BRAIN DATA ANALYSIS AND MANAGEMENT

R. Ilieva¹, PhD, Associate Professor in Automated Systems for Data Processing and Management
P. Georgieva², PhD, Head of Signal Processing Lab (SPL), IEETA
S. Petrova¹, MEng Student in Electronic Management, ELFE
Technical University of Sofia, Bulgaria¹,³
University of Aveiro, Portugal²

Machine Learning (ML) techniques have been extensively applied in bioinformatics. In this paper, we chose RapidMiner software to analyze brain data (EEG signals) in order to discriminate human emotions while subjects were observing images. Five ML classification algorithms were studied: k-Nearest Neighbor (kNN), Naive Bayes, Support Vector Machine, Artificial Neural Networks and Decision Tree. kNN and ensemble classifiers achieved above 80% accuracy with test data. This is a very encouraging result taking into account the fact that brain signals are highly non-stationary and noisy and therefore it is quite challenging data for analysis and management.

Keywords: bioinformatics, brain, Artificial Neural Networks, k-Nearest Neighbor (kNN), Naive Bayes, Support Vector Machine, Decision Tree.

Conference participants

I. Introduction

Machine Learning (ML) is a subarea of artificial intelligence connected with the design, analysis, implementation and application of programs which study on examples [2]. Learning from data is commercially and scientifically important. ML consists of methods that automatically extract interesting knowledge (patterns, models, relationships) in large databases.

The goal of this work is to find a reliable ML algorithm (or a combination of ML techniques) able to discriminate positive and negative human emotions based on Electroencephalogram (EEG) signals. The EEG data was collected while subjects were exposed to images typically provoking positive and negative emotions.

This paper is organized as follows. Section 2 briefly introduces the ML techniques studied. Section 3 describes the dataset, the acquisition process and the metrics used to analyze the classifier performance. In Section 4 classification results are summarized, and finally our conclusions are presented in Section 5.

II. Machine Learning techniques

The ML classification techniques applied in the present study are briefly introduced.

1. **K-Nearest Neighbor (k-NN)**

K-NN is a widely applied classifier [1]. The class of a new example (object) is defined based on the majority vote of the K nearest training examples with known class labels (Fig.1). K is usually a small positive integer unpaired number – this is to avoid that two classes have the same votes. kNN computes the most common class among the k Nearest Neighbors of the new example and assigns this class to the example. For K=1 the object is simply assigned to the class of its nearest neighbor.

2. **Support Vector Machine (SVM)**

SVM is a classification technique that defines the hyper-plane maximizing the margin between two classes [1]. There may be several options (lines) that separate two classes (+/- on Fig. 2). SVM determines the closest objects between the two classes (termed support vectors). The optimal separation line is the one that maximizes the distance between the classes. For linearly non separable classes (Fig. 3) the concept of kernel induced feature space is formulated. The Kernel SVM transforms the original data into a higher dimensional feature space where data is already linearly separable and then applies the same procedure as the one described above.

3. **Artificial Neural Network (ANN)**

ANN is a mathematical architecture inspired by the structure and functionality of biological neural networks [1]. ANN consists of layers: typically one input, one hidden and one output layer (Fig. 4). Each layer has a

Fig. 1. kNN classif. (K=3)

Fig. 2. Linear SVM

Fig. 3. Nonlinear SVM

Fig. 4. General ANN architecture
number of parallel processing elements (PE), termed neurons (or nodes), which are mathematical functions that mimic the dynamical behavior of biological neurons at a macro scale. Due to their properties of adaptation, noise filtering and parallel processing, the ANNs are a powerful framework for classification and regression. Their main disadvantage is the usually long processing time for networks with high dimension (high number of PE).

4. Decision Tree (DT)

DT is a classification technique based on the principle of complex division of a problem into a sequence of sub-problems and thus generation of a decision tree (Fig. 5).

III. EEG Signal Classification

1. EEG signals acquisition.

The goal of this study is to distinguish emotional bio-signals evoked by viewing selected affective pictures from the International Affective Picture System (IAPS) [5]. EEG-based brain-computer interfaces consist of very typical components, each of which performs its own critical function. Figure 1 shows the process cycle. Firstly, a stimulus set and test protocol are needed (1). During testing (2), the test subject will be exposed to the stimuli according to the test protocol. The resulting voltage changes in the brain are then recorded (3) as an electroencephalogram, from which noise and artifacts are removed (4). The resulting data will be analyzed (5) and relevant features (like power spectra) will be computed. Based on a test set from these features a classifier will be trained (6), and the rest of the data will be classified using this classifier. This step provides an interpretation of the original raw brain signals. The feedback step will not be used during this research. It is shown for the sake of completeness.

A total of 26 female volunteers participated in the study. 21 channels of EEG were recorded – Frontal and Parietal (FP), Frontal (F); Temporal (T), Central (C), Parietal (P) and Occipital (O) channels. A total of 24 of high arousal (> 6) images with positive valence (7.29+/–0.65) and negative valence (1.47+/–0.24) were selected. Each image was shown 3 times in a pseudo-random order and each trial lasted for 3500 ms: during the first 750 ms, a fixation cross was presented, then one of the images was shown for 500 ms and at last a black screen – for 2250 ms. Three schemes were implemented by choosing three different filters and detecting n maximums and n minimums at the output of the filters.

i) Butterworth filter of fourth order with passband [0.5 – 15] Hz. 12 features are stored according to the latency (time of occurrence) and amplitude of the first 3 maximums and minimums (Fig. 7a).

ii) Butterworth filter of fourth order with Delta band [0.5 – 4] Hz. 8 features are stored according to the latency and amplitude of the first 2 maximums and minimums (Fig. 7b).

iii) Butterworth filter of fourth order with Theta band [4 – 8] Hz. 12 features

![Fig. 5. Example of Decision Tree](image1)

![Fig. 6. Brain Study Spiral Model](image2)

![Fig. 7: Filtered signal (with fourth order Butterworth filter) and features detection: positive (line) & negative (dot)](image3)
are stored according to the latency and amplitude of the first 3 maximums and minimums (Fig. 7c).

2. Classifier performance metrics

The classifier performance is analyzed by the Confusion Matrix [3]. The basic structure of a Confusion Matrix for a two-class problem is presented in Table 1.

Where TP (true positive) and TN (true negative) are the numbers of correctly classified as positive and negative examples respectively, while FP (false positive) and FN (false negative) are the numbers of wrongly classified as positive and negative examples respectively. The following performance measures are determined from the Confusion Matrix.

- **Accuracy** is the fraction of all (positive and negative) correctly classified examples [3]:
  \[
  \text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}
  \]

- **Precision** is the fraction of correctly classified positive examples from all classified as positive:
  \[
  \text{Precision} = \frac{TP}{TP + FP}
  \]

- **Specificity** is the fraction of correctly classified negative examples from all classified as negative:
  \[
  \text{Specificity} = \frac{TN}{TN + FP}
  \]

- **Recall** is the fraction of positive examples correctly classified as positive examples from all positive examples:
  \[
  \text{Recall} = \frac{TP}{TP + FN}
  \]

While accuracy is a performance measure more typical for balanced data (data with similar number of examples of all classes), the other measures (precision, specificity and recall) are more adequate for the unbalanced data.

| ACTUAL CLASS | PREDICTED CLASS |
|--------------|-----------------|
| Class = Yes  | TP              |
| Class = No   | FN              |

Tab. 1

Confusion Matrix

Fig. 8. The process of data loading, filtering and storing

Fig. 9. X-Validation of five classifiers applied to ClassAB_Data

IV. EEG classification results with Rapid Miner (RM)

RM is an open-source data mining system [4]. It is available as a stand-alone application for data analysis and as a data mining engine for integration into other products. RM is an environment for machine learning, data mining, text mining, predictive analytics and business analytics.

1. Data Preprocessing

Before information extraction and classification the data was conveniently preprocessed, filtered and stored as follows.

- **Data normalization:** in order to avoid computational problems before loading into RM.
- **Data Cleaning and storing:** Fig. 8 depicts the process of: i) data loading into RM with the Read CSV operator; ii) filtering (removing) the lines with missing values (that correspond to lines with zeros) with the Filter Examples operator; and iii) storing the cleaned data into Brain Study Repository using the Store operator.

The same process was applied for the three datasets corresponding to Filter 1, Filter 2 and Filter 3.

2. Classification based on all attributes and all channels

The goal is to find the best binary (two-class) classifier based on all attributes and all channels. This scenario is implemented for three data sets Filter 1 (ClassAB_Data), Filter 2 (ClassABdelta_Data), Filter 3 (ClassABtheta_Data). Five classifiers were compared. Figure 9 shows the summarized process of simultaneous training and testing of five classifiers with the same data set loaded by the Retrieve Data operator. The Multiply operator provides the same data to each classifier represented by the X-Validation operator.

After playing the process, we obtained all values for accuracy, recall and precision from the confusion matrix. These performance measures are summarized in Table 2 for 5 classifiers, applied to 3 data sets. The highest classification rate is obtained for KNN classifier trained with CLassABdelta dataset.
C. Classifier optimization by feature selection of data set with all channels

Three datasets considered have between 8 and 12 features. Now we want to explore the possibility of improving the classification applying feature selection (reduction) procedures. Forward Selection and Backward Elimination are among the most typical operators in RM for extracting the most influential features.

KNN classifier optimization

In order to illustrate the plausibility of the feature selection procedure, KNN classifier is optimized by both operators. The performance results in terms of confusion matrix and the associated values for accuracy, precision and recall are summarized in Fig. 10. Here, Forward Selection leads to higher accuracy compared to the Backward Elimination.

NN (neural network) classifier optimization

The effect of feature selection procedure strongly depends on the classification algorithm. The results of NN classifier optimization are presented on Figure 11. In this case the latency 2 is the remover feature and the Backward Elimination leads to higher accuracy compared to the Forward Selection.

D. Classification based on all attributes and selected channels

Taking into account that Parietal and Occipital channels are responsible for visual processing in Group 1 we have isolated only Parietal channels, in Group 2 – only Occipital channels and in Group 3 – the combination of Parietal and Occipital channels from the complete dataset ABdelta data (Table 3, Table 4, Table 5). The highest classification rate was obtained for KNN classifier trained with ClassABdelta dataset.

E. Ensemble classification

Ensemble classification uses a combination of n-learned classifiers M1, M2, … Mn, in order to build an improved composite model [2]. In this study we applied the ensemble method called “Bagging” where the final result is obtained by the majority vote principle.
The sequence of nested processes aimed at implementation of ensemble Bagging classification with five classifiers is depicted on Fig. 12. In the previous study, DT and NB classifiers exhibited worse performance (see Table 2), therefore they were removed. The results are summarized on Fig. 13a) and b). The best ensemble classification (SVM, NN, KNN) does not provide significant advantages in comparison with the KNN from Table 3.

VI. Conclusions
In this paper a number of Machine Learning methods are studied and applied to a challenging classification problem of discriminating human emotions based on the EEG brain data. Among five classifiers, K-Nearest Neighbor (kNN) provided the best discrimination (84% accuracy) as an individual classification model. Ensemble classification (combination of SVM, NN and KNN classifiers) achieved slightly better results (85%). This study has shown that the preprocessing step on the raw data collected from the EEG machine is crucial for extracting discriminative patterns. First, we need to choose the frequency band that adequately reflects...
the affective brain states. Our conclusion is that apparently [0.5-4] Hz is the most corresponding band. Then, the most suitable features need to be identified. The classification is clearly affected by the choice of temporal (selected amplitudes and latencies) and spatial (selected channels) features. Last but not least, the classification algorithm is also important for the emotional patterns recognition. In case if no individual classifier is satisfactory, a mixture of weak classifiers (termed ensemble classification) may be a reasonably good alternative. All required data analysis steps fit nicely to the modular structure of RM software platform and make the pipeline of procedures quite clear and intuitive. We are confident that the methodology studied in this paper can be easily adapted to other problems in bioinformatics or other private and public sectors, such as banking, insurance and medicine.

Acknowledgement
The research, described in this paper, was carried out within the framework of the contract № IUNF 13000. All experiments are conducted during the Erasmus study period of the third author in the University of Aveiro (UA) under the supervision of the other two authors. The Erasmus scholarship provided by Technical University of Sofia and the excellent working conditions in UA are highly acknowledged. We would like to express gratitude to the PsyLab from UA, and particularly to Dr. Isabel Santos, for providing the data set.
References:

1. Han, J., M. Kamber, Data Mining Concepts and Technique, University of Illinois at Urbana. – Champaign., San Francisco., CA94111, Second Edition, 2006.
2. Bramer, M., Principles of Data Mining, Springer. – London., 2007.
3. Pang-Ning T., M. Steinback, V. Kumar, Introduction to Data Mining, 2003.
4. RapidMiner 5 software platform. Website: http://rapid-i.com/content/view/181/190/.
5. Frantzidis, C., et al., “On the classification of emotional biosignals evoked while viewing affective pictures: An integrated Data-Mining-Based approach for Healthcare applications.
6. Shoikova E., A. Peshev, M. Krumova, "ePortfolio – Identity and Professional Development", ICEIRD 2011, 5-7 May 2011. - Ohrid, Macedonia, Proceedings CD, pp. 1061-1068.
7. Nikolova, I., W. Jahn, Die Einführung und Konkurrenzvorteile des TQM, Wissenschaftliche konferenz Innovationen und wettbewerbsfähigkeit, Karlsruher Institut für Technologie, TU Sofia, TU Braunschweig, FOM Essen.
8. Tamošiūnienė, R., K. Angelov, Project and Programme Management and Evaluation, C., 2011.

Information about authors:

1. Roumiana Ilieva - Ph.D., Associate Professor in Automated Systems for Data Processing and Management Technical University of Sofia; address: Bulgaria, Sofia city; e-mail: rilieva@tu-sofia.bg
2. Petia Georgieva - Ph.D., Head of Signal Processing Lab (SPL), IEETA, University of Aveiro; address: Portugal, Aveiro city; e-mail: petia@ua.pt
3. Stanislava Petrova - MEng Student in Electronic Management, ELFE, Technical University of Sofia, address: Bulgaria, Sofia city; e-mail: stanislava_asparuhova@abv.bg