Proceedings of the 4th International
IEEE EMBS Conference on Neural Engineering
Antalya, Turkey, April 29 - May 2, 2009

Classification of EEG Signals Using Dempster Shafer Theory and a K-Nearest Neighbor Classifier

Ashkan Yazdani, Touradj Ebrahimi
Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
Email: {ashkan.yazdani, touradj.ebrahimi}@epfl.ch

Ulrich Hoffmann
Biorobotics Department
Fatronik - Tecnalia
Donostia - San Sebastián, Spain
Email: uhoffmann@fatronik.com

***Abstract*—A brain computer interface (BCI) is a communication system, which translates brain activity into commands for a computer or other devices. Nearly all BCIs contain as a core component a classification algorithm, which is employed to discriminate different brain activities using previously recorded examples of brain activity. In this paper, we study the classification accuracy achievable with a k-nearest neighbor (KNN) method based on Dempster-Shafer theory. To extract features from the electroencephalogram (EEG), autoregressive (AR) models and wavelet decomposition are used. To test the classification method an EEG dataset containing signals recorded during the performance of five different mental tasks is used. We show that the Dempster-Shafer KNN classifier achieves a higher correct classification rate than the classical voting KNN classifier and the distance-weighted KNN classifier.**

***Index Terms*—Dempster Shafer theory, BCI, nearest neighbor, classification, EEG**

I. Introduction

A BCI is a communication system, which allows a subject to act on his environment only by means of his brainwaves, without using the brain's normal output pathways of peripheral nerves and muscles. Like any communication system, a BCI has inputs (electrophysiological signals that result from brain activity) outputs (device actions), elements that transform inputs into outputs, and a protocol that determines its operation. The subject controls the active device by performing mental activities (MAs), which are associated with actions that are dependent on the BCI application. Typical BCI applications include control of cursor movement, control of a virtual apartment, spelling programs, control of external devices, and assistive technology for disabled users in general.

A typical BCI system employs a chain of several processing elements to translate brain activity into commands for an application. First, the brain activity of the subject is recorded with a signal acquisition device, for example with EEG electrodes and an EEG amplifier. Then, preprocessing algorithms are used to remove unwanted artifacts from the raw signals. Typical sources of artifacts in EEG signals are electrooculogram (EOG) activity, electromyogram activity (EMG), line noise, and slow baseline drifts. After preprocessing, features that are relevant for the classification of different MAs are extracted from the raw signals with a feature extraction method. Finally, a classification block uses the extracted features to decide and recognize which of the predefined MAs the subject performs. The output of the classification block can then be used to launch or control specific BCI applications.

One of the most important parts in BCI systems is the classification method. In [1], several classification methods have been studied for BCI applications and it has been shown that employing different classifiers may lead to considerably different system performances, depending on the structure and distribution of the data to be classified. Therefore, it is important to investigate different classification algorithms for BCI applications.

A particularly simple and popular classification algorithm that has so far not received much attention in the BCI community, is the k-nearest neighbor (KNN) method. The idea underlying the KNN method is to assign new unclassified examples to the class to which the majority of its k nearest neighbors belongs. One advantage of the KNN method over many other supervised learning methods is that it can easily deal with problems in which the number of classes is bigger than two. Furthermore, the KNN method allows adding examples to the training dataset without retraining the classifier. Clearly, the ability to deal with multiple classes as well as the ability to update classifiers online is important for BCI applications.

One potential problem inherent to the KNN approach is that it assumes that the k nearest neighbors of a test example are located at roughly the same distance from it. In other words, the KNN method does not take into account the fact that the k nearest neighbors of a test example might have largely differing distances from the test example. An intuitively appealing solution to this problem is to assign different degrees of importance to different nearest neighbors.

Dudani [2] proposed to assign different weights to the nearest neighbors as follows.

$$w^{(i)} = \begin{cases} \frac{d^{(k)}-d^{(i)}}{d^{(k)}-d^{(1)}} & \text{if } d^{(k)} \neq d^{(1)} \\ 1 & \text{if } d^{(k)} = d^{(1)} \end{cases} \quad (1)$$

Here the nearest neighbors are sorted by distance and $d^{(i)}$ denotes the distance of the i-th nearest neighbor from the test example. The decision rule in Dudanis approach is to assign the unknown example to the class which has the

978-1-4244-2073-5/09/$25.00 ©2009 IEEE

greatest sum of weights among the k nearest neighbors. An extension of the work by Dudani is the work of Denoeux [3] in which Dempster Shafer Theory is used to combine evidence coming from the k nearest neighbors of a test example. Besides the abovementioned points, the work of Denoeux also addresses ambiguity and distance rejection, and uncertainty and imprecision in class labels [3].

In this paper the performance of the Dempster Shafer theory based KNN classifier in a typical BCI application is studied. The rest of the paper is organized as follows. Section II briefly describes the Dempster Shafer theory of evidence and the KNN classifier based on this theory. In Section III, the data analysis methods including preprocessing and feature extraction techniques are described. In Section IV, the voting KNN classifier, distance-weighted KNN classifier, and Dempster Shafer KNN classifier are compared on an EEG dataset containing data recorded during the performance of five different mental tasks and the conclusion is given in Section V.

II. Materials and Methods

A. Dempster-Shafer Theory

The theory of evidence was introduced in the 1970s by G. Shafer [4] after the expansion of seminal works of A. Dempster [5]. This theory can be considered as a generalization of the probability theory [5], [6]. In the following a very brief introduction to the basic notions of the theory of evidence is given.

Considering a finite set (frame of discernment) Θ, a basic probability assignment (BPA) is a function $m : 2^{\Theta} \rightarrow [0, 1]$ so that $m(\emptyset) = 0$, $\sum_{A\subseteq\Theta} m(A) = 1$ and $m(A) \geq 0$ for all $A \subseteq \Theta$. The subsets of Θ which are associated with nonzero values of m are known as focal elements and the union of the focal elements is called core. The value of $m(A)$ expresses the proportion of all relevant and available evidence that supports the claim that a particular element of Θ belongs to the set A but to no particular subset of A. This value pertains only to the set A and makes no additional claims about any subsets of A. From this kind of mass assignment, the upper and lower bounds of a probability interval can be defined. Shafer defined the concepts of belief and plausibility as two measures over the subsets of Θ as follows.

$$\begin{aligned} \mathrm{Bel}(A) &= \sum_{B\subseteq A} m(B) \\ \mathrm{Pl}(A) &= \sum_{B\cap A\neq\emptyset} m(B) \end{aligned} \tag{2}$$

A BPA can also be viewed as determining a set of probability distributions P over Θ so that $\mathrm{Bel}(A) \leq P(A) \leq \mathrm{Pl}(A)$. It can be easily seen that these two measures are related to each other as $\mathrm{Pl}(A) = 1 - \mathrm{Bel}(\bar{A})$. Therefore, one needs to know only one of the three values of m, Bel, or Pl to derive the other two. Dempster's rule of combination can be used for pooling of evidence from two belief functions Bel_1 and Bel_2 over the same frame of discernment, but induced by different independent sources of information. The Dempster's rule of combination for combining two sets of masses, m_1 and m_2 is defined as follows.

$$\begin{aligned} m_{12}(\emptyset) &= 0 \\ m_{12}(A) &= \frac{1}{1-k} \sum_{B\cap C=A\neq\emptyset} m_1(B)m_2(C) \\ k &= \sum_{B\cap C=\emptyset} m_1(B)m_2(C) \end{aligned} \tag{3}$$

Here k is a measure of the amount of conflict between two evidences. If $k = 1$ the two evidences cannot be combined because their cores are disjoint. This rule is commutative, associative, but not idempotent or continuous.

B. KNN Classifier Based on Theory of Evidence

In [3], a new KNN classification method based on Dempster-Shafer theory of evidence has been proposed. In this section, we will briefly describe this method using the same notation as in [3]. Let us assume a feature matrix with the dimension of $P \times N$, where P denotes the dimension of feature vectors extracted from signal trials and N represents the number of feature vectors, a matrix L which indicates the true label of each feature vector, and a set of M different classes $C = \{C_1, \ldots, C_M\}$. The classification of an incoming sample x^s to be classified, using the KNN classifier based on theory of evidence will be described as follows. Denoting by Φ^s the set of k nearest neighbors of x^s, the label of each member of Φ^s (e.g. $\forall x^i \in \Phi^s, L_i = q$) can be considered as an evidence, which supports the hypothesis that x^s belongs to C_q. However, this piece of evidence is not 100% certain. Using the theory of evidence formalism, this can be expressed by assigning part of the belief to C_q and since this evidence does not support any other hypothesis, the rest of the belief will be assigned to the whole frame of discernment C. The BPA which can be employed here is a monotonically decreasing function of the distance between x^s and the elements of Φ^s . The rationale behind this is that the more distant elements of Φ^s provide the weaker belief that x^s belongs to their classes. Assuming that all $x^i \in \Phi^s$ are in class q, i.e. $L^i = q$, the BPA m^{si} is defined as follows.

$$\begin{aligned} m^{si}(\{C_q\}) &= \alpha \exp(-\gamma_q d^{\beta}) \\ m^{si}(C) &= 1 - m^{si}(\{C_q\}) \\ m^{si}(A) &= 0 \ \ \forall A \in 2^{\Theta}\backslash\{C, \{C_q\}\} \end{aligned} \tag{4}$$

Here $\gamma_q > 0$, $\beta \in \{1, 2, \ldots\}$, d is the Euclidean distance between x^s and x^i, and $0 < \alpha$ 1 implies that even if the distance between x^s and x^i is zero, it is not still certain that they belong to the same class. Simple heuristics for the choice of α and γ_q are presented in [3], whilst β is usually fixed to 1 or 2. This way, a BPA for each member of Φ^s can be defined, and combining these BPAs using Dempster's rule of combination will enable us make the final decision regarding the class assignment of x^s. Due to the fact that all of the belief functions have C as a focal element, it is always possible to

use this rule. If Φ_q^s represents the set of k nearest neighbors of x^s which belong to C_q, the combination of the corresponding BPAs can be done through $m_q^s = \bigoplus_{x^i \in \Phi_q^s} m^{si}$.

$$m_q^s(\{C_q\}) = 1 - \prod_{x^i \in \Phi_q^s} \left(1 - m^{si}(\{C_q\})\right) \tag{5}$$
$$m_q^s(C) = \prod_{x^i \in \Phi_q^s} m^{si}(C)$$

And if $\Phi_q^s = 0$, then $m_q^s(\{C\}) = 1$. Now that the BPAs m_q^s are at hand, a global BPA for all the M classes can be obtained through $m^s = \bigoplus_{q=1}^{M} m_q^s$.

$$m^s(\{C_q\}) = \frac{m_q^s(\{C_q\}) \prod_{r \neq q} m_r^s(C)}{K} \tag{6}$$
$$m^s(C) = \frac{\prod_{q=1}^{M} m_q^s(C)}{K}$$
$$K = \sum_{q=1}^{M} m_q^s(\{C_q\}) \prod_{r \neq q} m_r^s(C) + \prod_{q=1}^{M} m_q^s(C)$$

And the Bel and Pl functions can be defined as follows.

$$\mathrm{Bel}(\{C_q\}) = m^s(\{C_q\}) \tag{7}$$
$$\mathrm{Pl}(\{C_q\}) = m^s(\{C_q\}) + m^s(C)$$

The decision rule using such reasoning is to assign x^s to the class that has the greatest value of Bel (or Pl).

C. Datasets

The EEG data used in this study was recorded from 6 channels (C3, C4, P3, P4, O1, and O2) according to the 10-20 system of electrode placement. The EOG signal was recorded as well to monitor blinking artifacts. Trials were 10 seconds long and the data was sampled at 250 Hz with 12 bits of accuracy. The subjects were asked to perform five different mental tasks during different trials. These mental tasks are: baseline or total relaxation, multiplication i.e. silently multiplying two numbers (non-trivial), rotation i.e. imagination of rotation of an imagined 3-D block, counting i.e. visualization of numbers on an imaginary blackboard and incrementing, and finally letter composition i.e. mental composition without vocalizing. A complete description of this data can be found in [7].

III. Data Analysis

In this section, we present the methods used for cleaning, feature extraction, feature reduction and finally classifying windows of EEG signals coming from 10-second trials, into five different classes.

As preprocessing, first, the line noise was filtered out from the signals using a 60 Hz notch filter. In the next step independent component analysis (ICA) was used to remove the EOG artifacts from the signals using a method which was introduced in [8]. The FastICA algorithm was employed to compute EEG components which are independent of each other and more importantly are independent of the EOG component. At the end of this preprocessing block, a 6×2500 matrix for each trial was available where 6 indicates the number of artifact-free signals from different channels and 2500 equals the number of samples recorded during each trial. Figure 1 depicts a typical contaminated EEG signal and the artifact free signal after applying ICA.

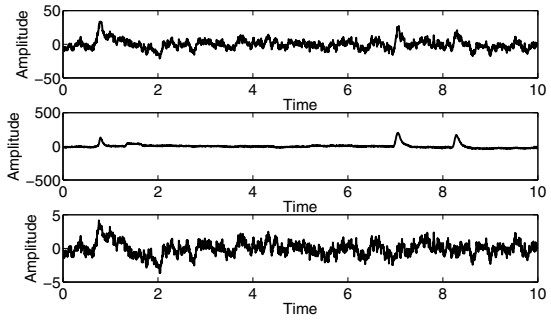

Fig. 1. From top to bottom: A typical EEG signal used in this study that is contaminated with EOG artifacts, the corresponding EOG signal, and the cleaned EEG signal after applying ICA.

The six artifact free signals for each task were divided into one-second windows, overlapping by half-seconds, producing 19 segments per trial. Each segment was represented as a matrix with dimensions 6×250. For feature extraction, AR coefficients were estimated for each of the one-second long segments using the Burg algorithm. Based on previous results in [9], [10], the order of six was used for AR modeling. Therefore, for each one-second long segment, 36 coefficients (6 AR coefficients for 6 different electrodes) formed the initial feature vector.

To extract more information from the signal, discrete wavelet transform (DWT) was employed. More specifically, the signal was decomposed into five levels using Daubechies wavelets. DWT decomposes the signal into a coarse approximation (A) and detail (D) information and allows analysis of the signal at different frequency bands. The Daubechies family of wavelets is known for its orthogonality property and efficient filter implementation and the Daubechies order 4 wavelet was found to be the most appropriate for analysis of EEG data [11]. Therefore, it has been used for time-frequency analysis of the EEG signals in this study. The extracted wavelet coefficients show the energy distribution of the EEG signal in time and frequency. As can be seen in table I, the A5 decomposition corresponds to the delta band of EEG signal, the D5 decompositions corresponds to the theta band of the EEG signal and the D4 and D3 decompositions correspond to the alpha and beta bands of EEG signals. Therefore, in feature extraction the A5, D5, D4, and D3 decompositions were used.

In order to reduce the number of features extracted with the DWT, only the following statistical features were considered [11].

1) Mean of the absolute values of the coefficients in each sub-band.
2) Average power of the wavelet coefficients in each sub-

band.

3) Standard deviation of the coefficients in each sub-band.

Features 1 and 2 provide information related to the power spectrum of the signal whereas feature 3 represents the changes in the power spectrum over time. In this manner, 72 features (12 statistical features for 6 different electrodes) were extracted from each EEG segment. Therefore, the total dimension of the extracted features from each EEG segment equals 108.

In the next step and after extracting features from all segments, the rows of the feature matrix were normalized so that each row had a mean value of zero and a standard deviation of one. This served to unify the dynamic range of all dimensions of the feature space. To evaluate classification performance, a 10-fold cross validation scheme was used for each subject. More precisely, the whole feature matrix was divided into ten partitions and one partition was used as test set while the other nine partitions where used as train (neighbors) set. This was repeated ten times, so that each partition was considered as a test set once.

Three different KNN classifiers, namely voting KNN (KNN), Distance weighted KNN (DWKNN) and KNN classifier based on Dempster-Shafer Theory of evidence (DSTKNN) were employed to assign the test points to one of the five classes based on the information in the neighbors set. In order to learn the optimal number of nearest neighbors k, we performed leave one out cross validation (LOOCV) within the train set, for values of k between one and fifty. The value of k which led to the maximum classification accuracy was selected and used for evaluation of performance on the test set. This learning method was performed separately for each of the three classifiers. For DSTKNN classifiers the value of α was set to 0.95, and the value of β was set to one, while the value of γ_q for each class was calculated from $\frac{1}{d_q^\beta}$, where d_q is the mean distance between two training vectors belonging to class C_q.

IV. Results

The classification performance achieved with the different classifiers for four subjects is shown in Table II. From this table it can be inferred that the DSTKNN classification algorithm results in higher correct classification rate than the DWKNN and KNN algorithms. The advantage of the classification method used in this study is that it has lower computational complexity in comparison with support vector machines (SVM) and neural networks and other classifiers, which are often used for classification. Furthermore, to the best of our knowledge, the results obtained in this paper using the aforementioned preprocessing, feature extraction, and classification methods, show some improvements in correct classification rate when compared to other studies on this dataset.

Decomposed Signal	Frequency Range (Hz)		
D1	62.5	-	125
D2	31.25	-	62.5
D3	15.625	-	31.25
D4	7.8125	-	15.625
D5	3.90625	-	7.8125
A5	0	-	3.90625

TABLE I
FREQUENCIES CORRESPONDING TO DIFFERENT LEVELS OF DECOMPOSITION FOR DAUBECHIES ORDER 4 WAVELET WITH A SAMPLING FREQUENCY OF 250 HZ

	KNN	DWKNN	DSTKNN
Subject 1	89.58±2.8	91.85±3.3	93.04±3.7
Subject 2	88.77±2.5	90.99±3.1	92.62±2.6
Subject 3	81.48±2.8	84.36±2.9	85.74±3.4
Subject 4	85.27±3.2	86.20±2.7	88.90±3.2

TABLE II
CORRECT CLASSIFICATION RATES OF THE THREE USED CLASSIFIERS FOR SUBJECTS 1,2,3, AND 4.

V. Conclusion

In this paper, Dempster-Shafer theory and KNN classification method were used for classification of five different mental tasks. To this end, after removing EOG artifacts, AR model coefficients and wavelet decomposition based statistical features were used to extract information from EEG signals. It has been shown that the DSTKNN classifier, which was used for the first time for classification of EEG signal and different mental tasks, will result in higher classification accuracy in comparison with other two KNN classifiers. Thanks to its simplicity and performance, we are now looking forward to evaluating the performance of this classifier with other BCI datasets and to employing it in online BCI applications.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control." *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun 2002.

[2] S. Dudani, "The distance weighted k-nearest neighbor rule," *IEEE Trans. Syst. Man Cybern.*, vol. 6, pp. 325–327, 1976.

[3] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst. Man Cybern.*, vol. 25, no. 5, pp. 804–813, 1995.

[4] G. Shafer, *A mathematical theory of evidence*. Princeton university press Princeton, NJ, 1976.

[5] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, vol. 38, pp. 325–339, 1967.

[6] F. Cuzzolin, "A geometric approach to the theory of evidence," *IEEE Trans. Syst. Man Cybern.*, vol. 38, no. 4, pp. 522–534, 2008.

[7] C. Anderson, S. Devulapalli, and E. Stolz, "Determining mental state from EEG signals using parallel implementations of neural networks," *Scientific programming*, vol. 4, no. 3, pp. 171–183, 1995.

[8] R. N. Vigário, "Extraction of ocular artefacts from EEG using independent component analysis." *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 3, pp. 395–404, Sep 1997.

[9] Z. A. Keirn and J. I. Aunon, "A new mode of communication between man and his surroundings." *IEEE Trans. Biomed. Eng.*, vol. 37, no. 12, pp. 1209–1214, 1990.

[10] C. Anderson, E. Stolz, and S. Shamsunder, "Discriminating mental tasks using EEG represented by AR models," in *IEEE 17th Annual Conference Engineering in Medicine and Biology Society, 1995.*, vol. 2, 1995.

[11] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform." *J Neurosci. Methods.*, vol. 123, no. 1, pp. 69–87, Feb 2003.