Epileptic Seizure Prediction Using Data Mining Algorithms.

Ilyas Zidane, Jamal Mhamdi and Nordine Zidane

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

November 18, 2019
Abstract— Epileptic seizures are caused by sudden electric discharges in neurons irregular transient operation; hence this generate cerebral neurological dysfunction. The seizures can occur in different regions of the brain and are reflected in various clinical expressions. As it happens, depending on the location and job of the affected neuron, seizures with different types occur: simple partial seizures (SPS), complex partial seizures (CPS), secondarily generalized seizures (SGS) with epilepsy locations classified in temporal lobe epilepsy (TLE), frontal lobe epilepsy (FLE), posterior epilepsy (PE including the parietal, occipital, occipito-temporal, temporo-occipito-parietal junction regions), operculo-insular epilepsy (OIE) and multifocal epilepsy (MFE).

In this paper, we propose a new approach of epilepsy seizure prediction by using data mining algorithms to analyze seizures of a huge number of patients EEGs. This method is most powerful and most accurate than the others prediction methods.

Keywords—Epilepsy, Seizure, EEG, Data mining, Supervised learning, Algorithms.

I. INTRODUCTION

Epilepsy is the most common neurological disease after migraine. It has been reported to affect 0.5% to 1% of the world's population, namely 50 million people in the world. It is estimated that about 430,000 people are affected in France, with 33,000 new cases each year. Nearly 5% of persons are experienced a seizure in their life time and 1% of persons suffer of epileptic seizure [1].

The statistics shows that Epilepsy increases the death risks. Although medicine generate many antiepileptic drugs, but they are neither curative and nor efficient to reduce seizures. Some patients are antiepileptic drugs resistant; in this case the solutions are surgical treatment to remove the epileptogenic area or cerebral electrical stimulation.

Epilepsy disease affects all ages, most affected are children and old persons, thus 75 % of diseases are established before 18th birthday. We analyze a hundred epileptics patients of all categories of age: Children (5-14 years), Youth (15-24 years), Adults (25-64 years) and Seniors (65 years and over). In babies, seizures might not be obvious to an observer. Many methods have been developed for seizure prediction over several years since 1995. The most of these methods used state similarity features for seizure prediction[2],also they employed independent component analysis to separate isolated epileptiform discharges from EEG background and analyse magnetoencephalography data to localize epileptic spikes using independent components analysis and spatio-temporal clustering [3],they used EEG spatiotemporal correlation structure with delay correlation and covariance matrices [4], they applied autoregressive modeling and least-squares parameter estimator [5], they studied effects of linear univariate features [6], Classification of patterns of EEG synchronization with bivariate features [7], they used EEG spatiotemporal correlation structure with delay correlation and covariance matrices [4], they applied autoregressive modeling and least-squares parameter estimator [5], they studied effects of linear univariate features [6], Classification of patterns of EEG synchronization with bivariate features [7], they used spectral power from raw and bipolar time-differential signals [8] and combination of mean phase coherence and the dynamic similarity index methods [9][10], they used cost-sensitive support vector machines in spike rate [11] and they employed Fourier transformation to perform signal processing of epileptic EEG [12].

II. DATA SET DESCRIPTION

The data set is issued from EEGs recordings of epileptic’s patients. The information in large clinical EEGs database has been entered during two years because of the huge number of patients and their time’s unavailability.

We choose only hundred patients to study the prediction of their healing according to the result of treatment using data mining algorithms so to predict life threatening caused by epileptic seizure and may be able to analyze signals in real-time and having a dashboard to generate alerts before the time of epileptic seizure.

The sampling frequency is 256 Hz with 128 channels of epileptic EEG signals which are recorded by focal electrodes form brain locations: temporal lobe epilepsy (TLE), frontal lobe epilepsy (FLE), posterior epilepsy (PE including the parietal, occipital, occipito-temporal, temporo-occipito-parietal junction regions), operculo-insular epilepsy (OIE) and multifocal epilepsy (MFE). The records may be taken before the seizure (Preictal ), during the seizure(ictal), between seizures (interictal ) and after the seizure(postictal). Preictal or prodromes the time before the seizure. It can last from minutes to days and make patient act and feel differently, for example, duration of ictal varies from a few seconds to 5 minutes. Therefore, ictal records during epileptic seizures can have at least 60 minutes of preictal signal preceding each seizure.
We have used the following fields of dataset table 1: Patient number(N), Age, Sex, type, Brain location, Total seizures(seiz) and Healing(H).

### III. DECISION TREE WITH FEW DATASET

In this paper, our method is to use an ameliorated supervised learning algorithm in order to build a decision tree which shows clearly the prediction patient’s response to treatment with a great precision better than other methods.

The problem is to find a function which would allow to represent the prediction?

The function would be minimum when the node is pure (all data are in the same class) and maximum when the data is equally distributed.

Two functions are often used in binary decision trees:

**Entropy**

\[
\text{Entropy}(p) = - \sum_{k=1}^{c} P(k|p) \times \log(P(k|p))
\]

**Gini**

\[
\text{Gini}(p) = 1 - \sum_{k=1}^{c} P(k|p)^2
\]

\[
= 2 \times \sum_{k \neq k'} P(k|p) \times P(k'|p)
\]

Also, we need a function to choose a test:

**Gain**

\[
\text{Gain}(p, \text{test}) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)
\]

\(p\) : position  \(\text{test} \) : test with n parameter  

\(P_j\) : part of sample items in p position moving to p\_j position.

\(i(p)\) : Entropy (p) or Gini (p)

It's obvious to remind some definition: Entropy is information’s measurement (or rather lack thereof). The information gain is calculated by making a split. Gini measurement is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch.

A decision tree contains a root node and the other nodes are growing upside down as new nodes by splitting the data into levels. A node of any level could have a node as parent and others as children nodes.

#### GENERIC ALGORITHM FOR BUILDING A TREE

**input:** sample

**begin**

Initialize to the empty tree

the root is the current node

repeat

Decide if the current node is terminal

if the node is terminal then

Assign a class

else

Select a test and create the subtree

end if

Skip to the undrawn tracking node if there is one until you get a decision tree

end

Table 2 below is a random extraction of 8 records sample from databases to apply Iterative Dichotomizer algorithm in order to generate decision tree.

### TABLE 2. PARTIAL DATASET

<table>
<thead>
<tr>
<th>N</th>
<th>A</th>
<th>S</th>
<th>T</th>
<th>L</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>Youth</td>
<td>Male</td>
<td>SPS</td>
<td>Temporal</td>
<td>yes</td>
</tr>
<tr>
<td>41</td>
<td>Youth</td>
<td>Female</td>
<td>SGS</td>
<td>Frontal</td>
<td>no</td>
</tr>
<tr>
<td>57</td>
<td>Adults</td>
<td>Female</td>
<td>CPS</td>
<td>Frontal</td>
<td>no</td>
</tr>
<tr>
<td>70</td>
<td>Youth</td>
<td>Male</td>
<td>CPS</td>
<td>Frontal</td>
<td>yes</td>
</tr>
<tr>
<td>75</td>
<td>Children</td>
<td>Male</td>
<td>SPS</td>
<td>Parietal</td>
<td>yes</td>
</tr>
<tr>
<td>80</td>
<td>Adults</td>
<td>Male</td>
<td>SGS</td>
<td>Parietal</td>
<td>no</td>
</tr>
<tr>
<td>82</td>
<td>Adults</td>
<td>Male</td>
<td>SPS</td>
<td>Parietal</td>
<td>no</td>
</tr>
<tr>
<td>85</td>
<td>Youth</td>
<td>Female</td>
<td>CPS</td>
<td>Temporal</td>
<td>no</td>
</tr>
</tbody>
</table>

Seizure Type : \(T = \{\text{SPS, CPS, CGS}\}\)

Age Range : \(A = \{\text{Children, Youth, Adults}\}\)

Brain Location : \(L = \{\text{Temporal, Frontal, Parietal}\}\)

SEX : \(S = \{\text{Female, Male}\}\)

Healing : \(H = \{\text{yes, no}\}\)

We want to build a decision tree that is able to determine if a patient is healing and therefore treatment responsive according to the attributes: A, S, T and L.

**Tree root (no test)**

**Label (3,5) for : (L(C,yes), L(C,no)).**

What is the first test to perform?
The test on T is not discriminatory as it is shown in figure 1.
The test on A is interesting on the branches Children and Adults.
The chosen test is the one with the biggest gain
Data processing using entropy:

A node p is terminal if: all the items associated with this node are in the same class or if no test could be selected.
We have chosen the test that maximizes:

\[ \text{Gain}(X) = \sum \frac{n_x}{n} \log \frac{n}{n_x} \]

So the test A(Age) is chosen because it has the maximal gain.
We use Iterative Dichotomiser supervised learning algorithms with the function below:

function Iterative_Dichotomiser (I, O, T)
I set of input attributes
O output attribute
T set of learning items
If (T is empty)
    Resend Error
If (all items of T belong to the same class)
    Return a node with the class label
If (I is empty)
    Returning a node with the most frequent label on the output attribute of T
Calculate the information gain for all attributes of I relative to T
X is the attribute with the largest gain
\{x_j / j = 1, 2, ..., m\} are the values of X
\{T_j / j = 1, 2, ..., m\} are the subsets of T decomposed with respect to x_j
Returning a tree with X as the root node's label and x_1, x_2, ..., x_m as labels of arcs going to trees

Iterative_Dichotomiser(I-\{X\},O,T_1),
Iterative_Dichotomiser(I-\{X\},O,T_2),...,
Iterative_Dichotomiser(I-\{X\},O,T_m).

After executing the iterative dichotomiser algorithm with all subtrees the final decision tree for epileptic predictive healing is:

Figure 3 shows clearly that Children (5-14 years) are responsive to treatment; Adults (25-64 years) are not. The Youth (15-24 years) male are responsive but not female. It is shown in figures 3 and 4 with percentages of responses to treatments.

Tree size: 5 nodes, 3 leaves
Edge widths: Relative to parent
Target class: H=yes

Tree size: 5 nodes, 3 leaves
Edge widths: Relative to parent
Target class: H=no
IV. DECISION TREE WITH GLOBAL DATASET

Our method predicts epileptic healing in this sample for only 8 patients dataset. We complete the experiment within 100 patients dataset and we find the same conclusion of decision tree.

We use Iterative Dichotomizer supervised learning algorithms with all patients dataset.

The settings of model parameters configuration are as follow:
- Pruning: at least six instances in leaves, at least two instances in internal nodes, maximum depth 10
- Splitting: Stop splitting when majority reaches 100% (classification only)
- Binary trees: Yes

Data:
- Data instances: 100
- Features: Age, Sex, type, Brain Location
- Target: Healing

We obtain 19 nodes with 10 leaves the tree is illustrated in figure 7 below:

V. Conclusion

This paper analyze a large epileptic seizures database containing a lot of records EEGs signal processing despite of its complexity. Most of predictive methods use total seizures or cost-sensitive support vector machines without analyzing the other classes and features, thus will involve inaccurate predictions.

Our method use all classes, features and parameters thanks to effective search easily and efficiently of pertinent and significant data using data mining algorithms.

The prediction using supervised learning algorithm to generate decision tree, applied to 8 records in the epileptic dataset is nearly the same applied to 100 records.

The challenge for the future is to predict the epileptic seizure time before at least several hours.
REFERENCES


