32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

EEG feature selection using mutual information and support vector machine: A comparative analysis

Carlos Guerrero-Mosquera, Michel Verleysen and Angel Navia Vazquez

Abstract— **The large number of methods for EEG feature extraction demands a good choice for EEG features for every task. This paper compares three subsets of features obtained by tracks extraction method, wavelet transform and fractional Fourier transform. Particularly, we compare the performance of each subset in classification tasks using support vector machines and then we select possible combination of features by feature selection methods based on forward–backward procedure and mutual information as relevance criteria. Results confirm that fractional Fourier transform coefficients present very good performance and also the possibility of using some combination of this features to improve the performance of the classifier. To reinforce the relevance of the study, we carry out 1000 independent runs using a bootstrap approach, and evaluate the statistical significance of the F_{score} results using the Kruskal-Wallis test.**

I. INTRODUCTION

Neuroelectric waveforms such as EEG and event related potential (ERP) recordings from multiple electrode arrays vary in their frequency content over time and across recording sites on the scalp. Accordingly, EEG and ERP data sets are nonstationary in both time and space. From this perspective, it is necessary to try to identify hidden dynamical patterns which could yield important insight into the underlying physiological mechanisms.

The electroencephalogram (EEG) is the tool with more diagnostic applications in clinical environment, and can be a good indicator of abnormality in the central nervous system because it is the record of the electrical activity of the neurons in the brain. EEG feature extraction has a widely variety of methods or techniques that could be different depending of the application, i.e., EEG epilepsy detection, EEG prediction and brain computer interface (BCI) [8]. Automatic detection of EEG seizures has been investigated for years. However, so far, no detector has demonstrated to have competitive sensitivity and specificity values, new alternatives being necessary to obtain new information to distinguish between real epileptic seizures during non-epileptic events (high specificity). Taking this requirement into account, EEG features extraction plays an important role in conjunction with methods that evaluate this features in different scenarios such as detection or classification of EEG signals. One form of extracting information from EEG signals is to apply a transform and try to choose the relevant information to reduce the computational burden of the detection or classification systems, with the objective of improving the performance, being this the aim of feature selection methods. In this paper, we compare features obtained by tracks extraction and wavelet transform (WT) with features from fractional Fourier transform (FrFT). Tracks extractions has shown to be suitable to non-stationary signals [3]. WT can be applied to extract the wavelet coefficients of discrete signals and characterize the behavior of dynamic signals [10] and the FrFT arises as a new alternative in EEG feature extraction and has gained more attention for various applications including time-frequency design to specific applications [1]. In order to know if the features introduced have a good performance in EEG classification task, we use a support vector machine (SVM) classifier that is an effective and state-of-art classification method. Then, we chose the relevant features by features selection algorithms based on mutual information that could improve both performance classification task and reduce the computational cost by dimension reduction of the features. We use several datasets and compute measures such as F_{score} and bootstrap evaluation [3].

This paper is organized as follows. Section II introduces the feature extraction, classification and feature selection methods. Section III and Section IV shows the material and results of the seizure classification method applied to real EEG data from epileptic patients. Finally, in Section V, the main results are discussed and the principal conclusions with further work are presented.

II. METHODS

A. The Tracks extraction method

Tracks extraction method performs an estimation by peak-matching based on the localization of peaks in energy on the time-frequency plane. By linking peaks which occur at similar frequencies, we can define tracks along the time-frequency plane [9].

This method proposes to use three features based on length, frequency and energy of the principal track by a discretized version of the k-th segment in the time frequency plane, $\vartheta_k(n,m)$, such that the track extraction procedure identifies the coordinates of every track with a dummy variable that is equal to 1 in those points:

$$T_{k,\ell}(n,m) = \begin{cases} 1, & if \quad \vartheta_k(n,m) \quad \text{belogs to the } \ell\text{-th track} \\ 0, & \text{otherwise} \end{cases}$$

This work has been funded by the Spanish Government under grant TEC2008-02473.

C. Guerrero and A. Navia is with the Signal Theory and Communications department, University Carlos III of Madrid Avda. Universidad, 30 28911 Leganes. Spain cguerrero@ieee.org

M. Verleysen is with the Machine Learning Group, Universite Catholique de Louvain, pl. du Levant 3, 1348 Louvain-la-Neuve, Belgium michel.verleysen@uclouvain.be

The lenght of every track is computed as:

$$L_{k,\ell} = \sum_n \sum_m T_{k,\ell}(n,m) \tag{1}$$

the average frequency is

$$F_{k,\ell} = (\sum_n \sum_m T_{k,\ell}(n,m)m)/L_{k,\ell} \tag{2}$$

and the energy is

$$E_{k,\ell} = (\sum_n \sum_m T_{k,\ell}(n,m)\vartheta_k(n,m))/L_{k,\ell} \tag{3}$$

It is possible to identify the principal track in segment k as the largest track:

$$\ell' = arg\max_{\ell}\{L_{k,\ell}\} \tag{4}$$

such that the final features for segment k are:

$$L_k = L_{k,\ell'} \tag{5}$$

$$F_k = F_{k,\ell'} \tag{6}$$

$$E_k = E_{k,\ell'} \tag{7}$$

The interested reader may refer to [3] for more detail.

B. Wavelet transform (WT)

Wavelets arise to overcome the drawback of a fixed time-frequency resolution of short time Fourier transforms. This tool performs a multiresolution analysis, $W_\Psi f(a,b)$ of a signal, $x(n)$ by convolving the mother function $\Psi(n)$ with the signal as given in [5], [7]:

$$W_\Psi x(b,a) = \sum_{n'=0}^{N-1} x(n')\Psi^*\left(\frac{n'-b}{a}\right) \tag{8}$$

where $()^*$ denotes complex conjugate , a is the scale coefficient, b the shift coefficient and $a, b \in \Re$, $a \neq 0$.
In the procedure of multiresolution decomposition of a signal $x(n)$, each stage consists of two digital filters and two downsamplers by 2. The bandwidth of the filter outputs are half the bandwidth of the original signal, which allows for the downsampling of the output signals by two without loosing any information according to the Nyquist theorem. The downsampled signals provide detail D1 and approximation A1 of the signal [10].
Once the mother wavelet is fixed, it is possible to analyze the signal at every possible scale a and translation b. The Daubechies' family of wavelets is one of the most commonly used orthogonal wavelets to non-stationary EEG signals presenting good properties and allowing reconstruction of the original signal from the wavelet coefficients [7].

C. The fractional Fourier transform (FrFT)

The fractional Fourier transform (FrFT) is a new change representation of the signal which is an extension of the classical Fourier transform. When fractional order increases gradually, the FrFT of a signal can offer much more time-frequency united representation than the classical Fourier transform. Moreover, FrFT provides a higher concentration than STFT and avoids the cross terms components produced by quadratics TFDs.
Fourier techniques employ chirp harmonics for the decomposition of signals with time-varying periodicity. It can be interpreted as the representation of a signal in neutral domain by means of the rotation of the signal by the origin in counter-clockwise direction with rotational angle α in time-frequency domain shown in Fig.1.

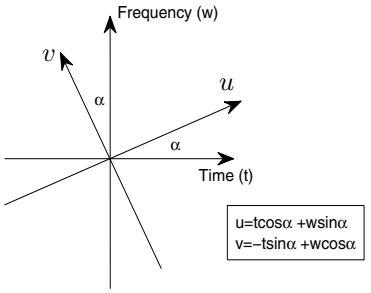

Fig. 1. The relation of fractional domain (u,v) with traditional time-frequency plane (t,w) rotated by an angle α.

Defining $K_\alpha(u,t)$ as the kernel function; the FrFT $X_\alpha(u)$ of $x(t)$ is given by

$$X_\alpha(u) = \int_{-\infty}^{\infty} x(t)K_\alpha(t,u)dt \tag{9}$$

$X_\alpha(u)$ could be expressed by means of the transformation kernel $K_\alpha(t,u)$ and is a linear transform [1], continuous in the angle α, which satisfies the basic conditions for being interpretable as a rotation in the time-frequency plane [1].

D. Support vector machine (SVM) classifier

The support vector machine (SVM) is a classification method rooted in statistical learning theory [11]. The motivation behind SVMs is to map the input into a high dimensional feature space, in which the data might be linearly separable. The construction of a hyperplane $\mathbf{w}^T x + b = 0$ ($\mathbf{w}$ the vector of hyperplane coefficients, b is a bias term) so that the margin between the hyperplane and the nearest point is maximized and can be posed as the quadratic-optimization problem transformed into a convex quadratic programming problem that is solved with standard techniques [12]. The

[1]

$$X_\alpha(u) = \begin{cases} \sqrt{(\frac{1-j\cot\alpha}{2\pi}} e^{j\frac{u^2}{2}\cot\alpha} \int_{-\infty}^{\infty} x(t)e^{j\frac{t^2}{2}\cot\alpha} e^{jut\csc\alpha}dt, \\ \text{if } \alpha \text{ is not a multiple of } \pi \\ x(t), \quad \text{if } \alpha \text{ is multiple of } 2\pi \\ x(-t), \quad \text{if } \alpha+\pi \text{ is multiple of } 2\pi \end{cases}$$

result is a discriminant function

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \cdot (\Phi(x) \cdot \Phi(x_i)) + b\right) = \text{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \cdot K(x, x_i) + b\right), \quad (10)$$

where $K(x_i, x_j)$ is a kernel function. This kernel trick is capable of producing arbitrary decision functions in input space, depending on the kernel function.
In many applications, SVMs have been shown to provide higher performance than traditional learning machines and have been introduced as powerful tools for solving classification problems including EEG classification. Two examples of suitable kernel functions are the polynomial kernel $K(x_i, x_j) = (x_i^T x_j + 1)^p$ and the radial basis function (RBF) kernel $K(x_i, x_j) = e^{(\frac{||x_i - x_j||^2}{2\sigma^2})}$, where σ^2 denotes the variance to be properly set [12].

E. Feature selection using mutual information

Usually feature selection relies on two main tasks: a relevance criterion and a searching procedure. The mutual information (MI) could be used as relevance criterion because it has been widely accepted as a good indicator to measure the importance of the feature. However, it is necessary to estimate the data probability distribution that usually is unknown. For this, we employ three estimations methods based on the Kraskov estimator, Parzen windows and K-nearest neighbors distances, all of them oriented to classification problems. The searching procedure used is the forward-backward procedure that acts in two steps: the first step, called forward stage, consists in adding features one by one. At each iteration, the feature chosen to be incorporated to the current subset is the one that most increases the mutual information. The process is stopped when adding any new feature actually decreases the mutual information. The second step, called backward stage, the features are eliminated one at a time. The feature that is excluded from the current feature subset is the feature that most increases the mutual information when it is discarded [2].

III. MATERIAL AND SETTINGS

This paper uses a database consisting of five sets (denoted as Z, O, N, F and S), each one containing 100 single-channel EEG segments each having 23.6 sec duration and sampling rate of 173.61 Hz [6]. In our experiments we use three classifications problems to evaluate our features.

1) The first problem called N1, two classes are examined: normal (Z) and seizure (S).
2) The second classification problem called N2, includes the classes normal, seizure-free and seizure (Z, F and S respectively).
3) In the third problem called N3, all the five classes are used.

More detail of the dataset and the problem classifications are described in [6].

Before feature extractions task, we have a feature matrix composed of: 3 features (LFE) from track extractions, 46 wavelets coefficients computed using the mother wavelet Daubechies 8 (db8) and detail D5; and 17 FrF coefficients obtained by varying the angle α from 0 to 4 (the increment step is 0.25).
All computation has been carried out off-line in a Pentium III computer, using the Matlab (V.6) programming environment and kernel RBF was used in the SVM classifier. The SVM's parameters was adjusted by cross validation.

IV. RESULTS

This section shows the performance analysis of the feature matrix analyzed separately in detection and classification tasks. We have worked with 7 experiments that involve computing all the possible feature combinations to evaluate the SVM performance.
The performance of the experiments set was based on a bootstrap experiment [4] using a function called "F_{score}" and defined as:

$$F_{score} = 2 * sensitivity * specifity / (sensitivity + specifity) \quad (11)$$

where sensitivity and specifity are defined as follows:

- $Sensitivity$: Percentage of EEG segments containing seizure activity correctly classified.
- $Specificity$: Percentage of EEG segments not containing seizure activity correctly classified.

Values in Table I correspond to F_{score} average values over the 1000 bootstrap runs. The statistical relevance of the results shown in Table I have been verified by means of a Kruskall-Wallis test, which is a sort of nonparametric ANOVA test that does not assume Gaussianity in the data under study. In all cases (except between Fractional Fourier (FrF) and LFE+Wavelets (W) in the N1 case) a p-value smaller than 0.01 has been obtained, thereby rejecting the null hypothesis that data come from the same distribution. Note in this table the difference in difficulty among N1 (easy), N2 and N3 (hard) problems.

TABLE I

F_{score} VALUES AVERAGE CORRESPOND TO OVER THE 1000 BOOTSTRAP RUNS.

	LFE	W	FrF	LFE+W	LFE+FrF	W+FrF	All
Dim	3	45	17	48	20	62	65
N1	99.36	**99.89**	98.70	98.70	99.23	99.74	99.66
N2	87.45	93.28	**99.18**	85.01	94.16	98.36	97.96
N3	86.27	82.54	83.59	81.35	88.18	**93.23**	92.25
Average	91.03	91.91	93.82	88.35	93.85	**97.11**	96.62

Fig.2, Fig.3 and Fig.4 show the F_{score} evolution vs number of coefficients. Finally, Table II shows the results in feature selection using all the features by forward-backward procedure with MI criteria. All the features have been normalized and each value in Table II represent the features chose. For example, {E,5 WC, 1 FrF} means feature E, 5 wavelets and 1 fractional coefficients.

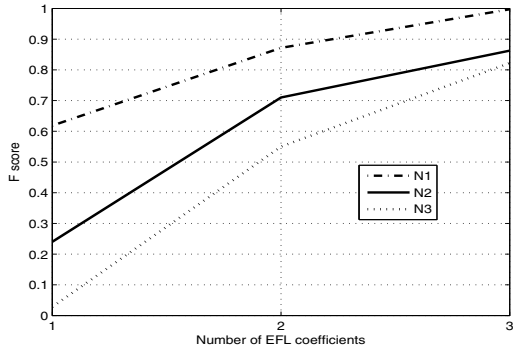

Fig. 2. F_{score} evolution vs number of LFE coefficients. Note how the feature F (correspond to number 2) increases the performance of the classifier in all the problems.

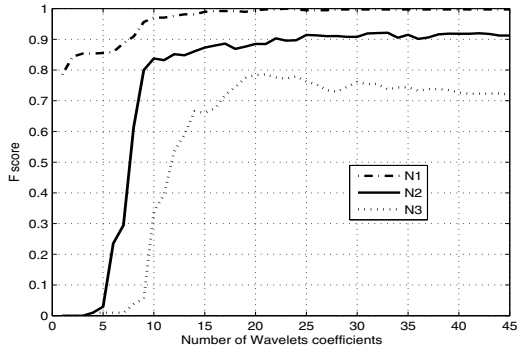

Fig. 3. F_{score} evolution vs number of wavelets coefficients. Note that all classification problems present a similar good performance using few coefficients.

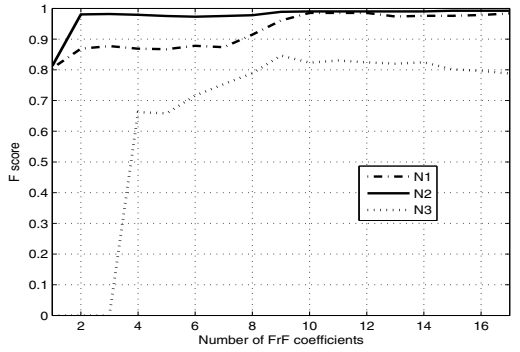

Fig. 4. F_{score} evolution vs number of FrF coefficients. The N2 problem presents a similar good performance using few coefficients as wavelets method (see Fig.3).

V. DISCUSSION AND CONCLUSIONS

The classification results clearly show the good performance of the FrF coefficients in all tasks (97.11% in Table I). However, we can assume that choosing a subset of features instead of all, turns out simpler the SVM classifier task increasing the classification F_{score} with less coefficients (93.23% in Table I).

Fig.2, Fig.3 and Fig.4 show a variation of the F_{score} measure related with the number of coefficients. These results could sense us other experiment consist in combining the feature subsets for improving the performance classification.

Table II shows the results after feature selection for each problem. Note in this table, how the best performance is reached with N1, reducing significantly the dimension of the features (one feature) and improving the performance of the classifier (100%). Results of N2 problem are quite similar compared to Table I, being the dimensional reduction and performance less notorious. For N3 problem, it is shown that exist a difficult of obtaining accurate estimators when we have unbalanced data.

In conclusion, FrF method really introduces new relevant information in our classification problems and opens the possibility of applying this method in other classification environments. Future works point to the study of data probability distribution estimators, feature selection parameters such as number of neighbors (K), effect of the size data and the sensibility with unbalanced data.

TABLE II

FEATURE SELECTION USING THREE DIFFERENT MI ESTIMATIONS. F_{score} VALUES AVERAGE CORRESPOND TO OVER THE 1000 BOOTSTRAP RUNS.

	Forward-backward selection		
	Kraskov	Parzen	Knn
N1	{E, 1 FrF} 99.63	{F} **100**	{ L,E,F,12 WC, 6 FrF} 99.75
N2	{L,F,E, 44 WC, 11 FrF} **99.37**	{L,E,F} 87.57	{L,E,F, 15 WC, 9 FrF} **99.2**
N3	{3 FrF} 85.35	{L,E,F} 86.34	{F,E,1 WC, 5 FrF} 85.59

REFERENCES

[1] Almeida L.B., "The fractional Fourier transform and Time-frequency representations," *IEEE Trans On signal Processing*, 1994, vol. 42, pp. 3084–3091

[2] Gomez V. Vanessa, Verleysen Michel and Jerome Fleury, "Information theoric feature selection for functional data classification," *Neurocomputing*, 2009, vol. 72, pp. 3580–3589

[3] Guerrero-Mosquera C., Malanda Trigueros A., Iriarte Franco J. and Navia Vazquez Angel, "New feature extraction approach for epileptic EEG signal detection using time-frequency distributions," *Med Biol Eng Comput*, 2009, vol. 48, pp. 321–330

[4] Harrell F.E., *Regression Modeling Strategies*, Springer, New York, 2001

[5] Latka M., Was Z., "Wavelet analysis of epileptic spikes," *Physical Rev. E*, 2003, vol. 67:052902

[6] Tzallas A.T., Tsipouras M.G. and Fotiadis D.I, "The use of Time-frequency distributions for epileptic seizure detection in EEG recordings," *Proceedings of the IEEE EMBS*, 2007, pp. 3–6

[7] Mallat S., *A wavelet tour of signal processing, Third edition: The sparse way*, Elsevier, Burlington USA, 2009

[8] McFarland D.J.,Anderson, K.R. Mller A. Schlgl and D.J. Krusienski, "BCI meeting 2005. Workshop on BCI signal processing: Feature extraction and traslation," *IEEE Trans On Neural Sys and Rehab Eng*, 2006, pp. 135–138

[9] McAulay R.J and Quatieri T.F, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustic, Speech, and Signal Processing*, 1986, vol. 34, pp. 744–754

[10] Ocak Hasan, "Optimal classification of epileptic seizures in EEG using wavelet analysis and genetic algorithm," *Signal processing*, 2008, vol. 88, pp. 1858–1867

[11] Vapnik, V, *The Nature of Statistical Learning Theory*, New York. Springer-Verlag, 1995

[12] Schlkopf B. and Smola A., *Learning with kernels*, Cambridge. The MIT Press, 2002