since wrappers are tuned to the specific interactions between the classifier and the dataset. The wrapper has a mechanism to avoid over-fitting due to the use of cross-validation measures of predictive accuracy. The disadvantages are slow execution and lack of generality: the solution lacks generality since it is tied to the bias of the classifier used in the evaluation function. The "optimal" feature subset will be specific to the classifier under consideration (Chrysostomou 2009). Pseudo code for a generalized wrapper algorithm is given in Algorithm 2 which is quite similar to the filter algorithm except that the wrapper utilizes a predefined mining or classification model A rather than independent measure M for subset evaluation. For each candidate subset S, the wrapper evaluates the goodness of the subset by applying the model A to the feature subset S. Therefore, changing the search function that generates the subset and prediction model A can result in different wrapper algorithms (Liu and Yu 2005).

Most wrapper methods used searching algorithms to find a set of optimal electrodes. The optimization function mostly rotates around to maximize the performance and minimize the

number of channels for a certain range of accuracy. Sequential forward and backward search (Shih et al. 2009; Kamrunnahar et al. 2009) as well as heuristic/random search (Millán et al. 2002; Wei and Wang 2011) are the widely used search algorithms in the literature. The following subsection presents key search algorithms in more detail.

2.2.1 Sequential selection algorithms

These algorithms iteratively search the whole feature space to select the best features. The most common algorithm was the sequential feature selection (SFS) (Pudil et al. 1994), which started with an empty set and added the feature that generates the maximum value for the objective function. In the next step, the remaining features were added individually, and the new subset was evaluated. The reverse of SFS was known as sequential backward selection (SBS) that started with all features and removed the feature whose removal minimally affects the objective function performance (Reunanen 2003). A more flexible algorithm was the sequential floating forward selection (SFFS), which introduced a backtracking step in addition to SFS (Reunanen 2003). The backtracking step removed one feature at a time from the subset and evaluated the new subset. If the removed feature maximized the objective function, the algorithm went back to the first step with the new reduced features. Otherwise, it repeated the steps until the required number of features or performance was achieved. The main problem with the SFS and SFFS algorithms was the nested effect, which meant that two features with a relatively high correlation might be included in the subset because they gave the maximum accuracy. To avoid this, adaptive SFFS (Somol et al. 1999) and Plus-L-Minus-r search method (Nakariyakul and Casasent 2009) were also developed.

2.2.2 Heuristic search algorithms

Heuristic algorithms were designed to solve a problem in a faster and more efficient way compared with the traditional methods. A heuristic is the rule of thumb that sacrifice optimality, accuracy, precision or completeness to find a solution. Usually, these algorithms are used to solve NP-complete problems. The most commonly used heuristic algorithms are genetic algorithms (Davis 1991), particle swarm optimization (Kennedy 2011), simulated annealing (Yang 2014), ant colony optimization (Dorigo et al. 2008), and differential evolution (Baig et al. 2017) etc.

A recent method that used the inconsistencies in classification accuracy after adding a noisy channel as a tool for selecting channels has been proposed (Yang et al. 2014). A predefined classification criterion was set and if such a criterion was met after including a channel, the channel would be selected. The technique used SVM, naive Bayes, LDA and decision trees classifier to test a channel. It recorded an average increase of maximum 4% in comparison with SVM, mutual information, CSP, and fisher criterion-based channel selection methods. It also found that the selected channels were mainly located on the motor and somatosensory association cortex area. In another work, Ghaemi et al. (2017) used an improved Binary Gravitation Search Algorithm (IBGSA) to automatically select the optimal channels for MI-based BCI. The algorithm used SVM as a classifier and both the time domain and the wavelet domain features were used. A classification accuracy of $55.82 \pm 8.30\%$ was achieved with all channels. The accuracy increased further to 60% after applying PCA for feature reduction. With the proposed channel selection algorithm, an accuracy of $76.24 \pm 2.78\%$ was achieved with an average of 7 channels.

Recently, researchers have been working on understanding the applications of evolutionary algorithms for channel selection. Most of the literature used the genetic algorithm and its

variations. Antonio et al. ([2012](#)) proposed a channel selection algorithm for imagined speech application. The method searched for a non-dominant channel combination using a multi-objective wrapper technique based on NSGA-II (Elitist Non-Dominated Sorting Genetic Algorithm) in the first stage called Pareto front. In the next stage, the algorithm assessed each channel combination and automatically select one combination using the Mamdani fuzzy inference system (FIS). The error rate and the number of channels selected were used as a multi-objective optimization function. In comparison to the classification accuracy of 70.33% with all channels, the algorithm achieved a classification accuracy of 68.18% with only 7 channels. A multi-objective genetic algorithm (GA) was used to optimize the number of channel selected and classification accuracy. The effectiveness of GA in channel selection applications was shown by considering three different variations of GA i.e. simple GA, steady-state GA, and NSGA-II (Kee et al. [2015](#)). The algorithms showed an increase of almost 5% in classification accuracy with an average of 22 selected channels.

Algorithm 3: Typical hybrid feature selection technique

Data:
$D(f_0, f_1, \ldots, f_{n-1})$ // Training dataset with n features
S_0 // Search starting subset
δ // Stopping Criterion
Result: S_{best} // Optimal subset
Initialize $S_{best} = S_0$;
$C_0 = cardinality(S_0)$;
$\gamma_{best} = evaluate(S_0, D, M)$;
// evaluate S_0 by an independent measure M
$\theta_{best} = evaluate(S_0, D, A)$;
// evaluate S_0 by a mining technique A
for $C = C_0 + 1$ *to* N **do**
  for $i = 0$ *to* $N - C$ **do**
    $S = S_{best} \cup F_j$;
    // generate a subset with cardinality C for evaluation
    $\gamma = evaluate(S, D, M)$;
    if γ *is better than* γ_{best} **then**
      $\gamma_{best} = \gamma$;
      $S'_{best} = S$;
  $\theta = evaluate(S'_{best}, S, A)$;
  if θ *is better than* θ_{best} **then**
    $S_{best} = S'_{best}$;
    $\theta_{best} = \theta$;
  return S_{best};

2.3 Hybrid techniques

Hybrid techniques are the combination of the above two techniques and eliminate the pre-specification of a stopping criterion. The hybrid model was developed to deal with large datasets. Algorithm [3](#) shows a typical hybrid technique algorithm that utilizes both an independent measure M and a mining algorithm A to evaluate the fitness of a subset. The algorithm starts its search from a given subset S_0 and tries to find the best subset in each round while also increasing the cardinality. The parameters/variables γ_{best} and θ_{best} correspond to cases

with and without classifier respectively, and are calculated in each round. The quality of results from a mining algorithm offers a natural stopping criterion (Liu and Yu 2005).

Li et al. (2011) used the L1 norm of CSP to first sort out the best channels and used the classification accuracy as an optimization function. With this method, an average accuracy of 90% was achieved with only 7 channels. The algorithm was tested on the Dataset IVa of BCI competition III. Handiru et al. proposed an iterative multi-objective optimization for channel selection (IMOCS) method that initializes the C3 and C4 channels as the candidates and updated the channel weight vector in each iteration. The optimization parameters were the ROI (motor cortex) and the classification accuracy. To terminate the iterative channel vector update function, a convergence metric based on the inter-trial deviation was used. The dataset used to evaluate the algorithm was Wadsworth dataset (Schalk et al. 2004). The proposed approach achieved a classification accuracy of 80% and a cross-subject classification accuracy of 61% for untrained subjects.

2.4 Embedded techniques

There is another type of technique known as embedded technique. In the embedded techniques, the channel selection depends upon the criteria created during the learning process of a specific classifier because the selection model is incorporated into the classifier construction (Dash and Liu 1997). Embedded techniques reduce the computation time taken up in reclassifying different subset, which is required in wrapper methods. They are less prone to over-fitting and require low computational complexity. Two commonly used embedded feature selection techniques are given below.

2.4.1 Recursive feature elimination for SVM (SVM-RFE)

This method performs iterative feature selection by training SVM. The features that have the least impact on the performance indicated by the SVM weights are removed (Chapelle and Keerthi 2008). Some other method uses statistical measures to rank the features instead of using a classifier. Mutual information and greedy search algorithm have been used to find the subset in (Battiti 1994). Kwak and Choi (2002) uses Parzen window to estimate mutual information. Peng et al. (2005) uses mRMR instead of mutual information to select features.

2.4.2 Feature selection-perceptron (FS-P)

Like SVM-RFE, multilayer perceptron network can be used to rank the features. In a simple neural network, a feedforward approach is used to update the weights of perceptron, which can be used to rank features (Romero and Sopena 2008). A cost function can be used to eliminate features randomly (Stracuzzi and Utgoff 2004).

There are some other methods that use unsupervised learning techniques to find the optimal subset of features. With unsupervised learning-based feature selection, a better description of data can be achieved. Ensemble feature selection technique is used in the literature to find a stable feature set (Haury et al. 2011; Abeel et al. 2009). The idea behind this technique is that different subsets are generated by a bootstrapping method and a single feature selection method is applied on these subsets. The aggregation method such as ensemble means, linear and weighted aggregation have been used to obtain the final results.

In this survey, we are focused on filtering techniques used in channel selection for MI-based EEG signals. A review of channel selection algorithms in different fields indicates the

application of channel selection algorithm in improving performance and reducing computation complexity. Channel selection algorithms have been applied in seizure detection and prediction, motor imagery classification, emotion estimation, mental task classification, sleep state and drug state classification (Alotaiby et al. 2015).

3 Channel selection for motor imagery classification

The analysis of motor imagery is of keen importance to the patients suffering from motor injury. Such analysis can be performed through EEG signals and may involve channel selection to deduce the channels that are the most related to a specific cognitive activity, as well as to reduce the overall computation complexity of the system.

Filtering techniques use autonomous criteria such as mutual information or entropy to evaluate the channel. Filtering is a common technique used in channel selection for MI classification because these techniques are based on signals statistics. These techniques can be divided into CSP-based and non-CSP based techniques.

3.1 CSP based filtering techniques

Common Spatial Pattern (CSP) filters are often used in the literature to study MI-EEG. The reason is that CSP has the ability to maximize the difference in variance between the two classes (Koles et al. 1990).

3.1.1 CSP variance maximization methods

In the literature, researchers have used CSP filters for the maximization of variance between classes. Wang et al. (2006) used maximum spatial pattern vectors from common spatial pattern (CSP) algorithms for channel selection. The channels were selected using the first and last columns of a spatial pattern matrix as they corresponded to the largest variance of one task and the smallest variance of the other task. The features selected for classification were averaged time course of event related de-synchronization (ERD) and readiness potential (RP) because they provide considerable discrimination between two tasks. Fisher Discriminant (FD) was applied as a classifier and a cross-validation of 10×10 was used to evaluate accuracy. The dataset used was from the BCI competition III, dataset IVa provided by Fraunhofer FIRST (Intelligent Data Analysis Group) and University Medicine Berlin (Neurophysics group) (Dornhege et al. 2004). The obtained results were divided into two parts based on the selected channels. In the first part, 4 channels were selected and achieved the combined classification accuracy was 93% and 91% for two subjects. In the case of 8 channels, classification accuracy was increased i.e. 96% and 93% for subject 1 and 2 respectively, at the cost of reducing the suitability of the system.

Another channel selection method based on the spatial pattern was developed by Yong et al. (2008). The authors redefined the optimization problem of CSP by incorporating an L1 norm regularization term, which also induced sparsity in the pattern weights. The problem was to find filters or weight vectors that could transform the signal into a one dimensional space where one class has maximum variance and other has minimum. The high variance corresponded to strong rhythms in EEG signal because of the RP and the low variance are related to attenuated rhythms, which are generated during a right hand imagined movement or when an ERD occurs. This was solved with an optimization process, in which the cost function

was defined as the variance of the projected class of one signal. Such a cost function was minimized while the combined sum of variance of both classes remained fixed. On average, 13 out of 118 channels were selected for classification, which generated a classification accuracy of 73.5%. The classification accuracy by incorporating all the channels in the classification procedure was 77.3%, so a small drop of 3.8 % was recorded while reducing the channels by an average of 80%. In this method, the regularization parameter was selected manually. To produce the optimal results, it should be selected automatically.

Meng et al. (2009) presented a heuristic approach to select a channel subset based on L1 norm scores derived from CSP. The channels with a larger score were retained for further processing. The CSP was implemented twice on the signals, making it a complex optimization problem that has to be solved heuristically in two stages because sometimes CSP can be affected by over-fitting due to a large number of channels. In stage one, the L1 norm was calculated to select channels and CSP was performed on the selected channels to generated features in the second stage. The score of each channel was evaluated based on L1 norm. The channels with the highest scores only were retained for further processing. In the second stage, CSP was applied and the features were passed to the classifier.

The method was compared with the commonly used γ^2 value in which each channel was scored based on a function dependent upon the samples, mean and standard deviation of the classes (Lal et al. 2004). Support vector machine (SVM) with a Gaussian radial basis function (RBF) was used as a classifier. 10×10 cross-validation was used to evaluate classification accuracy. The result of Meng et al. (2009) outperforms manual selection of electrodes and γ^2 value for all 5 subjects with an average classification accuracy of 89.68% with a deviation of 4.88%

3.1.2 Methods based on CSP variants

Some researchers modified the CSP algorithm for extracting the optimal number of channels. He et al. (2009) presented a channel selection method using the minimization of Bhattacharyya bound of CSP. Bayes risk for CSP was used as an optimization criterion. Calculating Bayes risk directly was a difficult task, so an upper bound of Bayes risk was applied as a substitution for measuring discriminability, which was known as Bhattacharyya bound. After finding the optimal index through Bhattacharyya bound of CSP a sequential forward search was implemented for extracting a subset of channels. Features were extracted through CSP with a dimension vector of 6. The dataset 1 of BCI competition IV was utilized in the experiment (Naeem et al. 2006). The authors selected data from subjects a, b, d and e each with 200 trials. Naïve Bayes classifier was applied for classification and an average classification accuracy of 90% with an average of $\sim$ 33 electrodes out of 59 using 10 times 3 fold cross validation.

Tam et al. (2011) selected channels by sorting the CSP filter coefficients, known as the CSP rank. The CSP generate two filters for two classes. These spatial filter coefficients then generated new filtered signals that were considered as weights assigned to different channels by the CSP. If the weight of a particular electrode was large, then it would be considered as contributing more towards the filtered signal, and hence the electrode was considered important. The first electrode was the one that had the largest value from the sorted coefficient of class 1 filter. The second channel was from the sorted coefficients of class 2 filter. If the channel was already selected, the algorithm would move to next largest coefficient in the same class until a new channel was selected. The data was recorded from five chronic stroke patients using a 64 channels EEG headset at a sampling rate of 250 Hz over 20 sessions of MI-tasks and each session was performed on a different day (Meng et al.

2008). The proposed algorithm showed an average classification of 90% with electrodes ranging from 8 to 36 from a total of 64 with Fisher linear Discriminant (FLD) classifier. The best classification was 91.7% with 22 electrodes.

The results showed that SCSP outperformed the existing channel selection algorithms including Fisher Discriminant (FD), Mutual Information, SVM, CSP and RCSP. The algorithm attained a 10% improvement in classification accuracy by reducing the number of channels compared to the three channel case (C3, C4 and Cz). The average classification accuracy of SCSP1 was 81.63 and 82.28% with an average number of channels 13.22 and 22.6 for dataset IIa and IVa respectively. The average classification rate of SCSP2 was 79.09 and 79.28% with an average number of channels 8.55 and 7.6 for dataset IIa and IVa respectively. The time taken by the algorithm to converge was 50.1 s with 1001 iteration on Dataset IVa and 2.63 s with 454 iterations on dataset IIa. The 10×10 cross-validation was used to evaluate classification performance with SVM classifier.

One of the hurdles encountered in analysing an EEG signal is its non-stationary nature. EEG signals differ from the session to session due to the preconditions of the subject. So there is a need of adaptive or robust algorithms to tackle these variations. Arvaneh et al. (2012) presented a Robust Sparse CSP (RSCSP) algorithm to deal with session channel selection problems in BCI. The pre-specified channel subset selection was based on experience. They replaced the covariance matrix in SCSP with a robust minimum covariance determinant (MCD) estimate that involved a parameter to resist the outlier.

To calculate the credibility of proposed algorithms, a comparison was carried out with five existing channel selection algorithms. The data was recorded from 11 stroke patients across 12 different sessions of motor imagery. Data was recorded over 27 electrodes at a sampling rate of 250 Hz (Ang et al. 2011). Eight electrodes were shortlisted with the RSCSP algorithm and the same channels were used across 11 other sessions. The results showed that the proposed algorithm outperformed the others such as SCSP by 0.88%, CSP by 2.85%, Mutual Information (MI) by 2.69%, Fisher Criterion (FC) by 4.85% and SVM by 4.58% over all 11 subsequent sessions. The overall average classification accuracy was 70.47% for RSCSP, which was not as good as it should be. A possible reason could be not using a search strategy to get an optimal subset of channels. Instead, subset definition were generated based on experience. Table 1 shows the summary of CSP based channel selection algorithms for motor imagery EEG applications.

3.2 Non-CSP based filtering techniques

Some researchers have proposed non-CSP based algorithms for MI channel selection. Some of the recent work is summarized in this section.

3.2.1 Information measure based methods

Information measures have been used frequently in selecting channels for EEG applications. He et al. (2013) proposed an improved Genetic Algorithm (GA) that involved Rayleigh coefficients (RC) for channel selection in motor imagery EEG signals. Maximization of Rayleigh coefficients was performed not only to maximize the difference in the covariance matrices but also to minimize the sum of them. Like common spatial patterns, the Rayleigh coefficients were also affected by redundant channels. The authors proposed a Genetic Algorithm based channel selection algorithm that utilized Rayleigh coefficient maximization. The first stage of algorithm was to delete some channels to reduce computation complexity using fisher

Table 1 Summary of CSP based channel selection method for motor imagery EEG

Techniques	Channel selection Strategy	Classifier	Performance metrics (average)	No. of channels selected/total no. of channels (average)	Dataset
Wang et al. (2006)	Max. of spatial pattern vector	Fisher discriminant (FD)	92.66% (max)	4/118	BCI competition III dataset IVa
			94.96%(max)	8/118	
Yong et al. (2008)	Maximizing variance of CSP	Linear discriminant analysis (LDA)	73.5%	13/118	BCI competition III dataset IVa
He et al. (2009)	Minimization of Bhattacharyya bound of CSP	Naïve Bayes	∼95%	∼33/59	BCI competition IV dataset I
Meng et al. (2009)	Heuristic algorithm based on $L1$ norm	SVM with Gaussian RBF	89.68%	20/118	BCI competition III dataset IVa
Tam et al. (2011)	CSP Rank	Fisher linear discriminant (FLD)	91.7% (max)	22/64	BCI-FES training platform
Arvaneh et al. (2011)	Recursive feature elimination using sparse CSP (SCSP)	SVM	81.63% (SCSP1)	13.22/22	BCI competition IV dataset IIa
			79.09% (SCSP2)	8.55/22	
			82.28% (SCSP1)	22.6/118	BCI competition III dataset IVa
			79.28% (SCSP2)	7.6/118	
Arvaneh et al. (2012)	Recursive feature elimination using robust sparse CSP (RSCSP)	SVM	70.47%	8/27	Stroke patients EEG dataset Ang et al. (2011)

ratio. The first channels subset was constructed using the electrodes with maximum fisher ratio. In the second stage, an improved genetic algorithm was purposed that utilized Rayleigh coefficient maximization for selecting the optimal channel subset.

The proposed algorithm was assessed on two sets of data. The first data was from BCI competition III dataset IVa (Dornhege et al. 2004).The second data was recorded from a 32 channel EEG system with a sampling rate of 250Hz. The performance of the proposed algorithm was compared with other algorithms such as Sequential Forward and Backward Search (SFS and SBS) and SVM-GA. The results showed that RC-GA achieved high classification accuracy with lower computational cost. The proposed algorithm achieved an accuracy of 88.2% for dataset 1 and 89.38% for dataset 2, with the average selected channels numbering 50 and 25 respectively. It was shown that RC-GA attained a more compact and optimal channel subset in comparison to other mentioned algorithms.

Yang et al. (2013) presented a novel method for channel selection by considering mutual information and redundancy between the channels. The technique used laplacian derivatives (LAD) of power average across the frequency bands starting from 4 to 44Hz with the bandwidth of 4Hz and overlap of 2 Hz. After calculating LAD power features, maximum dependency with minimum redundancy (MD-MR) algorithm was applied.

The proposed technique was applied on a dataset collected from 11 healthy performing motor imagery of walking. LAD was extracted from 22 channels selected symmetrically. Among these 22 channels, 10 channels were selected using the proposed technique MD-MR. The results were compared with other algorithms such as filter bank common spatial pattern (FBCSP), filter bank with power features (FBPF), CSP and sliding window discriminant CSP (SWD-CSP). A 10×10 cross-validation was used to evaluate results. The results showed an increase in classification accuracy of 1.78% and 3.59% compared to FBCSP and SWD-CSP respectively. The results were recorded from 4, 10 and 16 selected LAD channels with an average accuracy of $67.19\% \pm 2.78$, $71.45\% \pm 2.50$ and $71.64\% \pm 2.67$ respectively which was slightly less than the accuracy of 22 LAD channels. Overall the performance was not degraded much when fewer electrodes were used.

Shenoy and Vinod (2014) presented a channel selection algorithm based on prior information of motor imagery tasks and optimizes relevant channels iteratively. The EEG signals were band-passed using a Butterworth filter of order 4 into 9 overlapping bands with bandwidth of 5Hz and an overlap of 2Hz. The overlapping bands were then transformed using Common spatial patterns filter to new subspace followed by feature selection using minimum redundancy and maximum relevance (mRMR). The channel selection algorithm was divided into two stages.

In the first stage, by utilizing prior information about MI tasks, a reference channel was chosen using regions of neuropsychological significance of interest for MI. A kernel radius was initialized in this step through which channels that lie in the kernel were assigned a weight inversely proportional to the Euclidean distance from the reference channel. The weights were updated iteratively in the second stage. Finding the most optimal channel subset was presented as an optimization problem which is solved iteratively. Weights were updated iteratively by incorporating Euclidean distance and the difference of ERD and ERS band power values into the equation. Modified periodogram with hamming window was used instead of Welch method to reduce computation cost. The results showed that the proposed algorithm produces an average classification accuracy of 90.77% and surpasses FBCSP, SCSP1 and SCSP2 with only 10 channels for dataset 1. For the second dataset, the accuracy was around 80% which was less than SCSP1.

Shan et al. (2015) developed an algorithm for optimal channel selection that involves real-time feedback and proposed an enhanced method IterRelCen constructed on relief algorithm.

The enhancements were made in two aspects: the change of target sample selection strategy and the implementation of iteration. A Surface Laplacian filter was applied to the raw EEG signals and features were extracted from a frequency range of 5–35 Hz. The frequency range was decomposed into 13 partially overlapped bands with proportional bandwidth. The Hilbert transform was applied to extract the envelope from each band. The IterRelCen is an extension of Relieff algorithm. The pseudocode for Relieff algorithm was given in Algorithm 4 (Robnik-Šikonja and Kononenko 2003). IterRelCen was different than relief algorithm in two aspects.

1. The target sample selection rule was adjusted rather than randomly selecting target samples, samples close to the center of the database from the same class had the priority of being selected first.
2. The idea of iterative computation was borrowed to eliminate the noisy features. N features with the smallest weights were removed from the current feature set after each iteration.

Algorithm 4: Pseudo code of Relieff algorithm

Data:
F // Features
C // Class set
k // Number of nearest samples
m // Number of iteration <= number of samples
Result: *W* // Features weight set
$W[F] = 0$;
for $i = 1$ *to* m **do**
    Randomly select a sample S;
    Find k nearest hits (NHs);
    for *each* $C_{class}(S)$ **do**
        Find k nearest misses (NMs) from class C;
    for $f = 1$ *to* F **do**
        Update features weights W;
    return W;

Three different datasets were used in this research study. Dataset 1 was from the data analysis competition (Sajda et al. 2003). EEG was recorded from 59 electrodes with a sampling rate 100Hz. For the second dataset, the experiment was conducted in the Biomedical Functional Imaging and Neuroengineering Laboratory at University of Minnesota (Yuan et al. 2008). 62 channels EEG was used to record data from eight subjects at a sampling rate of 200Hz. Dataset 3 was also from the sae Lab (Yuan et al. 2008). The only difference was the four class control signal controls the movement of a cursor in four directions i.e. left and right hand, both hands and nothing moves the cursor in left, right, up and down direction.

Multiclass SVM was applied to classify the signals. Tenfold cross validation was used to evaluate results. One way ANOVA was employed to test significant performance improvement. The classification results were 85.2%, 91.4% and 83.2% for dataset 1, 2 and 3 respectively with an average number of selected channels, 14 ± 5 for dataset 1, 22 ± 6 for dataset 2 and 29 ± 8 for dataset 3.

3.2.2 Other methods

Shan et al. (2012) explained that increasing the number of channels gradually improves classification accuracy. The concept was to gradually increase the number of channels and opt for a time frequency spatial synthesis model. After preprocessing, 13 overlapping sub-bands were extracted with a constant proportional bandwidth using a 3rd order Butterworth filter. The envelope of each band was extracted using the Hilbert Transform and treated as a feature because it contained power modulation information available in frequency band. The trial to trial mean was calculated and applied as the input to classification algorithm, which was performed on weighted frequency pattern.

They tested the technique on two different sets of data for comparison. It was observed from the results that increasing channels gradually raised the classification accuracy for the first dataset but not for the second one. For dataset 1, the channels are gradually increased from 2 to 62 and the classification accuracy rose from 68.7 to 90.4% for the training case and for testing data it was increased from 63.7 to 87.7%. Classification accuracy for dataset 2 was increased for the training data from 77.5 to 91.6% when channels were gradually increased from 2 to 59 but the results were bad for testing datasets. It increased from 2 till channel 16 and then it dropped to 68.9% from 81.3% when channels were increased from 16 to 59. In general, it was concluded that increasing the number of channels for on-line BCI will increase the classification accuracy instead of off-line BCI but the problem with the mentioned technique was all the envelopes correspond to different frequency bands and did not contribute towards accurate classification.

Das and Suresh (2015) proposed a Cohen's d effect size CSP (E-CSP) based channel selection algorithm. The method eliminates channels that do not carry any useful information. The algorithm removes noisy trials from channel followed by Cohen's d effect size calculation for channel selection. The noisy trials were eliminated by calculating z-score, which measured the distance of trial from the trial mean. More z-score means trial was far away from the mean trial and could be declared as noisy. The threshold for z-score was calculated using cross validation. Cohen's d effect size was used to select channels. The Cohen's d distance was calculated using the equation:

$$d_i = \frac{|\overline{C_{1i}} - \overline{C_{2i}}|}{\sigma} \tag{4}$$

where

$$\sigma = \frac{(\sigma_{1i}^j + \sigma_{2i}^j)}{2} \tag{5}$$

σ_{li}^j represent the standard deviation of l class across the selected jth trials of channels i. $\overline{C_{li}}$ represents the mean of l class of ith channels. The channel was selected if d value was greater than δ.

$L = \{i \ : \ d_i > \delta\}; \forall \ i$ represents the set of selected channels. δ was calculated using cross-validation and the recommended range for δ is $[0.01 - 0.1]$. CSP was implemented to extract features. Two data-sets were utilized to evaluate the algorithm. One was from BCI competition IV, dataset IIa (Dornhege et al. 2004) and the other was from BCI competition III, dataset IVa (Naeem et al. 2006). SVM was used as a classifier and the results were compared with CSP, SCSP1 and SCSP2 algorithms. The results showed that for dataset 1, the proposed algorithm gained an average increment of 3% over other algorithms with an average classification accuracy of 83.61% using approximately 8 channels. For dataset 2, the average classification accuracy was 85.85% with an average of 9.20 channels. The results

concluded that the algorithm increased the classification accuracy as well as decreased the number of selected channels compared to other mentioned algorithms.

Qiu et al. (2016) used a modified sequential floating forward selection (SFFS) to select channels for feature extraction. The technique utilized the neighbouring channels as a feature for selection. As the SFFS searched in a continuous loop for channel selection and required more time for the large feature set, so the author presented a solution by considering the location of the channels in the cerebral cortex. The adjacent channels were considered as features and the complete feature set had fewer features that made add or delete a channel easier for SFFS. The results showed that without affecting the classification accuracy, the algorithm could select channels in a very small amount of time. The time for channel selection was reduced by 57% for data 1 and 65% for data 2.

4 Discussion and guidelines

This paper discussed MI-EEG channel selection algorithms based on filtering techniques taking into account different factors mentioned in literature for channel evaluation and search strategy. This survey paper introduced the basics of selection algorithms along with the procedures and presents a detail description of filtering techniques used for the channels in MI-based EEG applications. The comprehensive exploration of channel selection algorithms has shown that with a little pre-computation, it is possible to achieve decent performance metrics while utilizing a small subset of EEG channels. The survey study showed that by implementing a channel selection algorithm, the number of channels can be reduced up to 80% without significant effect on classification tasks. Reducing channels will result in low computational complexity and a reduction in setup time. It also increases the maintainability of the device with respect to the subject.

Tables 1 and 2 present the summary of filtering techniques applied in selecting channels for MI-based EEG applications explored in Sect. 3. These techniques have been carried out on a number of different databases. In order to determine the effectiveness of an algorithm, extensive analysis will be needed to find a technique that gives best results for all MI-based EEG applications. Finally, it is observed that filtering technique for channel selection has been used in a variety of applications that used MI-based EEG signals. It can also be depicted from the summary in Tables 1 and 2 that for BCI Competition III, Dataset IVa (Dornhege et al. 2004), the best filtering technique to generate maximum performance utilized mRMR strategy for channel selection (Shenoy and Vinod 2014). It uses SVM as a classifier and generates an average classification accuracy of 90.77% with only 10 channels out of 118. For BCI Competition IV, Dataset IIa (Naeem et al. 2006), Cohen's d effect size (Das and Suresh 2015) has shown a promising result of 83.61% with average 8.11 channels out of 22. To identify an algorithm to be the most effective is strongly dependent on the application.

Many comparative studies about channel selection techniques can be found in the literature. However, sometimes these studies lead to contradictory results if applied to some other datasets. The next subsection contains the guidelines extracted from reviewing existing literature, which can help researchers in selecting a suitable channel selection algorithm.

4.1 Adequate evaluation criteria

To choose an optimal channel subset for MI application, the criteria to evaluate a channel subset is required. There are two main types of measures that can be used as evaluation criteria, known as the Information-based and the classifier-based measures/criteria.

Table 2 Summary of non-CSP based channel selection method for motor imagery EEG

Techniques	Channel selection strategy	Classifier	Performance metrics (average) (%)	No. of channels selected/Total no. of channels (average)	Dataset
Shan et al. (2012)	Time frequency spatial synthesized model	Weighted frequency patterns	63.7	2/64	University of Minnesota
			81.3	16/59	NIPS 2001
He et al. (2013)	Rayleigh coefficient based genetic algorithm (RC-GA)	Fisher linear discriminant (FLD)	88.2	50/118	BCI competition III dataset IVa
			89.38	25/32	Online experiment with left and right hand MI movements
Yang et al. (2013)	Max. dependency min. redundancy of LAD power features	SVM	71.64	16/22	Walking motor imagery EEG
Shenoy and Vinod (2014)	Min. redundancy max. relevancy (mRMR)	SVM	90.77	10/118	BCI competition III dataset IVa
			81.22	10/22	BCI competition IV dataset IIa
Shan et al. (2015)	IterRelCen	Multiclass SVM	85.2	14/59	NIPS 2001
			94.1	22/62	University of Minnesota
			83.2	29/62	University of Minnesota
Das and Suresh (2015)	Cohen's d effect size	SVM	85.85	9.20/118	BCI competition III dataset IVa
			83.61	8.11/22	BCI competition IV dataset IIa
Qiu et al. (2016)	Sequential floating forward search based on channel locations	SVM	78	18/59	BCI competition IV dataset I
			83.3	31/118	BCI competition III dataset IVa

Information-based measures Information-based measures rely on the generic properties of the data such as channel ranking by a metric, interclass distance and probabilistic dependence to evaluate the channel subset. The measures for ranking channels included correlation coefficient (Guyon and Elisseeff 2003), mutual information (Lazar et al. 2012), symmetrical uncertainty (Kannan and Ramaraj 2010) etc. Dissimilarity measures for binary variables such as matching coefficient (Cheetham and Hazel 1969) and measures such as Euclidean distance and angular separation (Haralick et al. 1973) for the numerical variable were also used to measure interclass distances. For probabilistic dependence, measures such as Chernoff (Lan et al. 2007), and Bhattacharyya (He et al. 2009) were used to measure probabilistic dissimilarities.

The advantage of information-based measure is that these measures are not computationally complex, require less time to compute, and usually require one-time calculation only. For applications where time and computational complexity is a limitation, these measures are the feasible option. The drawback is that it does not guarantee the best performance.

Classifier-based measures These measures used classification accuracy for evaluation of a channel subset. A classifier is responsible for finding the separability measure and a channel is selected when the classifier performs well. The error rate of the classifier is a widely used measure to evaluate the classification algorithm. Other measures used in the literature include Chi-squared (Greenwood and Nikulin 1996), information gain, odds ratio, and probability ratio (Şen et al. 2014). The problem with classifier-based evaluation is that the results are specific to the classifier used and changing the classifier effects the performance. The problem with classifier-based evaluation is that the results are specific to the classifier used and changing the classifier effects the performance. The other problems are the computation and time complexity. However, it guarantees the best results.

4.2 Channel selection approaches

There are several characteristics of every channel selection method. Filter methods used statistical properties of channels to filter out weak channels. These methods are classifier independent and rely on information measure. Using information measures for evaluation of channels helps filter method to be more generic and computationally more efficient. However, these methods usually have a poor performance than other methods. Wrapper method, on the other hand, is computationally demanding as the channel subsets were evaluated by a classifier, so the approach is classifier dependent, but it tends to give much better classification accuracy.

Hybrid channel selection techniques have both the properties of filter and wrapper methods and thus can be used to create more generic channel selection methods. There is a new approach that uses different channel selection methods instead of selecting one and accepting its results as the final channel subset. Then combine the results with an ensemble approach to get the best channel subset because there can be more than one optimal channel subsets (Haury et al. 2011; Abeel et al. 2009).

4.3 Search algorithms

There are three main types of search algorithms: complete search, sequential search, and heuristic/random search. Each has its own pros and cons. The complete search methods guarantee to find the best channel subset according to the evaluation criteria. The most com-

mon complete search algorithm is the branch and bound (Burrell et al. 2007). In sequential search, channels are added and removed sequentially, and these methods are usually not optimal but simple to implement and fast. The common sequential search methods are SFS (Pudil et al. 1994), SBS (Reunanen 2003), plus l minus r (Nakariyakul and Casasent 2009), and SFFS (Reunanen 2003). The random search method introduces randomness into the above-mentioned search algorithms and generates the next subset randomly. Simulated annealing (Yang 2014) and genetic algorithms (Davis 1991) are the best examples of the random search. These approaches are useful when data is too big and computational time is short.

4.4 Feature extraction

In channel selection for EEG applications, features extraction also plays an important role in maximizing the performance of the system. Those feature extraction techniques should be used to maximize the interclass variability. The most commonly used features for MI applications can be divided into four categories: time domain, parametric model based, transformed domain and frequency-based features. In time domain, features such as mean, variance, Hjorth parameter (Vidaurre et al. 2009) were used but these features were considered weak features. Auto-regression model (Pfurtscheller et al. 1998) and common spatial patterns (CSP) (Wang et al. 2006) were the examples of model-based features. Fourier and Wavelet transform (Xu and Song 2008; Baig et al. 2016) were also used in the literature to extract features and these features lie under transform features category. For frequency domain features, power spectral density and spectral edge frequency-based features were mostly used in the literature (Baig et al. 2014; Pfurtscheller and Da Silva 1999). After reviewing the literature for channel selection algorithms, the most effective features are the ones extracted by CSP and its variations.

4.5 Feature selection

Feature selection is also important in improving the system performance and for the selection of suitable channels. The features extracted from the EEG might be of high dimensionality, redundant or contains outliers that degrade the performance of the system. To deal with these factors, the feature selection algorithm is the option. It is also essential to apply feature selection algorithm to reduce the computational cost if the EEG data is too big (Al-Ani and Al-Sukker 2006). Extracting and selecting the best feature that maximizes the classification accuracy or class variance will eventually lead toward selecting optimal and relevant channels for EEG application.

5 Conclusion

The evaluation of channel selection algorithms is a challenging task. Various parameters, such as time, complexity and accuracy can be used to evaluate different channel selection algorithms. For real-time applications, time and accuracy are the most important parameters to consider. The performance of wrapper and hybrid selection techniques heavily rely on the selection of classifier and the subject. However, filter techniques are subject and classifier independent as these techniques use other signal information related parameters to select the best channel subset. The other main question in channel selection is how to determine the optimal number of channels to be selected. The answer to this question is very complicated

because the human brain is the most complex thing in the world. The generalization of the EEG decoding methods is very difficult, a slight change in the experiment will affect the signal processing, feature extraction, and classification methods. The case of channel selection is the same, in which the optimal channel set is highly dependent on the application, features, evaluation criteria and classifiers. Traditional approaches used the criteria based on the convergence of the classification accuracy using cross-validation or an analytical solution to an optimization problem etc. to select the optimal number of channels but the idea for the optimal number of channels are the ones that can preserve information of a task more than the others. In the end, it is a trade-off between the number of channels, system performance, time and computation cost. After reviewing the literature, we conclude that the brain cortex regions that are relevant to application often appear in the optimal channel subset. To design the optimal channel selection algorithm, we have to perform an extensive and deep analysis of each technique. We also need to study the performance sensitivity based on different types of tasks and classifiers. The application of evolutionary algorithms in selecting channels for EEG is under research and is open for further investigation.

References

Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2009) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3):392–398

Al-Ani, A, Al-Sukker A (2006) Effect of feature and channel selection on EEG classification. In: 2006 28th Annual international conference of the IEEE engineering in medicine and biology society, EMBS'06. IEEE, pp 2171–2174

Alotaiby T, El-Samie FEA, Alshebeili SA, Ahmad I (2015) A review of channel selection algorithms for EEG signal processing. EURASIP J Adv Signal Process 2015(1):66

Ang KK, Guan C, Chua KSG, Ang BT, Kuah CWK, Wang C, Phua KS, Chin ZY, Zhang H (2011) A large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain–computer interface. Clin EEG Neurosci 42(4):253–258

Antonio TGA, Alberto RGC, Luis VP (2012) Toward a silent speech interface based on unspoken speech. In: The 5th international joint conference on biomedical engineering systems and technologies

Arvaneh M, Guan C, Ang KK, Quek C (2011) Optimizing the channel selection and classification accuracy in EEG-based BCI. IEEE Trans Biomed Eng 58(6):1865–1873

Arvaneh M, Guan C, Ang KK, Quek C (2012) Robust eeg channel selection across sessions in brain–computer interface involving stroke patients. In: The 2012 international joint conference on neural networks (IJCNN). IEEE, pp 1–6

Atoufi B, Lucas C, Zakerolhosseini A (2009) A survey of multi-channel prediction of EEG signal in different EEG state: normal, pre-seizure, and seizure. In: Proceedings of the seventh international conference on computer science and information technologies, Yerevan, 28 Sept.–2 Oct. 2009

Baig MZ, Javed E, Ayaz Y, Afzal W, Gillani SO, Naveed M, Jamil M (2014) Classification of left/right hand movement from EEG signal by intelligent algorithms. In: 2014 IEEE symposium on computer applications and industrial electronics (ISCAIE). IEEE, pp 163–168

Baig MZ, Mehmood Y, Ayaz Y (2016) A BCI system classification technique using median filtering and wavelet transform. In: Kotzab H, Pannek J, Thoben KD (eds) Dynamics in logistics. Springer, Cham, pp 355–364

Baig MZ, Aslam N, Shum HP, Zhang L (2017) Differential evolution algorithm as a tool for optimal feature subset selection in motor imagery EEG. Expert Syst Appl 90:184–195

Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550

Blum A, Rivest RL (1989) Training a 3-node neural network is np-complete. In: Advances in neural information processing systems, pp 494–501

Burrell L, Smart O, Georgoulas GK, Marsh E, Vachtsevanos GJ (2007) Evaluation of feature selection techniques for analysis of functional MRI and EEG. In: DMIN, pp 256–262

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chapelle O, Keerthi SS (2008) Multi-class feature selection with support vector machines. In: Proceedings of the American statistical association

Cheetham AH, Hazel JE (1969) Binary (presence-absence) similarity coefficients. J Paleontol 43:1130–1136

Chrysostomou K (2009) Wrapper feature selection. In: Encyclopedia of data warehousing and mining, second edition. IGI Global, pp 2103–2108

Coyle SM, Ward TE, Markham CM (2007) Brain–computer interface using a simplified functional near-infrared spectroscopy system. J Neural Eng 4(3):219

Daly DD, Pedley TA (1990) Current practice of clinical electroencephalography. Raven Press, New York

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(3):131–156

Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151(1–2):155–176

Das A, Suresh S (2015) An effect-size based channel selection algorithm for mental task classification in brain computer interface. In: 2015 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 3140–3145

Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York

Dorigo M, Birattari M, Blum C, Clerc M, Stützle T, Winfield A (2008) Ant colony optimization and swarm intelligence: 6th international conference, ANTS 2008, Brussels, 22–24 Sept 2008, Proceedings, vol 5217. Springer

Dornhege G, Blankertz B, Curio G, Muller K-R (2004) Boosting bit rates in noninvasive eeg single-trial classifications by feature combination and multiclass paradigms. IEEE Trans Biomed Eng 51(6):993–1002

Duun-Henriksen J, Kjaer TW, Madsen RE, Remvig LS, Thomsen CE, Sorensen HBD (2012) Channel selection for automatic seizure detection. Clin Neurophysiol 123(1):84–92

Faller J, Scherer R, Friedrich EV, Costa U, Opisso E, Medina J, Müller-Putz GR (2014) Non-motor tasks improve adaptive brain-computer interface performance in users with severe motor impairment. Front Neurosci 8:320

Faul S, Marnane W (2012) Dynamic, location-based channel selection for power consumption reduction in EEG analysis. Comput Methods Program Biomed 108(3):1206–1215

Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5(Nov):1531–1555

Garrett D, Peterson DA, Anderson CW, Thaut MH (2003) Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Trans Neural Syst Rehabilit Eng 11(2):141–144

Gastaut H (1952) Electrocorticographic study of the reactivity of rolandic rhythm. Rev Neurol 87(2):176–182

Ghaemi A, Rashedi E, Pourrahimi AM, Kamandar M, Rahdari F (2017) Automatic channel selection in EEG signals for classification of left or right hand movement in brain computer interfaces using improved binary gravitation search algorithm. Biomed Signal Process Control 33:109–118

Greenwood PE, Nikulin MS (1996) A guide to chi-squared testing, vol 280. Wiley, New York

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182

Handiru VS, Prasad VA (2016) Optimized bi-objective EEG channel selection and cross-subject generalization with brain–computer interfaces. IEEE Trans Hum Mach Syst 46(6):777–786

Haralick RM, Shanmugam K, Dinstein I et al (1973) Textural features for image classification. IEEE Trans Systems Man Cybern 3(6):610–621

Haury A-C, Gestraud P, Vert J-P (2011) The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS One 6(12):e28210

He L, Yu Z, Gu Z, Li Y (2009) Bhattacharyya bound based channel selection for classification of motor imageries in EEG signals. In: Control and decision conference, 2009. CCDC'09. Chinese. IEEE, pp 2353–2356

He L, Hu Y, Li Y, Li D (2013) Channel selection by rayleigh coefficient maximization based genetic algorithm for classifying single-trial motor imagery EEG. Neurocomputing 121:423–433

Hill NJ, Lal TN, Schröder M, Hinterberger T, Widman G, Elger CE, Schölkopf B, Birbaumer N (2006a) Classifying event-related desynchronization in EEG, ECoG and MEG signals. In: Joint pattern recognition symposium. Springer, pp 404–413

Hill NJ, Lal TN, Schroder M, Hinterberger T, Wilhelm B, Nijboer F, Mochty U, Widman G, Elger C, Scholkopf B et al (2006b) Classifying EEG and ECoG signals without subject training for fast BCI implementation:

comparison of nonparalyzed and completely paralyzed subjects. IEEE Trans Neural Syst Rehabilit Eng 14(2):183–186

Kamrunnahar M, Dias N, Schiff S (2009) Optimization of electrode channels in brain computer interfaces. In: 2009 Annual international conference of the IEEE engineering in medicine and biology society. EMBC 2009. IEEE, pp 6477–6480

Kanayama N, Sato A, Ohira H (2007) Crossmodal effect with rubber hand illusion and gamma-band activity. Psychophysiology 44(3):392–402

Kannan SS, Ramaraj N (2010) A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm. Knowl Based Syst 23(6):580–585

Kee C-Y, Ponnambalam S, Loo C-K (2015) Multi-objective genetic algorithm as channel selection method for p300 and motor imagery data set. Neurocomputing 161:120–131

Kennedy J (2011) Particle swarm optimization. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning. Springer, New York, pp 760–766

Kirmizi-Alsan E, Bayraktaroglu Z, Gurvit H, Keskin YH, Emre M, Demiralp T (2006) Comparative analysis of event-related potentials during Go/NoGo and CPT: decomposition of electrophysiological markers of response inhibition and sustained attention. Brain Res 1104(1):114–128

Koles ZJ, Lazar MS, Zhou SZ (1990) Spatial patterns underlying population differences in the background EEG. Brain Topogr 2(4):275–284

Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: European conference on machine learning. Springer, pp 171–182

Kwak N, Choi C-H (2002) Input feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159

Lal TN, Schroder M, Hinterberger T, Weston J, Bogdan M, Birbaumer N, Scholkopf B (2004) Support vector channel selection in BCI. IEEE Trans Biomed Eng 51(6):1003–1010

Lal TN, Schröder M, Hill NJ, Preissl H, Hinterberger T, Mellinger J, Bogdan M, Rosenstiel W, Hofmann T, Birbaumer N et al (2005) A brain computer interface with online feedback based on magnetoen-cephalography. In: Proceedings of the 22nd international conference on Machine learning. ACM, pp 465–472

Lan T, Erdogmus D, Adami A, Mathan S, Pavel M (2007) Channel selection and feature projection for cognitive load estimation using ambulatory EEG. Comput Intell Neurosci 2007:74895. https://doi.org/10.1155/2007/74895

Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen V, Duque R, Bersini H, Nowe A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 9(4):1106–1119

Li M, Ma J, Jia S (2011) Optimal combination of channels selection based on common spatial pattern algorithm. In: 2011 International conference on mechatronics and automation (ICMA). IEEE, pp 295–300

Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502

Meng F, Tong K, Chan S, Wong W, Lui K, Tang K, Gao X, Gao S (2008) BCI-FES training system design and implementation for rehabilitation of stroke patients. In: 2008 IEEE international joint conference on neural networks, IJCNN 2008 (IEEE world congress on computational intelligence). IEEE, pp 4103–4106

Meng J, Liu G, Huang G, Zhu X (2009) Automated selecting subset of channels based on CSP in motor imagery brain–computer interface system. In: IEEE international conference on robotics and biomimetics (ROBIO), 2009. IEEE, pp 2290–2294

Millán JdR, Franzé M, Mouriño J, Cincotti F, Babiloni F (2002) Relevant EEG features for the classification of spontaneous motor-related tasks. Biol Cybern 86(2):89–95

Moran D (2010) Evolution of brain–computer interface: action potentials, local field potentials and electro-corticograms. Curr Opin Neurobiol 20(6):741–745

Naeem M, Brunner C, Leeb R, Graimann B, Pfurtscheller G (2006) Seperability of four-class motor imagery data using independent components analysis. J Neural Eng 3(3):208

Nakariyakul S, Casasent DP (2009) An improvement on floating search algorithms for feature subset selection. Pattern Recognit 42(9):1932–1940

Niedermeyer E (1997) Alpha rhythms as physiological and abnormal phenomena. Int J Psychophysiol 26(1–3):31–49

Ortiz-Rosario A, Adeli H (2013) Brain–computer interface technologies: from signal to action. Rev Neurosci 24(5):537–552

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Pfurtscheller G, Da Silva FL (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 110(11):1842–1857

Pfurtscheller G, Neuper C, Schlogl A, Lugger K (1998) Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. IEEE Trans Rehabilit Eng 6(3):316–325

Pineda JA (2005) The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing". Brain Res Rev 50(1):57–68

Pudil P, Novovičová J, Kittler J (1994) Floating search methods in feature selection. Pattern Recognit Lett 15(11):1119–1125

Qiu Z, Jin J, Lam H-K, Zhang Y, Wang X, Cichocki A (2016) Improved SFFS method for channel selection in motor imagery based BCI. Neurocomputing 207:519–527

Reunanen J (2003) Overfitting in making comparisons between variable selection methods. J Mach Learn Res 3(Mar):1371–1382

Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Mach Learn 53(1–2):23–69

Romero E, Sopena JM (2008) Performing feature selection with multilayer perceptrons. IEEE Trans Neural Netw 19(3):431–441

Sajda P, Gerson A, Muller K-R, Blankertz B, Parra L (2003) A data analysis competition to evaluate machine learning algorithms for use in brain–computer interfaces. IEEE Trans Neural Syst Rehabilit Eng 11(2):184–185

Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR (2004) BCI 2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans Biomed Eng 51(6):1034–1043

Şen B, Peker M, Çavuşoğlu A, Çelebi FV (2014) A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J Med Syst 38(3):18

Shan H, Yuan H, Zhu S, He B (2012) EEG-based motor imagery classification accuracy improves with gradually increased channel number. In: 2012 Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 1695–1698

Shan H, Xu H, Zhu S, He B (2015) A novel channel selection method for optimal classification in different motor imagery BCI paradigms. Biomed Eng Online 14(1):93

Shao S-Y, Shen K-Q, Ong CJ, Wilder-Smith EP, Li X-P (2009) Automatic EEG artifact removal: a weighted support vector machine approach with error correction. IEEE Trans Biomed Eng 56(2):336–344

Sharbrough F (1991) American electroencephalographic society guidelines for standard electrode position nomenclature. J Clin Neurophysiol 8:200–202

Shenoy HV, Vinod AP (2014) An iterative optimization technique for robust channel selection in motor imagery based brain computer interface. In: 2014 IEEE international conference on systems, man and cybernetics (SMC). IEEE, pp 1858–1863

Shih EI, Shoeb AH, Guttag JV (2009) Sensor selection for energy-efficient ambulatory medical monitoring. In: Proceedings of the 7th international conference on mobile systems, applications, and services. ACM, pp 347–358

Somol P, Pudil P, Novovičová J, Paclık P (1999) Adaptive floating search methods in feature selection. Pattern Recognit Lett 20(11–13):1157–1163

Stracuzzi DJ, Utgoff PE (2004) Randomized variable elimination. J Mach Learn Res 5(Oct):1331–1362

Tam W-K, Ke Z, Tong K-Y (2011) Performance of common spatial pattern under a smaller set of EEG electrodes in brain–computer interface on chronic stroke patients: a multi-session dataset study. In: 2011 Annual international conference of the IEEE engineering in medicine and biology society, EMBC. IEEE, pp 6344–6347

Torkkola K (2003) Feature extraction by non-parametric mutual information maximization. J Mach Learn Res 3(Mar):1415–1438

Van Gerven M, Farquhar J, Schaefer R, Vlek R, Geuze J, Nijholt A, Ramsey N, Haselager P, Vuurpijl L, Gielen S et al (2009) The brain–computer interface cycle. J Neural Eng 6(4):041001

Vidaurre C, Krämer N, Blankertz B, Schlögl A (2009) Time domain parameters as a feature for EEG-based brain–computer interfaces. Neural Netw 22(9):1313–1319

Wang Y, Gao S, Gao X (2006) Common spatial pattern method for channel selection in motor imagery based brain–computer interface. In: 27th Annual international conference of the engineering in medicine and biology society, 2005. IEEE-EMBS 2005. IEEE, pp 5392–5395

Weiskopf N, Mathiak K, Bock SW, Scharnowski F, Veit R, Grodd W, Goebel R, Birbaumer N (2004) Principles of a brain–computer interface (BCI) based on real-time functional magnetic resonance imaging (FMRI). IEEE Trans Biomed Eng 51(6):966–970

Wei Q, Wang Y (2011) Binary multi-objective particle swarm optimization for channel selection in motor imagery based brain-computer interfaces. In: 2011 4th International conference on biomedical engineering and informatics (BMEI)

Wolpaw JR, Birbaumer N, Heetderks WJ, McFarland DJ, Peckham PH, Schalk G, Donchin E, Quatrano LA, Robinson CJ, Vaughan TM (2000) Brain–computer interface technology: a review of the first international meeting. IEEE Trans Rehabilit Eng 8(2):164–173

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM (2002) Brain-computer interfaces for communication and control. Clin Neurophysiol 113(6):767–791

Xu B, Song A (2008) Pattern recognition of motor imagery EEG using wavelet transform. J Biomed Sci Eng 1(01):64

Yang X-S (2014) Nature-inspired optimization algorithms. Elsevier, Amsterdam

Yang H, Guan C, Wang CC, Ang KK (2013) Maximum dependency and minimum redundancy-based channel selection for motor imagery of walking EEG signal detection. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1187–1191

Yang H, Guan C, Ang KK, Phua KS, Wang C (2014) Selection of effective EEG channels in brain computer interfaces based on inconsistencies of classifiers. In: 2014 36th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 672–675

Yong X, Ward RK, Birch GE (2008) Sparse spatial filter optimization for EEG channel reduction in brain–computer interface. In: IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE, pp 417–420

Yuan H, Doud AJ, Gururajan A, He B (2008) Cortical imaging of event-related (de) synchronization during online control of brain–computer interface using minimum-norm estimates in frequency domain. IEEE Trans Neural Syst Rehabil Eng 16(5):425

Zhao Z, Liu H (2007) Searching for interacting features. In: IJCAI, vol 7, pp 1156–1161