A Comparative study of ANN and SVM for EEG Classification

Mrs. KavitaMahajan,
Assistant Professor
Department of Electronics & communication Engineering
S.S.V.P.S B.S.Deore College of Engineering, Dhule

Mrs. Sangita M. Rajput
Assistant Professor
Department of Electronics & communication Engineering
S.S.V.P.S B.S.Deore College of Engineering, Dhule

Abstract

One of the research areas in biomedical signal processing is EEG signal processing. Epileptic seizures are disclosure of epilepsy. Brain disorders can be studied with the help of an electroencephalogram (EEG) signal for detection of epilepsy. In this proposed method, EEG signal is decomposed using DWT. Various dimension reduction methods are used for dimension reduction of decomposed data. The Classification is done with two classifiers for data as normal or abnormal. Performance of classifiers is compared to show the improved method.

1. Introduction

About 60 million people worldwide are affected by Epilepsy, the most common neurological disorders. Two third people get control on their seizureswith the help of proper medication and surgery. Remaining 25% people continue to get seizures even after medical treatment. Brain activity can be detailed with the help of Electroencephalogram (EEG). Forunderstanding epilepsy EEG recordings can provide valuable information. The seizuresdetection is possible by observing theEEGs help in the diagnosis and treatment ofepilepsy. Thus, automatic research is needed to understand the mechanisms causing epileptic disorders.

Analyses of brain activities started with the recording of EEG form human scalp after Hans Berger reporting activity in 1924 for first time.Formerly the inspection of EEG was done visually to qualitatively distinguish normal EEG activity from generalized abnormal activities. The advent of computers and the technologies associated with them has made it possible to effectively apply a host of methods to quantify EEG changes [2].

The EEG spectrum has four frequency bands: delta (<4 Hz), theta (4-8 Hz), alpha (8-13 Hz) and beta (13-30 Hz). Since the EEG signals are non-stationary, the parametric methods are not suitable for frequency decomposition of these signals. The wavelet transforms (WT) is a powerful method that was proposed in the late 1980s to perform time-scale analysis of signals. Since the WT is appropriate for analysis of non-stationary signals and this represents a major advantage over spectral analysis, it is well suited to locating transient events, which may occur during epileptic seizures. Adeli et al. [4]gave an overview of the discrete wavelet transform (DWT) developed for recognizing and quantifying spikes, sharp waves and spike-waves. They used wavelet transform to analyze and characterize epileptiform discharges in the form of 3-Hz spike and wave complex in patients with absence seizure. In the present study for epileptic seizure detection in patients with absence seizures (petit mal), the WT was used for feature extraction from the EEG signals belonging to the normal and the patient with absence seizure.

Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA) are wellknownmethods for feature extractionare used to reduce the dimension of data. Then these features were used as an input given to neural network and support vector machine. The accuracy ofthe various classifiers will be assessed and cross-compared, andadvantages and limitations of each technique will be discussed.

2. Feature Extraction Methods

2.1 The Wavelet Transform

A signal is said to be stationary if it does not change much over time. Fourier transform can be applied to the stationary signals. However, like EEG, plenty of signals may contain non-stationary or transitory characteristics. Thus it is not ideal to directly apply Fourier transform to such signals. In such a situation time–frequency methods such as wavelet transform must be used. In wavelet analysis, a variety of different probing functions may be used. This concept leads to the defining equation for the continuous wavelet transform (CWT):

$$W(a,b) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt$$

where*b* acts to translate the function across x(t), and the variable *a* acts to vary the time scale of the probing function, ψ. If *a* is greater than one, the wavelet function, ψ, is stretched along the time axis, and if it is less than one (but still positive) it contacts the function. While the probing function ψ could be any of a number of different functions, it always takes on an oscillatory form, hence the term ‘‘wavelet.” The normalizing factor ensures that the energy is the same for all values of *a*. In applications that require bilateral transformations, it would be preferred a transform that produces the minimum number of coefficients required to recover accurately the original signal [1]. The discrete wavelet transform (DWT) achieves this parsimony by restricting the variation in translation and scale, usually to powers of 2. For most signal and image processing applications, DWT-based analysis is best described in terms of filter banks. The use of a group of filters to divide up a signal into various spectral components is termed sub-band coding. This procedure is known as multi-resolution decomposition of a signal *x*[n]. Each stage of this scheme consists of two digital filters and two down-samplers by 2. The first filter, *h* [·] is the discrete mother wavelet, high-pass in nature, and the second, g[·] is its mirror version, low-pass in nature. The down-sampled outputs of first high-pass and low-pass filters provide the detail, D1 and the approximation, A1, respectively [4].

Selection of appropriate wavelet and the number of levels of decomposition is very important in analysis of signals using DWT. The number of levels of decomposition is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies required for classification of the signal are retained in the wavelet coefficients. Since the EEG signals do not have any useful frequency components above 30 Hz, the number of levels was chosen to be 5. Thus the signal is decomposed into the details D1–D5 and one final approximation, A5. The ranges of various frequency bands are shown in Table 1. The approximation and detail records are reconstructed from the Daubechies 4 (DB4) wavelet filter [5]. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the EEG signal in time and frequency. Table 1presents frequencies corresponding to different levels of decomposition for Daubechies order 4 wavelet with a sampling frequency of 173.6 Hz.

Table 1 Frequencies corresponding to different levels of decomposition for Daubechies 4 filter wavelet with a sampling frequency of 173.6 Hz.

Decomposed Signal	Frequency range (Hz)
D1	43.4–86.8
D2	21.7–43.4
D3	10.8–21.7
D4	5.4–10.8
D5	2.7–5.4
A5	0–2.7

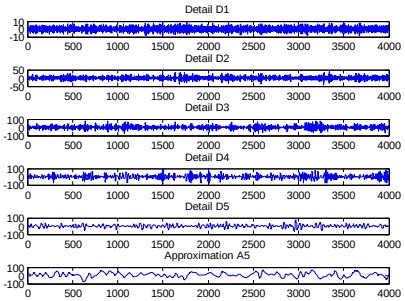

Fig. 1 Approximate and detailed coefficients of EEG signal taken from a healthy subject.

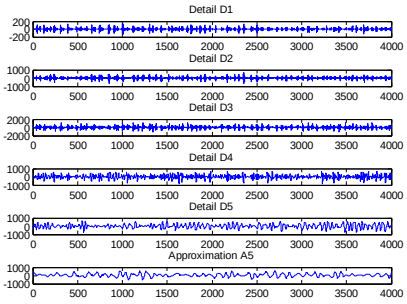

Fig. 2 Approximate and detailed coefficients of EEG signal taken from unhealthy subject (epileptic patient).

2.2 Independent component analysis

Assume that *n* linear mixtures $x_1,\ldots,x_n$ of n independent components were observed:

$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n, j = \overline{1,n}$(1)

In this equation the time has been ignored. Instead, it was assumed that each mixture $x_j$as well as each independent component s_i are random variables and $x_j(t)$ and $s_i(t)$ are samples of these random variables. It is also assumed that both the mixture variables and the independent components have zero mean [8].

If not subtracting the sample mean can always center the observable variables x_i. This procedure reduces the problem to the model zero-mean:

$\hat{x} = x - E(x)$ (2)

Let *x* be the random vectors whose elements are the mixtures and let *s*be the random vector with the components $s_1,\ldots,s_n$. Let A be the matrix containing the elements a_{ij}. The model can now be written:

x = As or $x = \sum_{i=1}^{n} a_i s_i$ (3)

The above equation is called independent component analysis or ICA. The problem is to determine both the matrix A and the independent components *s*, knowing only the measured variables *x*. The only assumption the methods take is that the components $s_i$are independent. It has also been proved that the components must have nongaussian distribution. Whitening can be performed via eigenvalue decomposition of the covariance matrix:

$VDV^T = E[\hat{x}\hat{x}^T]$ (4)

whereV is the matrix of orthogonal eigenvectors and D is a diagonal matrix with the corresponding eigenvalues. The whitening is done by multiplication with the transformation matrix P:

$\tilde{x} = P\hat{x}$ (5)

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

The matrix for extracting the independent components from$\tilde{x}$is $\widetilde{W}$, where $P = \widetilde{W}P$.

2.3 Fast Ica for N Units

A unit represents a processing element, for example an artificial neuron with its weights W. To estimate several independent components, the weights $w_1,\ldots.w_2$must be determined. The problem is that the outputs $w_1^T x, \ldots\ldots, w_n^T x$ must be done as independent as possible after each iteration in order to avoid the convergence to the same maxima. One method is to estimate the independent components one by one [9].

Algorithm:

i) Initialize w_i

ii) Newton phase

$w_i = E\{\tilde{x}g(w_i^T\tilde{x})\} - E\{\acute{g}(w_i^T\tilde{x})\}w_i$ (6)

where*g* is a function with one of the following form:

$$g_1(y) = \tanh(a_1 y),$$

$$g_2(y) = y\exp\left(-\frac{1}{2}y^2\right),$$

$g_3(y) = 4\,y^3$ (7)

iii) Normalization

$w_i = \frac{1}{\|w_i\|}w_i$(8)

iv) Decorrelation

$w_i = w_i - \sum_{j=1}^{i-1} w_i^T w_j w_j$(9)

v) Normalization (like in the step iii)

vi) Go to step ii) if not converged.

2.4 Principal component analysis (PCA)

Given a set of centered input vectors $\boldsymbol{x_t}$(t= 1,. . . ,land $\sum x_{t=0}$) **,** each of which is of *m* dimension$x_t = (x_t(1), x_t(2), \ldots, x_t(m))^T$(usually $m < l$)**,** PCAlinearly transforms each vector x_t**,** into a new ones$_t$**,** by

$s_t = U^T x_t$(1)

where***U*** is the *mxm*orthogonal matrix whose ithcolumn $u_i$is the ith eigenvector of the samplecovariance matrix $C = \frac{1}{l}\sum_{t=1}^{l} x_t x_t^T$.

In other words, PCA firstly solves the eigenvalue problem (2).

$\lambda_i u_i = Cu_i, i = 1, \ldots, m$ (2)

where$\lambda_i$is one of the eigenvalues of $\boldsymbol{C}$ **.** $u_i$is thecorresponding eigenvector. Based on the estimated u_i**,** thecomponents of $s_i$are then calculated **as** the orthogonaltransformations ofx_t**,**

$s_t = u_i^T x_t, i = 1, \ldots, m$(3)

The new components are called principal components. Byusing only the first several eigenvectors sorted indescending order of the eigenvalues, the number ofprincipal components in s_t can be reduced. This is thedimensional reduction characteristic of PCA [7].

2.5 LINEAR DISCRIMINANT ANALYSIS (LDA)

The aim of LDA is to create a new variable that is a combinationof the original predictors. This is accomplished by maximizing thedifferences between the predefined groups, with respect to thenew variable. The goal is to combine the predictor scores in sucha way that, a single new composite variable, the discriminantscore, is formed. This can be viewed as an excessive data dimensionreduction technique that compresses the p-dimensional predictorsinto a one-dimensional line. At the end of the process it

ishoped that each class will have a normal distribution of discriminantscores but with the largest possible difference in mean scoresfor the classes. In reality, the degree of overlap between the discriminantscore distributions can be used as a measure of the successof the technique. Discriminant scores are calculated by adiscriminant function which has the form:

$$D = w_1 Z_1 + w_2 Z_2 + w_3 Z_3 + \cdots w_p Z_p$$

As a result a discriminant score is a weighted linear combination ofthe predictors. The weights are estimated to maximize the differencesbetween class mean discriminant scores. Generally, thosepredictors which have large dissimilarities between class meanswill have larger weights, at the same time weights will be smallwhen class means are similar [9].

2.6. Support vector machines

Support vector machines (SVMs) are one of the most recently developed classifiers and build on developments in computational learning theory. They are finding applications in bioinformatic applications, because of their accuracy and their ability to deal with a large number of predictors. Most of the previous classifiers separate classes using hyperplanes that split the classes, using a flat plane, within the predictor space. SVMs extend the concept of hyperplane separation to data that cannot be separated linearly, by mapping the predictors onto a new, higher-dimensional space (called the feature space) in which they can be separated linearly. The method's name derives from the support vectors, which are lists of the predictor values obtained from cases that lie closest to the decision boundary separating the classes and are, therefore, potentially the most difficult to classify.It is reasonable to assume that these cases have the greatest impact on the location of the decision boundary [9].

Computationally, finding the best location for the decision plane is an optimization problem that makes uses of a kernel function constructs linear boundaries through non-linear transformations, or mappings, of the predictors. The 'clever' part of the algorithm is that it finds a hyperplane in the predictor space which is stated in terms of the input vectors and dot products in the feature space. A dot product is the cosine of the angle between two vectors (lists of predictor values) that have normalized lengths. The dot product can then be used to find the distances between the vectors in this higher dimensional space. A SVM locates the hyperplane that separates the support vectors without ever representing the space explicitly. Instead a kernel function is used that plays the role of the dot product in the feature space.

The support vector classifier has many advantages. A unique global optimum for its parameters can be found using standard optimization software. Nonlinear boundaries can be used without much extra computational effort. Moreover, its performance is very competitive with other methods. A drawback is that the problem complexity is not of the order of the dimension of the samples, but of the order of the number of samples [6].

3. Results

3.1 The Dataset

The publicly available data described in [2] is used for the experiment. There are five sets (denoted A–E)each containing 100 single-channels EEG segments of 23.6-sec duration. Sets A and B consisted of segments taken from surface EEG recordings that were carried out on five healthy volunteers in an awake state with eyes open are in set Aand eyes closed are in set B respectively. Segments in set D were recorded from within the epileptogenic zone and those in set C from the hippocampal formation of the opposite hemisphere of the brain while set E only contained seizure activity. All EEG signals were recorded with the same 128- channel amplifier system,using an average common reference. The datawere digitized at 173.61 samples per second using 12 bit resolution.Band-pass filter settings were 0.53–40 Hz (12 dB/oct). Four datasets (A,C, D and E) of the complete dataset are used for the experiment.

3.2 Experimental Result

In this experiment, the neural network and support vector machine classifiers are used to classify EEG signal as normal or epileptic. The EEG signal is first decomposed using wavelet decomposition. Then this signal dimensions are reduced by using ICA, PCAand LDA. The statistical features of this reduced signal are obtained which are used as an input to classification system based on SVM and Neural Network.

The two layered, five perceptron feed forward back propagation algorithm neural network classifier was used to train features extracted using PCA, ICA and LDA. For developing neural network classifier, feature vectors of normal data are used for training the classifier and for testing the classifier various data feature vectors are used. For SVM based classification samples are randomly selected and used for training theneural networks, and the remaining samples are used for testing the developed

models.Gaussian radial basis function(RBF) kernel is used for SVM.

Performance analysis of classifier is tested with parameters such as sensitivity (true positive ratio) and specificity (true negative ratio) calculated by using confusion matrix. The sensitivityvalue (true positive, same positive result as the diagnosis of expert neurologists) is calculated by dividing the total of diagnosisnumbers to total diagnosis numbers that are stated by the expertneurologists. Sensitivity, also called the true positive ratio, iscalculated by the formula:

$$sensitivity = TPR = \frac{TP}{TP + FN}\ X\ 100\%$$

On the other hand, specificity value (true negative, same diagnosisas the expert neurologists) is calculated by dividing the totalof diagnosis numbers to total diagnosis numbers that are stated bythe expert neurologists. Specificity, also called the true negativeratio, is calculated by the formula:

$$specificity = TNR = \frac{TN}{TN + FP}\ X\ 100\%$$

The procedure is repeated on EEG recordings of all different sets for combination of reduction methods and classifiers. The results obtained are shown in Table 2. As seenin Table 2, the classification rate with LDA feature extraction ishighest (100%) and ICA came second (99.50%). The PCA had lowestcorrect classification percentage (97.75%) compared to LDA andICA.

Table 2 The values of statistical parameters of the ICA, PCA and LDA models for EEG signal classification using Neural Network and Support Vector Machine.

DATA SET	PARAMETERS	A & C	A & D	A & E
PCA + ANN	ACCURACY (%)	**94.50**	**96.13**	**95.13**
	SENSITIVITY (%)	**75.00**	**87.07**	**75.26**
	SPECIFICITY (%)	**97.81**	**97.66**	**97.87**
ICA + ANN	ACCURACY (%)	**99.62**	**99.50**	**99.38**
	SENSITIVITY (%)	**99.35**	**99.13**	**99.13**
	SPECIFICITY (%)	**100**	**100**	**99.70**
LDA + ANN	ACCURACY (%)	**94.17**	**95.00**	**93.33**
	SENSITIVITY (%)	**96.00**	**98.00**	**96.00**
	SPECIFICITY (%)	**100**	**100**	**98.57**
PCA + SVM	ACCURACY (%)	**98.00**	**97.50**	**97.75**
	SENSITIVITY (%)	**96.16**	**96.16**	**96.15**
	SPECIFICITY (%)	**100**	**98.96**	**99.48**
ICA + SVM	ACCURACY (%)	**96.50**	**97.75**	**99.50**
	SENSITIVITY (%)	**100**	**96.04**	**99.50**
	SPECIFICITY (%)	**93**	**99.50**	**97.01**
LDA + SVM	ACCURACY (%)	**100**	**100**	**100**
	SENSITIVITY (%)	**100**	**100**	**100**
	SPECIFICITY (%)	**100**	**100**	**100**

4. Conclusion

Visual inspection of the signals does not provide much information regarding the health of individual. In this implemented system, following conclusions are drawn. The ANN classifies the EEG signal with overall accuracy of 97% correct rate whereas the SVM classifier classifies the EEG signal with overall accuracy of 98.67%.

- The SVM gives improved result for LDA as compared to ICA and PCA (100 %).

  Combination LDA+SVM produced more consistent results than combination of PCA+SVM and ICA+SVM. The excellence of LDA is also shown by the number of Support Vectors which is reduced and smaller than PCA and ICA.
- Different EEG signals (epileptic and non-epileptic) are applied to ANN and SVM and found that the SVM gives better result for all the different types of input EEG signals than ANN.

  Electroencephalogram obtained from the scalp of human body is basically combination of different random signals. ANN and SVM classifiers are used to classify EEG signals. When ANN classifier is used for classification, it is found that the artificial neural network using back-propagation training suffers from its slow convergence. They may have larger testing (statistic) errors as compared to support vector machines due to the Empirical Risk Minimization (ERM) approach employed by the former. The advantage of SVM over ANN is its better generalization ability due to Structural Risk Minimization (SRM) principle. The nonexistence of local minimum in SVM learning is also another reason why SVM is more superior.

  ANN is known to overfit data unless cross-validation is applied whereas SVM does not overfit data and thus ‘curse of dimensionality’ is avoided. In ANN learning, the topology is fixed but in SVM, learning actually is to learn the topology.

5. References

[1] Subasi A., M. Ismail Gursoy, "*EEG signal classification using PCA, ICA, LDA and support vector machines",* Expert Systems with Applications, 37,(2010), pp. 8659–8666.

[2] Bronzino, J. D., *"Principles of electroencephalography (2nd ed.)"*. In J. D. Bronzino (Ed.), The biomedical engineering handbook. Boca Raton: CRC Press LLC(2000).

[3] EEG time series are available under: http://www.meb.unibonn.de/ epileptologie/science/physik/eegdata.html.

[4] Adeli, H., Zhou, Z., &Dadmehr, N., "*Analysis of EEG records in an epileptic patient using wavelet transform".* Journal of Neuroscience Methods, 123, (2003), pp.69–87.

[5] RobiPolikar, online wavelet tutorial http://engineering.rowan.edu/~polikar/ WAVELETS /WTtutorial.html, (2006).

[6] Cortes, C. and Vapnik, V., *"Support vector networks"*. Machine Learning, 20,(1995), pp.273-297.

[7] Jolliffe I. T.,*"Principal component analysis"*. 2nd edition, Springer New York, (2002), pp.1-9.

[8] AapoHyvärinen and ErkkiOja*"Independent Component Analysis:Algorithms and Applications"*Neural Networks Research Centre, Helsinki University of Technology, Finland Neural Networks, 13(4-5): (2000),pp. 411-430.

[9] Hyvärinen A. and Oja E., "*A fast fixed-point algorithm for independent component analysis*, Neural Computing, **9**, (1997), pp.1483–1492.

[10] S. Balakrishnama, A. Ganapathiraju *"Linear Discriminant Analysis - A Brief Tutorial".* Mississippi State University.