Two-Stage Feature Selection Algorithm Based on Supervised Classification Approach for Automated Epilepsy Diagnosis

Mechmeche S1*, Salah RB2 and Ellouze N1
1National Engineering School of Tunis, University of El Manar, UR SITI, Tunis, Tunisia
2Prince Sattam Bin Abdulaziz University, Biomedical Technology Department, Riyadh, Saudi Arabia

Abstract

Epileptic diagnosis is generally achieved by visual scanning of Interictal Epileptiform Discharges (IEDs) using EEG recordings. The main objective of this research is to select a smallest relevant feature subset from the original dataset in order to reduce the diagnosis time and increase classification accuracy by removing irrelevant and redundant features. For this purpose we suggest a two-stage feature selection algorithm based on supervised classification approach adopting successively a wrapper feature selection and a wrapper feature subset selection method. Matlab simulation results illustrate that through comparing the two classifiers, the high-dimensionality is reduced at only one relevant feature that showed classification metrics of 100%. The epilepsy diagnosis is successfully tested in the discriminant Fisher-space with the single-best relevant feature.

Keywords: Cross-validation; Classification metrics; EEG; Feature selection; IED's; LDA; Mahalanobis distance classifier; QDA; Supervised classification

Introduction

Epileptic is a neurological disorder marked by sudden recurrent episodes of sensory disturbance, loss of conscience, convulsions, associated with abnormal electrical activity in the brain. The confirmation of the existence of an epileptic diseases is based on visual detection of isolated Interictal Epileptiform Discharges (IEDs) (spikes or spike-waves complex), using EEG (Electroencephalogram) signal recordings in certain brain areas, for example, the confirmation of the epileptic-absence type is based on presence of a spike-waves rhythmic at 3 Hz [1-3]. This technique is inaccurate, fastidious and too time consuming. The aim of our research is to establish an automated diagnosis of epileptic disease employing a supervised classification approach (Figure 1).

To create a training set, we need to build a knowledge database composed of normal EEG sample and epileptic EEG sample. Feature extraction is an essential pre-processing step to pattern recognition and machine learning problems. To build the training set, the signal pattern may be described by three field analysis: Time field [4-11], frequency field [11-13], and time-frequency field [4,7,11,14-16]. In this article, EEG-signal pattern is described in high dimensionality in the three previous fields. To reduce the dimensionality at a SRFS (Smallest Relevant Feature Subset), we have proposed two-stage feature selection algorithm using wrapper-based method in supervised classification [17]: The first stage uses the IFE (individual feature evaluation) method and the second stage uses the SBS (sequential Backward Selection) method.

A Mahalanobis Distance-based Classifier (MDC) is suggested to classify the unknown EEG signal into "Normal" or "Epileptic" classes. For an optimal visualization of both of them, the samples are projected in the linear Fisher space [18,19] using Fisher linear Discriminant Analysis (FDA) that consists of seeking the optimal directions that are efficient for discrimination [20,21].

Methods

Knowledge database

The population selected is composed of 20 labeled single-EEG signals (derived from the Neurology department of University Hospital of Sousse-Tunisia), sampled at a frequency $F = 200$Hz, segmented at 1 second epoch, and filtered from artifacts, divided into two groups: 10 normal signals for the first group and 10 epileptic signals for the second group. These signals will be modeled by a set of features to form the training set that will be used in the feature selection process.

Feature extraction

In feature extraction process, we have adopted the statistical analysis approach from each single-EEG signal. Feature vector is composed of normal EEG sample and epileptic EEG sample. Feature extraction is an essential pre-processing step to pattern recognition and machine learning problems. To build the training set, the signal pattern may be described by three field analysis: Time field [4-11], frequency field [11-13], and time-frequency field [4,7,11,14-16]. In this article, EEG-signal pattern is described in high dimensionality in the three previous fields. To reduce the dimensionality at a SRFS (Smallest Relevant Feature Subset), we have proposed two-stage feature selection algorithm using wrapper-based method in supervised classification [17]: The first stage uses the IFE (individual feature evaluation) method and the second stage uses the SBS (sequential Backward Selection) method.

A Mahalanobis Distance-based Classifier (MDC) is suggested to classify the unknown EEG signal into “Normal” or “Epileptic” classes. For an optimal visualization of both of them, the samples are projected in the linear Fisher space [18,19] using Fisher linear Discriminant Analysis (FDA) that consists of seeking the optimal directions that are efficient for discrimination [20,21].

Methods

Knowledge database

The population selected is composed of 20 labeled single-EEG

*Corresponding author: Mechmeche S, National Engineering School of Tunis, University of El Manar, UR SITI, Tunis, Tunisia, Tel: 21671872253; E-mail: s.biotech@laposte.net

Received: March 31, 2016; Accepted: April 13, 2016; Published: April 21, 2016

Citation: Mechmeche S, Salah RB, Ellouze N (2016) Two-Stage Feature Selection Algorithm Based on Supervised Classification Approach for Automated Epilepsy Diagnosis. J Bioengineer & Biomedical Sci 6: 183. doi:10.4172/2155-9538.1000183

Copyright: © 2016 Mechmeche S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
composed of 48 features that are extracted from time, frequency and
time-frequency fields (Table 1):
LPC: Linear Predictive Coefficients
DFTC: Discret Fourier Transformation Coefficients
CC: Cepstral Coefficients
DHTC: Discret Hilbert Transformation Coefficients
WC: Wavelet Coefficients
STFTC: Short Time Fourier Transformation Coefficients

Training dataset
The training dataset is represented as (nxd) data pattern, it is
defined as:
\[ X_{TR}^{TR} = \begin{bmatrix} x_{1,k} \end{bmatrix}, \]
\[ 1 \leq i \leq n, 1 \leq k \leq d \]
\[ n: \text{Total number of samples}; d: \text{dataset dimensionality} \]
\[ x_{i,k}: \text{General term of training dataset} \]
The signals are manually labeled and ordered into two groups,
normal and epileptic, by an expert neurologist.
The "normal" group is defined by the following dataset:
\[ X_{TR}^{TR} = \begin{bmatrix} x_{1,k} \end{bmatrix}, \]
\[ 1 \leq i \leq N_{X_{1}^{TR}} \]
\[ N_{X_{1}^{TR}}: \text{Samples number of first group} \]
The "Epileptic" group is defined by the following dataset:
\[ X_{TR}^{TR} = \begin{bmatrix} x_{1,k} \end{bmatrix}, \]
\[ \left\lfloor N_{X_{1}^{TR}} + 1 \right\rfloor \leq i \leq N_{X_{2}^{TR}} \]
\[ N_{X_{2}^{TR}}: \text{Samples number of second group} \]

Feature selection algorithm
For the classification difficulty, wrapper feature selection consists
of selecting the features that maximize the classifier performance and
capable of discriminating samples that belong to different classes.
In this research, the classifier performance is evaluated from the
confusion matrix that derives the important metrics, such as, Accuracy,
Sensitivity and Specificity. The feature selection algorithm is composed
of the two following stages (Figure 2):

- IFE (Individual Feature Evaluation) stage,
- SBS (Sequential Backward Selection) stage.

Individual feature evaluation stage: In the first algorithm stage,
a wrapper feature selection method is used by applying the Individual
Feature Evaluation technique. The choice of the features is accorded
to the highest metrics that have been selected. Two classifiers have
been evaluated for this process: LDC (Linear Discriminant Classifier)
and QDC (Quadratic Discriminant Classifier) that provide the two
following Relevant Feature Subsets (RFS):
\[ F_{RFS - LDC}: \text{Relevant Feature Subset corresponding to the higher} \]
ranked-LDC metrics
\[ F_{RFS - QDC}: \text{Relevant Feature Subset corresponding to the higher} \]
ranked-QDC metrics.
In the output of the first stage, the algorithm compare between the
higher ranked LDC metrics and the higher ranked QDC metrics in
order to select the Smallest-Best Relevant Feature Subset (SBRFS):
\[ F_{SRFS - LDC}: \text{Smallest-Best Relevant Feature Subset usingLDC} \]
classifier
\[ F_{SRFS - QDC}: \text{Smallest-Best Relevant Feature Subset usingQDC} \]
classifier.
The output of the second stage provides the smallest relevant
feature subset \( F_{SRFS} \) that is finally obtained by selecting the best
smallest size between \( F_{SRFS - LDC} \) and \( F_{SRFS - QDC} \).

Mahalanobis distance classifier (MDC): Mahalanobis Distance
Classifier computes the distance \( d(x_{unk}, m_k) \) between unknown EEG
feature vector and the two classes “Normal” and “Epileptic” as follow:
\[ d(x_{unk}, m_k) = (x_{unk} - m_k)^{T}^{-1} (x_{unk} - m_k) \] (4)
\[ X_{unk}: \text{Unknown feature vector}; \]
\[ m_k: \text{Mean of the kth class}; \]
\[ T: \text{Covariance matrix of the learning dataset } X_{TR}. \]
Results and Discussion

First-stage experimental results

In the first part of individual feature evaluation (IFE) stage a 5-fold cross-validation procedure is used in LDC-classifier in order to estimate the metrics (Accuracy, Sensitivity and Specificity) of each feature (Figure 3). The algorithm chooses only the features having the higher metrics (Table 2 and Figure 3).

The feature subset deduced from the first stage using LDC-classifier will therefore be defined as:

$$F_{FRFS-LDC} = \{f_{16}, f_{19}, f_{20}, f_{21}\}$$

In the second part of individual feature evaluation (IFE) stage, a 5-fold cross-validation procedure is applied in QDC-classifier in order to estimate the metrics of each feature (Figure 4) and the algorithm selects only the features having the higher metrics (Table 3).

The feature subset deduced from the first stage using QDC-classifier will therefore be defined as:

$$F_{FRFS-QDC} = \{f_{16}, f_{19}, f_{20}, f_{21}, f_{28}, f_{42}\}$$

At the end of the first algorithm stage the smallest best relevant feature subset (SBRFS) have been selected by comparing between the metrics and the size of both $$F_{FRFS-LDC}$$ and $$F_{FRFS-QDC}$$ subsets, the SBRFS will therefore be defined as:

$$F_{SBRFS} = \{f_{16}\}$$

Second-stage experimental results

To reduce the dimensionality of $$F_{SBRFS}$$ we have used the SBS (Sequential Backward Selection) search method that starts with all features and removes a single feature at each step until the desired dimension with the highest metrics is reached. For each step a 5-fold cross validation is applied for the feature subset selection process. In the first part of the second-algorithm stage, the experimental results using LDC-classifier illustrates that the SBRFS (Smallest Best Relevant Feature Subset) is composed of the 16th feature:

$$F_{SBRFS-LDC} = \{f_{16}\}$$

In the second part of the second-algorithm stage, the experimental results using QDC-classifier illustrates that the SBRFS (Smallest Best Relevant Feature Subset) is also composed of the 16th feature:

$$F_{SBRFS-QDC} = \{f_{16}\}$$

The output of the second stage selects the smallest relevant feature subset comparing both the metrics and the size of $$F_{SBRFS-LDC}$$ and $$F_{SBRFS-QDC}$$ subsets, the SRFS (Smallest Relevant Feature Subset) is deduced as:

$$F_{SRFS} = \{f_{16}\}$$

The final experimental result of the two-stage algorithm feature selection is resumed in the following figure (Figure 5).

The combination of these two techniques (IFE and SBS) leads to reduce the dimensionality of the original feature set at only one best relevant feature that will be used in epilepsy diagnosis.

The diagnostic result was successfully tested (Figure 6) on EEG
A two-stage feature selection algorithm has been proposed in this article in order to remove the redundancy and to reduce the dimensionality of the dataset at the relevant feature subset. The mRMR (Minimum-Redundancy Maximum-Relevance) approach was successfully confirmed and tested in the first algorithm stage using IFE method, and the dimensionality of the relevant feature subset selected was successfully reduced in the second stage using SBS method at only one single best relevant feature that may be reduce considerably the processing time of the diagnostic. The performance of the results can be improved by using other robust dataset features and other classifier types for validation, such as the ANN (Artificial Neural Network), SVM (Support Vector Machine) and GA (Genetic Algorithm) methods. Using the automated IED’s diagnosis the doctor will no longer need to scan visually EEG signal leads.

### Literature Review

Table 4 show a comparative study on IED’s classification metrics in recent years, regardless of the number of features used: Our feature selection algorithm improves the classification metrics for both LDC and QDC classifier using the single-best relevant feature selected (Table 4).

### Conclusion

A two-stage feature selection algorithm has been proposed in this article in order to remove the redundancy and to reduce the dimensionality of the dataset at the relevant feature subset. The mRMR (Minimum-Redundancy Maximum-Relevance) approach was successfully confirmed and tested in the first algorithm stage using IFE method, and the dimensionality of the relevant feature subset selected was successfully reduced in the second stage using SBS method at only one single best relevant feature that may be reduce considerably the processing time of the diagnostic. The performance of the results can be improved by using other robust dataset features and other classifier types for validation, such as the ANN (Artificial Neural Network), SVM (Support Vector Machine) and GA (Genetic Algorithm) methods. Using the automated IED’s diagnosis the doctor will no longer need to scan visually EEG signal leads.

### References