

EEG data classification using wavelet features selected by Wilcoxon statistics

Thanh Nguyen · Abbas Khosravi ·
Douglas Creighton · Saeid Nahavandi

Received: 8 August 2014 / Accepted: 9 December 2014 / Published online: 24 December 2014
© The Natural Computing Applications Forum 2014

Abstract This paper introduces a method to classify EEG signals using features extracted by an integration of wavelet transform and the nonparametric Wilcoxon test. Orthogonal Haar wavelet coefficients are ranked based on the Wilcoxon test's statistics. The most prominent discriminant wavelets are assembled to form a feature set that serves as inputs to the naïve Bayes classifier. Two benchmark datasets, named Ia and Ib, downloaded from the brain–computer interface (BCI) competition II are employed for the experiments. Classification performance is evaluated using accuracy, mutual information, Gini coefficient and F-measure. Widely used classifiers, including feedforward neural network, support vector machine, k-nearest neighbours, ensemble learning Adaboost and adaptive neuro-fuzzy inference system, are also implemented for comparisons. The proposed combination of Haar wavelet features and naïve Bayes classifier considerably dominates the competitive classification approaches and outperforms the best performance on the Ia and Ib datasets reported in the BCI competition II. Application of naïve Bayes also provides a low computational cost approach that promotes the implementation of a potential real-time BCI system.

Keywords Wavelet transformation · Wilcoxon test · EEG signal classification · BCI competition II · Naïve Bayes classifier

1 Introduction

BCIs are tools which use brain activity in humans or animals to activate external devices without participation of peripheral nerves and muscles. They are designed to enable a direct connection between brain and external devices. BCIs can be applied to different fields and particularly useful in the treatment of neuromuscular disorders of paralysed patients because of their ability to bypass the motor system using only brain activity. For individuals for whom conventional methods are ineffective, BCIs are a useful communication-and-control option.

EEG is the recording of electrical activity of the cerebral cortex nerve cells in the brain. Examination of the EEG signals to understand brain activity is a fundamental problem of a BCI. Therefore, constructing a usable and reliable BCI requires an accurate and effective classification of multichannel EEG signals.

There have been a number of methods introduced for EEG signal classification in the literature from low-cost techniques such as linear discriminant analysis (LDA) [1–4], logistic regression [5–7], k-nearest neighbour [8–10], to computationally expensive approaches such as support vector machine (SVM) [11–13], artificial neural networks [14–17], and Adaboost ensemble learning [1, 18].

Subasi and Gursoy [19] proposed the use of discrete wavelet transform for feature extraction and SVM for classifying EEG signals. Gandhi et al. [20] on the other hand investigated wavelet function among existing members of the wavelet families for EEG data analysis.

An approach called clustering technique-based least-square SVM for the classification of EEG signals is presented in [21]. Sun et al. [22] applied Bayesian classifiers with Gaussian mixture models for online EEG data processing. Alternatively, Hsu [23] examined an adaptive

T. Nguyen (✉) · A. Khosravi · D. Creighton · S. Nahavandi
Centre for Intelligent Systems Research (CISR),
Deakin University, Waurn Ponds, VIC, Australia
e-mail: thanh.nguyen@deakin.edu.au

EEG analysis system for two-session, single-trial classification of motor imagery data. The adaptive LDA dynamically tunes its parameters by the Kalman filter when the left- and right-hand motor imagery data are classified.

Likewise, Hu et al. [24] suggested a combination of coefficients of joint regression model and spectral powers of EEG data at two specific frequencies to separate different motor imagery patterns. Cinar and Sahin [25] in another approach utilized particle swarm optimization and radial basis function networks for classifying EEG signals.

EEG signal classification in general requires investigation of a feature extraction. A survey of EEG signal analysis techniques was presented in [26]. Autoregressive (AR) models, Fourier transform (FT), time–frequency analysis and wavelet transform (WT) are broadly used to extract salient features [27]. AR, FT and conventional time–frequency methods commonly assume that EEG signal is stationary. However, this assumption is often invalid in practice. Therefore, WT is recommended rather than AR and FT for non-stationary transient signals like EEG. WT provides combined information in time–frequency domain that can enhance the performance of EEG classification. WT has been successfully applied to a number of problems including those of medical data analysis, e.g. see [28–31].

A method using Haar WT and Wilcoxon test is proposed for EEG signal feature extraction. Haar wavelets are chosen due to their compact support and orthogonality, which allows the discriminative features of data samples to be expressed with a few wavelet coefficients. The Wilcoxon test is employed as a filter approach by ranking the significant levels of all features. The orthogonal characteristic of Haar wavelets ensures the proper and effective employment of the Wilcoxon criterion.

This paper introduces a combination between WT and naïve Bayes classifier for EEG signal classification in a BCI system. Naïve Bayes assumes that the value of a particular feature is independent to the presence or absence of any other feature. This assumption is often invalid in practice. However, the employment of Haar wavelet features with orthogonal characteristic allows naïve Bayes to be carried out efficiently. Therefore, the use of naïve Bayes in combination with Haar wavelets for BCI motor imagery data classification is advocated.

Through this study, we examine and compare performance of the wavelet-naïve Bayes model with classification methods frequently applied in the literature. Experiments are conducted using two benchmark datasets, named Ia and Ib, downloaded from the brain–computer interface (BCI) competition II to make sure conclusions driven out of this study are valid and general. The competition winners of the Ia and Ib datasets, respectively, were Mensh et al. [32] and Bostanov [33]. We also show the dominance of the

proposed approach against these winner methods, which are briefly presented in the following.

Mensh et al. [32] used gamma-band power as a potential control signal for BCIs due to its correlation with high-level mental states. Although most frequency-based BCIs are based on the mu and beta rhythms, the authors recognized that most of the useful frequency information for classification in the Ia dataset is in the gamma band, with essentially none below 24 Hz. The discriminant analysis was utilized efficiently in this dataset despite its linearity limitation.

Bostanov [33] in another approach used the continuous WT and Student's two-sample *t*-statistic for EEG data feature extraction. The method performs fully automated detection and quantification of event-related brain potential components in the timescale plane. The classical linear discriminant analysis is then employed for the classification.

The arguments of the paper are organized as follows. The next section presents the main methodology where wavelet coefficients are selected by the Wilcoxon test and naïve Bayes classifiers are used for classification. Section 3 is devoted for experimental results and discussions, followed by concluding remarks in Sect. 4.

2 Naïve Bayes and wavelets selected by Wilcoxon test for EEG signal classification

Naïve Bayes classifier [34] requires a small number of training data to estimate the parameters. This is an advantage of the naïve Bayes classifier when applied for motor imagery data classification as there are often a limited number of trials in these applications.

The proposed methodology is diagrammed in Fig. 1. WT is applied to data of each channel separately to extract the information contained in the signals. The wavelet coefficients are then ranked based on the Wilcoxon test statistics to select the most discriminative coefficients of each channel. In the next step, these selected coefficients are combined to produce a feature set that serves as inputs to the naïve Bayes classifier. The following presents in detail WT and the Wilcoxon test for wavelet coefficient selection.

2.1 WT for feature extraction

WT represents a signal in a time–frequency fashion [35]. WT eliminates the requirement of signal stationarity that often applies to conventional methods. Once the wavelets (the mother wavelet) $\varphi(x)$ is fixed, translations and dilations of the mother wavelet can be formed $\{\varphi(\frac{x-b}{a}), (a, b) \in \mathbb{R}^+ \times \mathbb{R}\}$. It is convenient to take special values for a and b as $a = 2^{-j}$ and $b = 2^{-j}k$ where j and k

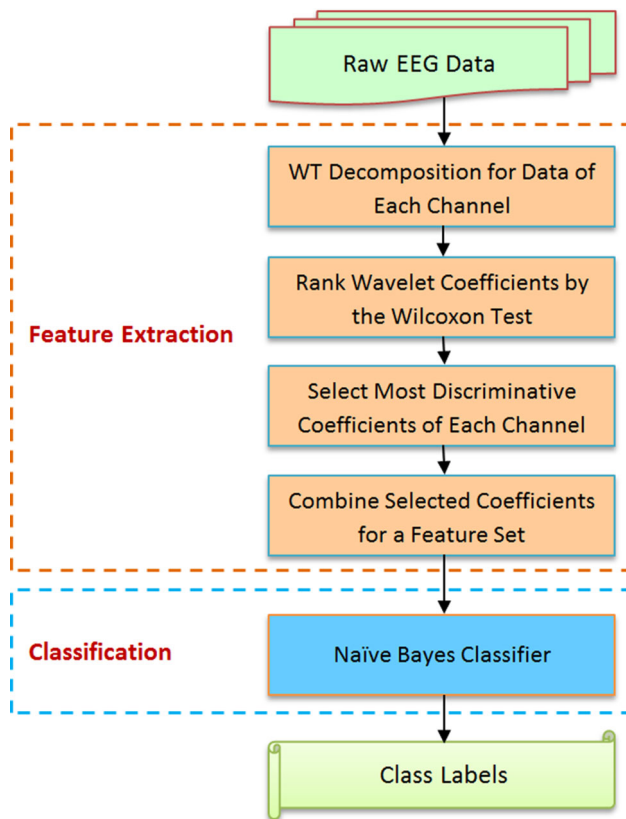


Fig. 1 Wavelets selected by Wilcoxon test as inputs to Naïve Bayes classifier

are integers. One of the simplest wavelets is the Haar wavelet, which has been used to solve various practical problems. Haar functions can uniformly approximate any continuous function. Dilations and translations of the function φ , which is $\varphi_{jk}(x) = const.\varphi(2^jx - k)$, define an orthogonal basis in $L^2(R)$. This means that any element in $L^2(R)$ may be represented as a linear combination of these basis functions. The scaling function in Haar wavelet is simply unity on the interval $[0,1)$ as $\phi(x) = 1(0 \leq x < 1)$.

Once the transformation is completed, a procedure to select coefficients that best separate the different classes is performed. A conventional approach to this procedure is the maximum variance (MV) criterion. MV selects coefficients (features) that have greatest variance. Quiroga et al. [36] argued that coefficients with the largest variance do not necessarily show the best discrimination among classes. Accordingly, selected coefficients should have the largest deviation from normality for the best discrimination. For this end, Quiroga et al. [36] suggested using the Lilliefors modification of a Kolmogorov–Smirnov (KS) test for normality. The test compares the cumulative distribution function of the data $F(x)$ with that of a Gaussian distribution $G(x)$ given a dataset x . Deviation from normality is measured by $\max(|F(x) - G(x)|)$.

Nevertheless, the KS test follows an unsupervised strategy that does not emphasize the difference or the discrimination of the classes. It is important to note that even in a single class, features may still present a large deviation from normality. If this context occurs, the KS test may nominate these features although they do not refer to the difference among the classes. Thus, the information used by KS test may not be appropriate to guarantee good discrimination properties of a feature passing the test. In this paper, we introduce a method using the Wilcoxon test to select elite wavelet coefficients for classification. Unlike the MV or KS test, the Wilcoxon method provides information about the equality of population locations of the classes. It involves a supervised approach that takes into account class labels to separate features of different classes. The following subsection scrutinizes backgrounds of the Wilcoxon method.

2.2 Wilcoxon method

Wilcoxon rank sum test is equivalent to the Mann–Whitney U test, which is a nonparametric test for equality of population locations (medians). The null hypothesis is that two populations enclose identical distribution functions, whereas the alternative hypothesis refers to the case two distributions differ regarding the medians. The normality assumption regarding the differences between the two samples is not required. That is why this test is used instead of the two-sample t test in many applications when the normality assumption is concerned.

The main steps of the Wilcoxon test are summarized below [37, 38]:

1. Assemble all observations of the two populations and rank them in the ascending order.
2. The Wilcoxon statistic is calculated by the sum of all the ranks associated with the observations from the smaller group.
3. The hypothesis decision is made based on the p value, which is found from the Wilcoxon rank sum distribution table.

In the application of the Wilcoxon test for wavelet coefficient selection, the absolute values of the standardized Wilcoxon statistics are utilized to rank coefficients. Note that the Haar wavelets are orthogonal. This ensures that the higher ranking coefficients are more prominent.

3 Experiments and discussions

3.1 Performance evaluation metrics

Accuracy, F1 score statistics (F-measure), Gini coefficient and mutual information are metrics used to evaluate

classification performance in the experiments. The F-measure is a single measure of a classification procedure's usefulness. The F-measure considers both the "Precision" and "Recall" of the procedure to compute the score. Precision is the number of correct positive results divided by the number of predicted positive results. On the other hand, Recall is the number of correct positive results divided by the number of actual positive results. The F-measure is the harmonic mean of Precision and Recall expressed as follows:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

The higher the F-measure, the better the predictive power of the classification technique. A score of 1 (or 100 %) means the classification procedure is perfect.

Gini coefficient (index) is an empirical measure of classification performance based on the area under a receiver operating characteristic curve (AUC). It is a linear rescaling of AUC: $2 \times \text{AUC} - 1$. The greater the Gini index, the better performance is the classifier.

The mutual information (MI) between estimated and true labels is calculated by:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

where $p(x, y)$ is the joint probability distribution function of estimated and true class labels X and Y , and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y , respectively.

3.2 Datasets

Experiments in this study are deployed using the two widely used Ia and Ib datasets downloaded from the BCI Competition II. The data were generated by Birbaumer et al. [39].

Description of the Ia dataset is available on the competition website as follows. "The dataset was taken from a healthy subject. The subject was asked to move a cursor up and down on a computer screen, while his cortical potentials were taken. During the recording, the subject received visual feedback of his slow cortical potentials (Cz-Mastoids). Cortical positivity leads to a downward movement of the cursor on the screen. Cortical negativity leads to an upward movement of the cursor. Each trial lasted 6 s. During every trial, the task was visually presented by a highlighted goal at either the top or bottom of the screen to indicate negativity or positivity from second 0.5 until the end of the trial. The visual feedback was presented from second 2 to second 5.5. Only this 3.5-s interval of every trial is provided for training and testing. The sampling rate of 256 Hz and the recording length of 3.5 s results in 896 samples per channel for every trial".

Number of training trials is 268 where 135 trials are of class "1" and 133 trials are of class "2", which correspond to moving a cursor up and down. Number of testing trials of this dataset is 293. "EEG data were taken from the following positions: Channel 1: A1-Cz (10/20 system) (A1 = left mastoid), Channel 2: A2-Cz, Channel 3: 2 cm frontal of C3, Channel 4: 2 cm parietal of C3, Channel 5: 2 cm frontal of C4, Channel 6: 2 cm parietal of C4".

On the other hand, "the Ib dataset was taken from an artificially respired ALS patient". The subject was asked to move a cursor up and down on a computer screen, while his cortical potentials were taken. During the recording, the subject received auditory and visual feedback of his slow cortical potentials (Cz-Mastoids). Cortical positivity leads to a downward movement of the cursor on the screen. Cortical negativity leads to an upward movement of the cursor. Each trial lasted 8 s.

During every trial, the task was visually and auditorily presented by a highlighted goal at the top (for negativity) or bottom (for positivity) of the screen from second 0.5 until second 7.5 of every trial. In addition, the task ("up" or "down") was vocalised at second 0.5. The visual feedback was presented from second 2 to second 6.5. Only this 4.5-s interval of every trial is provided for training and testing. The sampling rate of 256 Hz and the recording length of 4.5 s results in 1,152 samples per channel for every trial. EEG data were taken from the following positions: Channel 1: A1-Cz (10/20 system) (A1 = left mastoid), Channel 2: A2-Cz, Channel 3: 2 cm frontal of C3, Channel 4: 2 cm parietal of C3, Channel 5: vEOG artefact channel to detect vertical eye movements, Channel 6: 2 cm frontal of C4, Channel 7: 2 cm parietal of C4".

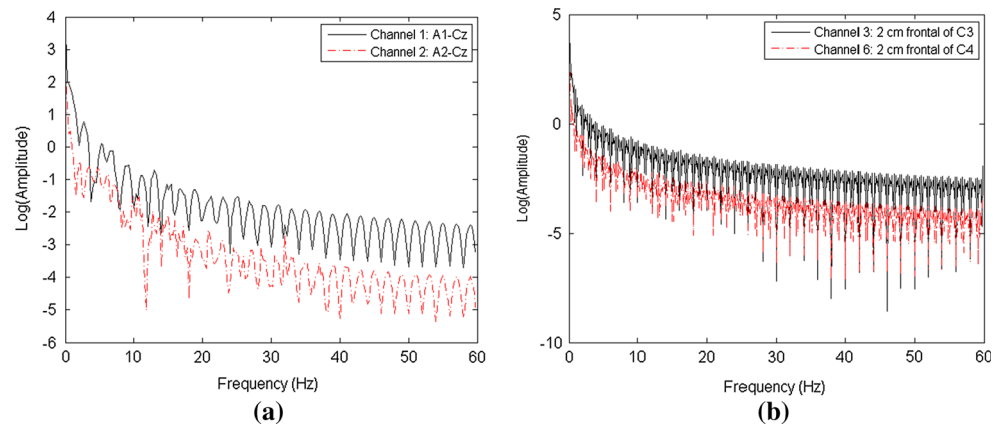
Number of training trials in this dataset is 200 where classes "1" and "2" both have the same number of trials at 100. Number of testing samples is 180.

EEG signals recorded in the Ia and Ib datasets exhibit noisy, embedded outliers, non-stationary and multidimensional characteristics. Data of individual channels show different spectra. Plots presented in Fig. 2 demonstrate more detailed the differences among the channels. For example, in the Ia dataset, signals in channel 1 (A1-Cz) are recorded in greater amplitudes than those in channel 2 (A2-Cz) at the same frequency. Similarly, in the Ib dataset, signal amplitudes in the channel 3 (2 cm frontal of C3) are larger than those in channel 6 (2 cm frontal of C4).

3.3 Feature extraction

The Haar WT at level 4 is implemented for each channel. Then, some filter approaches are applied to remove coefficients with low absolute values, little variation, small ranges or low entropy. For each criterion, the threshold for removing coefficients may be varied. For example, the 10th

Fig. 2 Amplitude comparisons between channels of the **a** Ia dataset and **b** Ib dataset



percentile is chosen as the threshold for every criterion. It means that coefficients are removed if they have absolute values in the lowest 10 % of the dataset. After that, coefficients with variances, ranges and entropy values less than the 10th percentile are also disregarded. These coefficients are generally not of interest because they have a low potential to discriminate the classes. Taking into account, these features may enhance noise and computational burden of the feature selection. As the aim is to select the most prominent feature of each channel to form the feature set, these filters often do not affect the selection because salient features are not removed by these procedures. They rather help to reduce the number of features before performing the Wilcoxon test and thus diminish computational costs of the test. These processes are applied to both training and testing sets simultaneously to ensure features that are removed or retained in both sets are analogous.

Figure 3a, b displays the distributions of wavelet coefficients obtained by WT on channel 1 of the Ia dataset and channel 6 of the Ib dataset, respectively. The original signal is a sum of the coarse approximation component A4 and four detail components D1–D4. Each component corresponds to a particular frequency bandwidth. The approximations are the low frequency of the signal, whereas the details are the high-frequency components. The blue triangular marks indicate the most discriminative coefficients selected through the statistical test based on the Wilcoxon method. Alternatively, the blue diamond marks specify the coefficients chosen by the KS test.

The Wilcoxon statistics of wavelet coefficients of channel 1 of the Ia dataset and channel 6 of the Ib dataset are exemplified in Fig. 4a, b, respectively. The most informative coefficient is indicated by ranking the Wilcoxon statistics of all coefficients. The coefficient with the greatest statistic of each channel is selected to form the feature set.

Distributions of the first feature (i.e. wavelet coefficient of channel 1) of the feature set for the Ia and Ib datasets are

illustrated in Fig. 5a, b, respectively. It can be seen that there is a disturbance and vague distinction of two classes in both datasets.

The feature set in the Ia dataset consists of six features corresponding to six channels. Likewise, there are seven features resulting from seven channels of the Ib dataset. Figure 6 displays the example 3D projections of the feature sets derived from the training samples of the Ia and Ib datasets, respectively. Obviously, there is a huge overlap between the two classes in both datasets. The feature set in the Ia dataset shows a clearer distinction between the two classes: “1” and “2”. Accordingly, we would expect to obtain a greater classification performance on the Ia dataset compared to that of the Ib dataset.

3.4 Results and discussions

For comparison, the following procedures: feedforward neural network (FFNN), support vector machine (SVM), k-nearest neighbours (kNN), ensemble learning Adaboost and adaptive neuro-fuzzy inference system (ANFIS) are also executed. Tables 1 and 2 present results of the proposed wavelet-naïve Bayes method and the comparable techniques deployed on the Ia and Ib datasets, respectively. With nondeterministic classifiers (i.e. FFNN, ANFIS), results are the average values over 30 independent trials. For these methods, standard deviation statistics are also reported adjacent to the means. Note that all results are displayed in percentage.

We also report here the best results of the competition on the two datasets, which can be seen at <http://www.bbc.de/competition/ii/results/index.html>. The winners of the Ia and Ib datasets, respectively, were Mensh et al. [32] and Bostanov [33].

Mensh et al. [32] obtained the greatest accuracy on the Ia dataset at 88.7 %. However, with the same method, the authors were just able to obtain 43.9 % accuracy on the Ib dataset. On the other hand, the method of Bostanov [33]

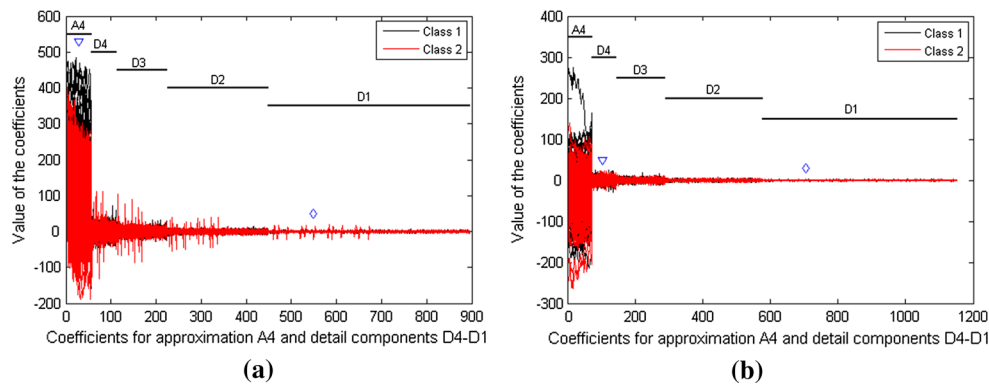


Fig. 3 Wavelet coefficients of **a** channel 1 of the Ia dataset and **b** channel 6 of the Ib dataset

Fig. 4 Wilcoxon statistics to select coefficients: **a** channel 1, Ia dataset, **b** channel 6, Ib dataset

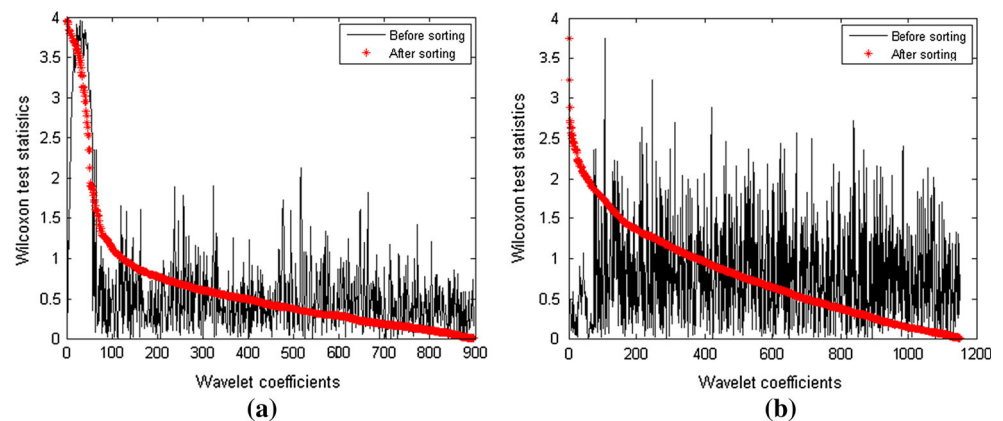
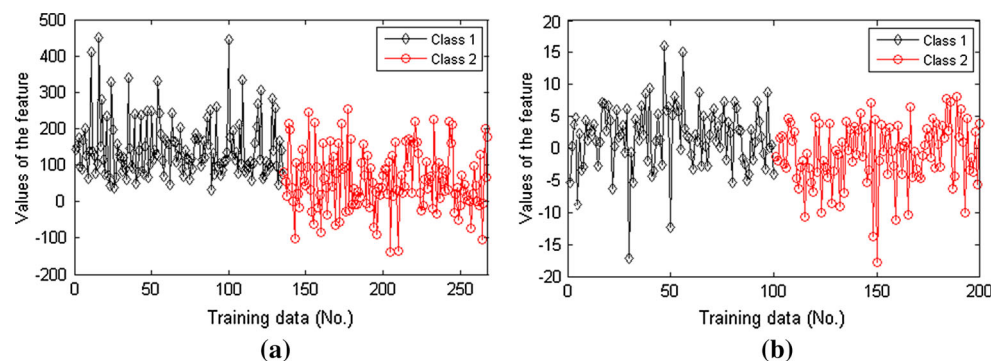


Fig. 5 Distribution of the first feature (Channel 1) of the **a** Ia dataset, **b** Ib dataset



derived the best performance on the Ib dataset with the accuracy at 54.4 %. This method, however, could just produce the accuracy at 82.6 % on the Ia dataset. These statistics reveal the fact that none of the two competition-winner methods performs effectively on both Ia and Ib datasets.

In contrast, the proposed wavelet-naïve Bayes clearly outperforms both competition-winner methods in both datasets. Wavelet-naïve Bayes obtains 90.10 and 56.11 % accuracy in the Ia and Ib dataset, respectively. The comparisons among the classifiers also highlight the superiority of the naïve Bayes against the competitive classifiers. The

dominance of wavelet-naïve Bayes is shown not only in accuracy but also in other performance measures, i.e. F-measure, Gini index and MI (Tables 1, 2).

Figure 7 presents the accuracy comparisons when performing classifiers using features selected by KS test and Wilcoxon test. Noticeably, the performance of applications of the Wilcoxon features is superior to those of the KS features through all classifiers. This is understandable as the KS test selects features without a reference to the class labels (i.e. an unsupervised approach). Inversely, the Wilcoxon method takes into account the class labels in

Fig. 6 Examples of 3D projection of the **a** Ia feature set, **b** Ib feature set

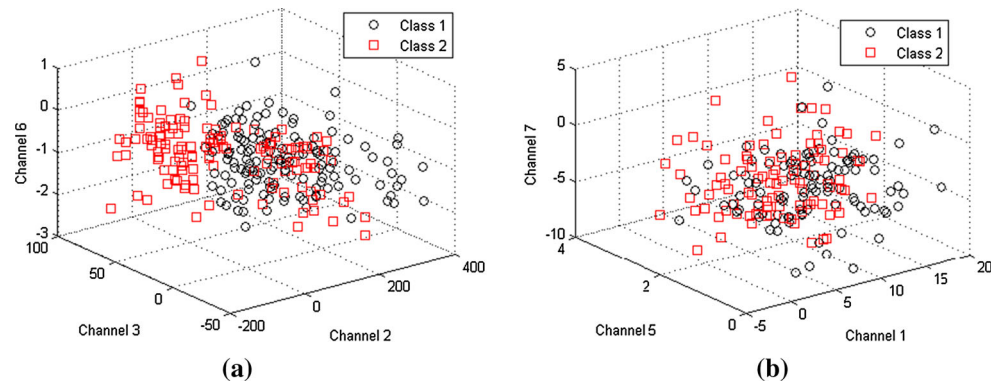


Table 1 Performance on the Ia dataset

	F-measure	Gini index	MI	Accuracy
The best result of competition by Mensh et al. [32]				88.70
The accuracy obtained by Bostanov [33]				82.60
SVM	85.71	68.67	40.42	84.30
kNN	81.57	67.83	41.34	83.96
Adaboost	83.97	65.92	35.19	82.94
FFNN	80.17	62.49	32.47	81.26
	(±5.65)	(±9.90)	(±9.81)	(±4.95)
ANFIS	86.91	73.88	44.35	86.94
	(±1.86)	(±3.97)	(±5.08)	(±1.99)
Naïve Bayes	89.90	80.19	53.57	90.10

Table 2 Performance on the Ib dataset

	F-measure	Gini index	MI	Accuracy
The best result of competition by Bostanov [33]				54.40
The accuracy obtained by Mensh et al. [32]				43.90
SVM	54.84	6.67	0.32	53.33
kNN	56.22	10.00	0.72	55.00
Adaboost	55.14	7.78	0.44	53.89
FFNN	55.42	8.07	0.70	54.04
	(±3.64)	(±5.60)	(±0.66)	(±2.80)
ANFIS	57.49	10.70	0.91	55.35
	(±1.71)	(±3.16)	(±0.48)	(±1.58)
Naïve Bayes	57.75	12.22	1.09	56.11

examining the features and also provides the information about the equality of population locations of the classes so that it is a more efficient feature selection.

Fivefold cross-validation is applied to further verify performance of the proposed method. The training and testing data samples in each dataset, which were split for the BCI competition II (see Sect. 3.2 for descriptions), are merged to yield a large dataset before applying the cross-validation strategy. Each newly formed dataset is divided

randomly into fivefolds where fourfolds of data are used for training and the last fold is for testing. This process is repeated 30 times for each classifier using either of the two feature sets selected by KS and Wilcoxon tests, and the average accuracy results are reported in Tables 3 and 4 for the Ia and Ib datasets, respectively.

To draw convincing conclusions in evaluating feature sets attained by KS and Wilcoxon tests, we implement the Kruskal–Wallis test [40] for comparing two sets of accuracy results. The Kruskal–Wallis test is a nonparametric version of the classical one-way ANOVA. As the results over 30 running times may not be normally distributed, they may violate the normal assumption of the ANOVA. Therefore, the use of Kruskal–Wallis test is appropriate. The test returns the *p* value for the null hypothesis that all samples in two sets of results are drawn from the same population. For example, the value in the last column of the SVM row in Table 3 (0.000) is the *p* value of the Kruskal–Wallis test performed on two sets (populations) of classification results: one set consists of 30 outcomes of SVM using the KS features and another set comprises 30 outcomes of SVM using the Wilcoxon features. In detail, this *p* value, which is smaller than 0.01, demonstrates that there is a significant difference (at 1 % level) between the two populations. It also means that the Wilcoxon feature set significantly dominates the KS feature set at the 1 % significance level.

It is seen from Tables 3 and 4 that all classifiers when employing the Wilcoxon feature set are considerably superior to when using the KS feature set. On average, the accuracy difference between these two feature sets is 16.33 and 8.55 % in the Ia and Ib datasets, respectively. This difference is maximally up to 20.53 % in the Ia dataset (the case of SVM) or 11.75 % in the Ib dataset (the case of naïve Bayes). Noticeably, the proposed combination of naïve Bayes classifier with Haar wavelets selected by Wilcoxon test outperforms all other competitive methods. It achieves the average accuracy at 82.59 %, which is the maximum value in the Ia dataset (see Table 3). In the Ib dataset (Table 4), it also attains the maximum accuracy at 62.89 %.

Fig. 7 Accuracy obtained by features selected by KS test and Wilcoxon test in the **a** Ia dataset, **b** Ib dataset

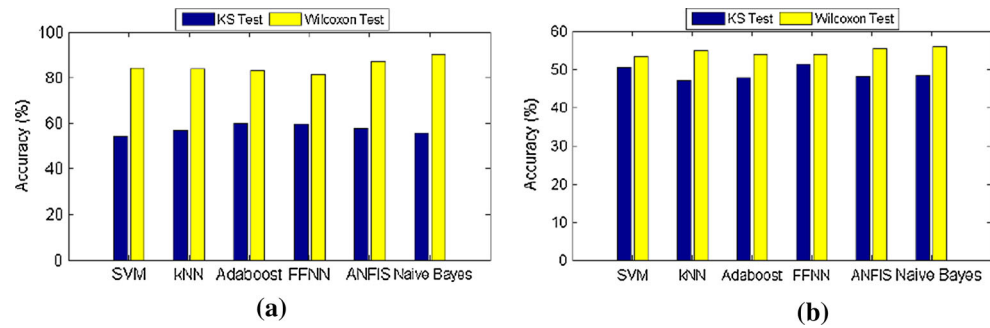


Table 3 Average accuracy of 30 running times with fivefold cross-validation in the Ia dataset

Classifiers	KS	Wilcoxon	Kruskal–Wallis <i>p</i> values
SVM	58.13 ± 4.26	78.65 ± 3.82	0.0000
kNN	66.72 ± 4.14	82.54 ± 2.88	0.0000
Adaboost	67.43 ± 4.11	80.64 ± 3.42	0.0000
FFNN	64.13 ± 5.94	80.62 ± 2.79	0.0000
ANFIS	67.30 ± 4.29	81.75 ± 3.45	0.0000
Naïve Bayes	65.10 ± 4.58	82.59 ± 3.75	0.0000

Table 4 Average accuracy of 30 running times with fivefold cross-validation in the Ib dataset

Classifiers	KS	Wilcoxon	Kruskal–Wallis <i>p</i> values
SVM	51.01 ± 4.82	62.32 ± 4.04	0.0000
kNN	50.48 ± 5.60	58.55 ± 5.79	0.0000
Adaboost	50.75 ± 4.86	57.15 ± 5.46	0.0000
FFNN	50.61 ± 5.94	57.41 ± 5.85	0.0001
ANFIS	50.70 ± 5.90	57.63 ± 5.54	0.0001
Naïve Bayes	51.14 ± 4.61	62.89 ± 4.89	0.0000

Results also show that the *p* values of the Kruskal–Wallis pairwise tests are all smaller than 0.01 (see last column of Tables 3, 4). Therefore, the Kruskal–Wallis tests reject the null hypothesis that results of two feature selection methods come from the same distribution at the 1 % significance level. This shows the statistically significant dominance and robustness of Wilcoxon method against the KS test method introduced in [36].

The processing time of classification methods is reported in Table 5 and graphically illustrated in Fig. 8a, b for the Ia and Ib datasets, respectively. The experiments are carried out on a computer that has the Intel(R) Core(TM) i7-2600K CPU @ 3.40 and 3.70 GHz with RAM at 16.0 GB running on the 64-bit Windows 7 Operating System. Note that the reported statistics measure the time

Table 5 Average processing time (in seconds) of classifiers

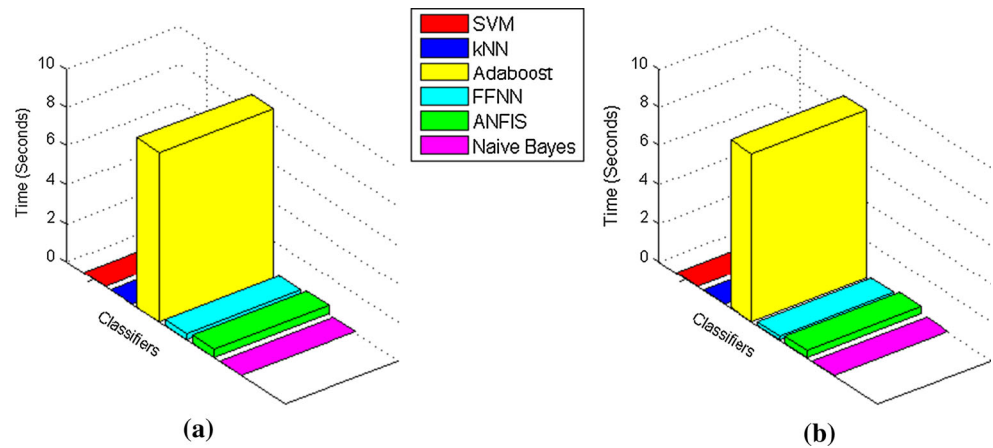
Classifiers	Ia dataset	Ib dataset
SVM	0.03	0.02
kNN	0.01	0.01
Adaboost	8.84	8.73
FFNN	0.31	0.23
ANFIS	0.46	0.4
Naïve Bayes	2.48×10^3	2.39×10^3

for modelling and classifying the entire testing datasets. Being an ensemble learning method, Adaboost spends the largest amount of time at approximately 9 s to deal with each dataset. FFNN, SVM, kNN and ANFIS accomplish classification of the data very quickly with less than a second. Naïve Bayes is the fastest method and thus it shows a processing time advantage against the other competitive classifiers. It needs less than 2.5 ms to process the entire Ia and Ib datasets with 293 and 180 testing trials, respectively. Obviously, the amount of time required by naïve Bayes to classify a single trial is remarkably small, which is approximately equivalent to 0.014 ms. This implies that the proposed method can be applied into a real-time EEG signal analysis system.

4 Conclusions

This paper introduces a method for EEG data classification using wavelets selected by Wilcoxon test and naïve Bayes classifier. Well-known methods such as MV and KS tests for selecting discriminative wavelet coefficients examine features in an unsupervised approach without reference to the class labels. This does not guarantee the separability of the feature set. The proposed method in this study suggests using the Wilcoxon test, which performs in a supervised strategy. The Wilcoxon test separates data samples according to the class labels and select features by evaluating the population locations of the classes. The Wilcoxon test is employed by ranking wavelet coefficients. As Haar

Fig. 8 Processing time of classifiers in the **a** Ia dataset, **b** Ib dataset



wavelets are orthogonal, this employment of the Wilcoxon criterion as a filter approach is advocated.

Experimental results on two benchmark datasets downloaded from the BCI competition II exhibit the superiority of the Wilcoxon feature selection against the KS test approach. Results of the Kruskal–Wallis test through the fivefold cross-validation procedure strengthen the robustness of the Wilcoxon feature set against the KS feature set. The proposed combination between Haar wavelets and Wilcoxon test therefore provides an effective approach to EEG signal feature extraction.

Results also show great performance dominance of naïve Bayes against other comparable classifiers, including FFNN, SVM, kNN, Adaboost and ANFIS. There is a synergy between Haar wavelets and naïve Bayes as their combination generates the greatest classification accuracy among competitive methods. This synergy is understandable because the most crucial assumption of naïve Bayes about the independence of features is absolutely satisfied by the orthogonal characteristic of Haar wavelets. Noticeably, naïve Bayes in combination with wavelets outperform the two winning methods of both benchmark Ia and Ib datasets in the BCI competition II by 1.40 and 1.71 %, respectively.

With regard to processing time, naïve Bayes classifier requires a low computational expense. Less than a millisecond is needed to classify a signal trial. This enables an online real-time BCI system to be implemented effectively. High accuracy and low computational cost clearly demonstrate the win–win benefit of the proposed algorithm. It therefore can be utilized in the development of a fully automated motor imagery EEG signal classification system, which is simple, accurate and reliable, for BCI applications. Clinically, the proposed approach can be applied as an authentic indicator to observe mental states of paralysed and disabled people and diagnose different brain related diseases in medical practice.

As two datasets used in this research are cortical positivity or negativity controlled BCIs, which are somehow particular because this controlling method is not popular in comparison with other kinds of BCI complementation, it would be worth to apply the proposed approach to various kinds of BCI complementation in a future research.

References

- Sabeti M, Katebi SD, Boostani R, Price GW (2011) A new approach for EEG signal classification of schizophrenic and control participants. *Expert Syst Appl* 38(3):2063–2071
- Vidaurre C, Kawanabe M, von Bunau P, Blankertz B, Muller KR (2011) Toward unsupervised adaptation of LDA for brain–computer interfaces. *IEEE Trans Biomed Eng* 58(3):587–597
- Li Y, Koike Y (2011) A real-time BCI with a small number of channels based on CSP. *Neural Comput Appl* 20(8):1187–1192
- Zhang R, Xu P, Guo L, Zhang Y, Li P, Yao D (2013) Z-score linear discriminant analysis for EEG based brain–computer interfaces. *PLoS ONE* 8(9):e74433
- Hosseinifard B, Moradi MH, Rostami R (2013) Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Comput Methods Programs Biomed* 109(3):339–345
- Prasad P, Halahalli H, John J, Majumdar K (2014) Single-trial EEG classification using logistic regression based on ensemble synchronization. *IEEE J Biomed Health Inform* 18(3):2014
- Li Y, Wen PP (2014) Modified CC-LR algorithm with three diverse feature sets for motor imagery tasks classification in EEG based brain computer interface. *Comput Methods Programs Biomed* 113(3):767–780
- Wang D, Miao D, Xie C (2011) Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection. *Expert Syst Appl* 38(11):14314–14320
- Guo L, Rivero D, Dorado J, Munteanu CR, Pazos A (2011) Automatic feature extraction using genetic programming: an application to epileptic EEG classification. *Expert Syst Appl* 38(8):10425–10436
- Acharya UR, Molinari F, Sree SV, Chattopadhyay S, Ng KH, Suri JS (2012) Automated diagnosis of epileptic EEG using entropies. *Biomed Signal Process Control* 7(4):401–408

11. 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 Rehabil Eng* 11(2):141–144
12. Vatankeh M, Asadpour V, Fazel-Rezai R (2013) Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage. *Appl Soft Comput* 13(5):2537–2546
13. Joshi V, Pachori RB, Vijesh A (2014) Classification of ictal and seizure-free EEG signals using fractional linear prediction. *Biomed Signal Process Control* 9:1–5
14. Subasi A, Erçelebi E (2005) Classification of EEG signals using neural network and logistic regression. *Comput Methods Programs Biomed* 78(2):87–99
15. Ting W, Guo-zheng Y, Bang-hua Y, Hong S (2008) EEG feature extraction based on wavelet packet decomposition for brain computer interface. *Measurement* 41(6):618–625
16. Guo L, Rivero D, Dorado J, Rabunal JR, Pazos A (2010) Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. *J Neurosci Methods* 191(1):101–109
17. Özbay Y, Ceylan R, Karlik B (2011) Integration of type-2 fuzzy clustering and wavelet transform in a neural network based ECG classifier. *Expert Syst Appl* 38(1):1004–1010
18. Ahangi A, Karamnejad M, Mohammadi N, Ebrahimpour R, Bagheri N (2013) Multiple classifier system for EEG signal classification with application to brain–computer interfaces. *Neural Comput Appl* 23(5):1319–1327
19. Subasi A, Gursoy MI (2010) EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Syst Appl* 37(12):8659–8666
20. Gandhi T, Panigrahi BK, Anand S (2011) A comparative study of wavelet families for EEG signal classification. *Neurocomputing* 74(17):3051–3057
21. Siuly, Li Y, Wen PP (2011) Clustering technique-based least square support vector machine for EEG signal classification. *Comput Methods Programs Biomed* 104(3):358–372
22. Sun S, Lu Y, Chen Y (2011) The stochastic approximation method for adaptive Bayesian classifiers: towards online brain–computer interfaces. *Neural Comput Appl* 20(1):31–40
23. Hsu WY (2011) EEG-based motor imagery classification using enhanced active segment selection and adaptive classifier. *Comput Biol Med* 41(8):633–639
24. Hu S, Tian Q, Cao Y, Zhang J, Kong W (2013) Motor imagery classification based on joint regression model and spectral power. *Neural Comput Appl* 23(7–8):1931–1936
25. Cinar E, Sahin F (2013) New classification techniques for electroencephalogram (EEG) signals and a real-time EEG control of a robot. *Neural Comput Appl* 22(1):29–39
26. Subha DP, Joseph PK, Acharya R, Lim CM (2010) EEG signal analysis: a survey. *J Med Syst* 34(2):195–212
27. McFarland DJ, Anderson CW, Muller K, Schlogl A, Krusienski DJ (2006) BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. *IEEE Trans Neural Syst Rehabil Eng* 14(2):135
28. Li D, Pedrycz W, Pizzi NJ (2005) Fuzzy wavelet packet based feature extraction method and its application to biomedical signal classification. *IEEE Trans Biomed Eng* 52(6):1132–1139
29. Tan Y, Li G, Duan H, Li C (2014) Enhancement of medical image details via wavelet homomorphic filtering transform. *J Intell Syst* 23(1):83–94
30. Torbati N, Ayatollahi A, Kermani A (2014) An efficient neural network based method for medical image segmentation. *Comput Biol Med* 44:76–87
31. Nguyen T, Khosravi A, Creighton D, Nahavandi S (2014) Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Syst Appl*. doi:10.1016/j.eswa.2014.10.027
32. Mensh BD, Werfel J, Seung HS (2004) BCI competition 2003-data set Ia: combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals. *IEEE Trans Biomed Eng* 51(6):1052–1056
33. Bostanov V (2004) BCI competition 2003-data sets Ib and IIb: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Trans Biomed Eng* 51(6):1057–1061
34. Bishop CM (2006) *Pattern recognition and machine learning*. Springer, New York
35. DeVore RA, Lucier BJ (1992) Wavelets. *Acta Numer* 1(1):1–56
36. Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput* 16(8):1661–1687
37. Deng L, Pei J, Ma J, Lee DL (2004) A rank sum test method for informative gene discovery. In: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pp 410–419
38. Lehmann EL, D’Abrera HJ (2006) *Nonparametrics: statistical methods based on ranks*. Springer, New York
39. Birbaumer N, Flor H, Ghanayim N, Hinterberger T, Iverson I, Taub E, Kotchoubey B, Kübler A, Perelmouter J (1999) A brain-controlled spelling device for the completely paralyzed. *Nature* 398:297–298
40. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 47(260):583–621