Applied Computer Science, vol. 16, no. 4, pp. 56–63
doi:10.23743/acs-2020-29

Submitted: 2020-11-23
Revised: 2020-11-30
Accepted: 2020-12-17

machine learning, EEG signal, classification,
data balancing, feature extraction

Amina ALYAMANI [0000-0002-0286-7105]*,
Oleh YASNIY [0000-0002-9820-9093]**

CLASSIFICATION OF EEG SIGNAL BY METHODS OF MACHINE LEARNING

Abstract

Electroencephalogram (EEG) signal of two healthy subjects that was available from literature, was studied using the methods of machine learning, namely, decision trees (DT), multilayer perceptron (MLP), K-nearest neighbours (kNN), and support vector machines (SVM). Since the data were imbalanced, the appropriate balancing was performed by Kmeans clustering algorithm. The original and balanced data were classified by means of the mentioned above 4 methods. It was found, that SVM showed the best result for the both datasets in terms of accuracy. MLP and kNN produce the comparable results which are almost the same. DT accuracies are the lowest for the given dataset, with 83.82% for the original data and 61.48% for the balanced data.

1. INTRODUCTION

Hypnotic therapy is a method of psychotherapy that helps to heal a large number of disorders, including stress, depression, anxiety, pain (Provençal et al., 2018; Wood & Bioy, 2008), and eliminating the unwanted memories in patient mind (Terhune et al. 2017). Hypnosis can also enhance thought suppression by minimizing the effect of cognitive load (Bryant & Sindicich, 2007). The influence of hypnosis on the human being can be assessed by registering brain signals.

The widely used technique for recording brain signal is electroencephalogram (EEG) (Sanei & Chambers, 2007). EEG signal is a miniature amount of electrical flow in a human brain that holds and controls the entire body (Haykin 2009).

* Omar Al-Mukhtar University, Faculty of Engineering, Department of Computer Science, West Shiha, Dernah, Libya, amina.alyamani@yahoo.com
** Ternopil Ivan Pul'uj National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Mathematical Methods in Engineering, Ruska 56, 46001, Ternopil, Ukraine, oleh.yasniy@gmail.com

EEG signal can be employed to diagnose Alzheimer disease (Podgorelec, 2012), to predict epileptic seizure (Satapathy, Jagadev & Dehuri, 2017), to detect mental disorders (Dvey-Aharon et al., 2015; Thilakvathi et al., 2017). The approaches of machine learning give the possibility to analyze the EEG signal and draw appropriate conclusions based on the results of performed analysis. In particular, various classification methods can help to diagnose the mentioned above diseases. In paper (Parvinnia et al., 2014), EEG signals were classified using adaptive weighted distance nearest neighbor algorithm. In the study (Amin et al., 2017), the pattern recognition approach was employed to classify the EEG signals. Often EEG signals contain artifacts which should be found and treated respectively. Paper (Lawhern et al., 2012) gives the methods to detect and classify the subject-generated artifacts in EEG signals using auto-regressive models. Results, obtained in the mentioned above study, indicate reliable classification among several different artifact conditions across subjects.

However, despite the numerous application of machine learning approaches to the EEG signals, there still remains a wide range of potential activity and many interesting problems to be solved by means of computer and respective algorithms.

The present study utilizes the dataset obtained in the frames of Horizon 2020 program, which is available at (Real & Kübler, 2014). EEG signals of two healthy subjects were recorded (S01: right-handed male; S02: right handed female) (Real & Kübler, 2014). The subject sat in a comfortable chair. Stimuli were presented in two conditions. In an active condition, the subject was told to listen to a tone stream, and that he/she should count the occurrence of the odd (low) tones. In a passive condition, the subject was told to listen to a series of tones and that he/she would just have to listen to the tones. After the first recording ("PRE"), the subject listened to an Erickson-type hypnotic induction, where the subject imagined being on a ship in dense fog, making hearing and seeing difficult. Then, the EEG experiment was repeated (datasets "POST"). Finally, the subject listened to instructions designed to take back a hypnotic state (Real & Kübler, 2014).

2. METHODOLOGY

2.1. Dataset preparation

The dataset S01 No5 consists of four EEG data matrices, Xs, where the columns of each X show outputs of 31 recording electrodes and rows of each X contain recorded data in every 0.001953125 second (or 0.002 for simplification) from each electrode. Since the sampling rate is 512 Hz, the sampling period is around 0.002 second. Also, the duration of each stimuli is 50 ms, which means that every data-point is a 25 by 31 submatrix (each stimulus is 50 ms and by dividing it to the sample period 2ms on gets the number 25).

The dataset provides the time period when each stimulus is presented. This time period is available in the vector ‘trial’. The class label of each stimuli is available in the vector ‘y’. So, it is possible to find each data-point along with its label. For example, the first stimuli starts at time period 19898 ms, so to find the EEG response of this stimuli one needs to extract the submatrix of X, i.e. the row number $19898 \div 2$ of matrix X to row number $(19898 \div 2 + 25)$ of matrix X.

Also, the label of this data-point is provided in the first entry of vector ‘y’ in data cell number 1 of the S01 file. This label is ‘2’ which means that this stimulus is from class ‘frequent’. There are two classes of data: ‘odd’ and ‘frequent’. The ‘odd’ class is represented with label ‘1’.

There was written the code to perform the mentioned-above procedure in order to prepare the dataset. This code reads all datasets S01 and S02 and concatenates them according to the described procedure.

After the data preparation step was performed and for the better understanding of the samples of dataset, 9 samples of the dataset were visualized from both male and female data. Fig. 1 shows the visualized results. In each subplot, it is possible to see one sample of stimuli along with its label on the top. Each sample consists of 31 stimuli which have been drawn with different colors. The horizontal and vertical axes show time period and amplitude of the stimuli; respectively. As it is possible to see, the amplitude value of each stimulus changes with time and this change is significant. Another important point is that it can be observed from the figure that there are some distinct patterns in the behaviors of stimuli. For the male data, i.e. S01.mat, the stimuli of the ‘odd’ class somehow shows a descending behavior during the time while the stimuli of the other class show an ascending and then descending behavior in their shapes. Similarly, it is possible to note some distinctive patterns in the female data, i.e. S02.mat.

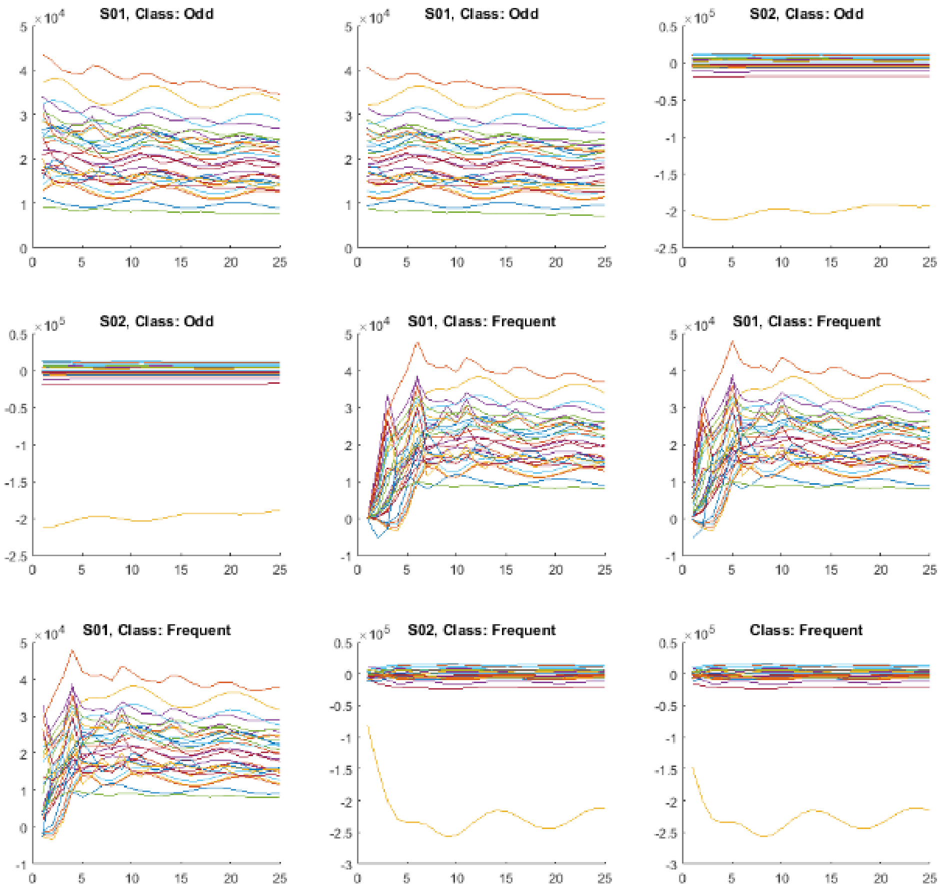

Fig. 1. A graphical view of the stimuli samples of dataset no. 5 along with their labels

2.2. Feature extraction

Features extraction is a common way to extract meaningful features from the EEG data (Li et al., 2015; Sun et al., 2019). In this way, an Autoencoder was used to find a new representation of the data in a lower dimensional space. Autoencoder is an unsupervised method of machine learning that provides a new representation of data in a lower dimensional space (Hinton & Salakhutdinov, 2006). In other words, an autoencoder is a type of artificial neural network used to learn efficient data encoding in an unsupervised manner. The aim of the Autoencoder is to learn a representation for a set of data, typically for dimensionality reduction. Along with the reduction side, a reconstructing side is learnt, where the autoencoder tries to generate the representation as close as possible to its original input from the reduced encoding.

The following model is used for the autoencoder network. The input size was $25 \cdot 31$, which is equal to 775. Hidden layer size is 64. Output layer size was $25 \cdot 31$, that is the same as input size.

2.3. Balancing the data

Another difficulty of the data is that it is an imbalanced data. In other words, the ratio of its classes is highly different that comes from the fact that the 'frequent' class has much more data than the 'odd' class. Originally, there are more than 3000 data points for the 'frequent' class and 480 data points for the 'odd' class of data, i.e. 3000 >> 480.

In this experiment, two scenarios were employed. In the first scenario, the original data were used. In the second scenario, there was an attempt to turn data into a balanced data. Both scenarios were implemented and the results were presented:

1. For the first scenario the data were not changed and all of the samples were kept.
2. In the second scenario, an under-sampling technique was employed to turn the data into a balanced one. To accomplish that, Kmeans clustering algorithm (MacQueen, 1967) was applied over the data points of the 'frequent' class. The number of clusters was set to 1000, and Kmeans algorithm was started. After the learning, the centers of clusters were treated as the new data points for the 'frequent' class of data. The number 1000 was determined empirically.

After that step, 1000 data points were obtained for the 'frequent' class of data and 480 data points were for the 'odd' class of data. Although the dataset is not exactly balanced, nevertheless, this ratio of classes samples is more balanced than the original one.

2.4. MLP structure

The multilayer perceptron (MLP) neural network was employed which uses the gradient decent back-propagation for tuning the parameters of the network (Haykin, 2009). The MLP topology is the following.

Input size, that is the number of nodes in input layer was 64, number of hidden layers was 1, number of neurons in each hidden layer was 32, number of neurons in output layer was 1.

3. RESULTS

To verify the models, 10-fold cross validation method was performed. In this way, the data were split into ten parts and each time one part was taken as the test data and the rest was treated as training data.

The classification was performed by four methods: support vector machines (SVM) (Cortes & Vapnik, 1995), Decision Trees (Quinlan, 1986), K-nearest neighbors (Altman, 1992) and MLP (Haykin, 2009). Table 1 presents the results in terms of accuracies along with the standard deviations. According to the results shown in Table 1, it can be seen that SVM performs better in both datasets and provide higher accuracies with accuracy 87.47% for the original data and 66.95% for the balanced data. Also, MLP and kNN produce the comparable results which are almost the same. DT accuracies are the lowest for the given dataset, with 83.82% for the original data and 61.48% for the balanced data.

Tab. 1. Comparison of different methods in terms of accuracies

Dataset	SVM	kNN (k = 8)	DT	MLP
Original data	87.47 ± 1.51	87.108 ± 1.56	83.82 ± 1.58	87.03 ± 1.77
Balanced data	66.95 ± 1.89	64.05 ± 4.03	61.48 ± 3.24	65.74 ± 2.06

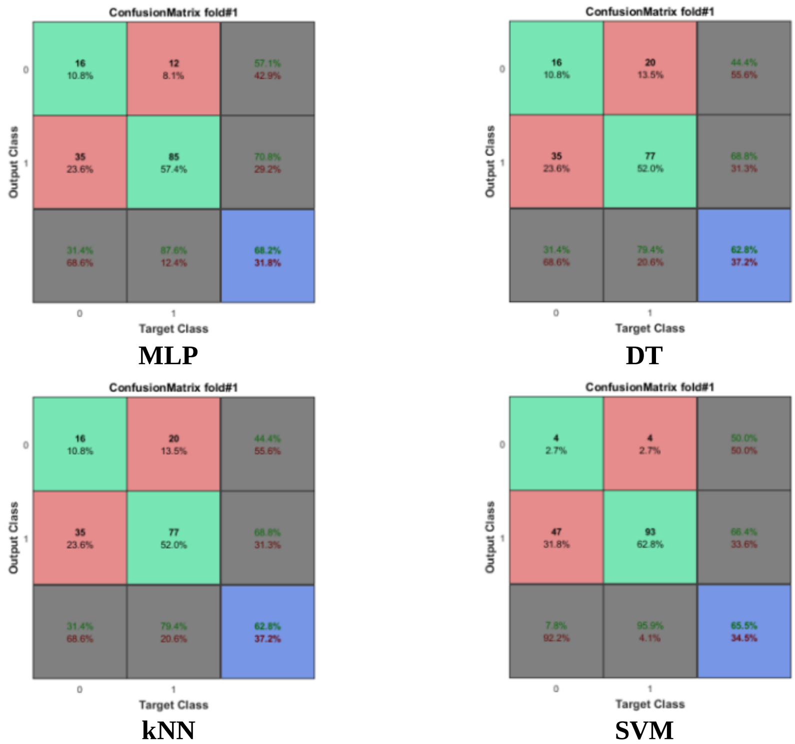

Fig. 2. Confusion matrices over the balanced data (Fold #1)

Fig. 3 presents the confusion matrices over the original dataset, i.e. the imbalanced version of dataset.

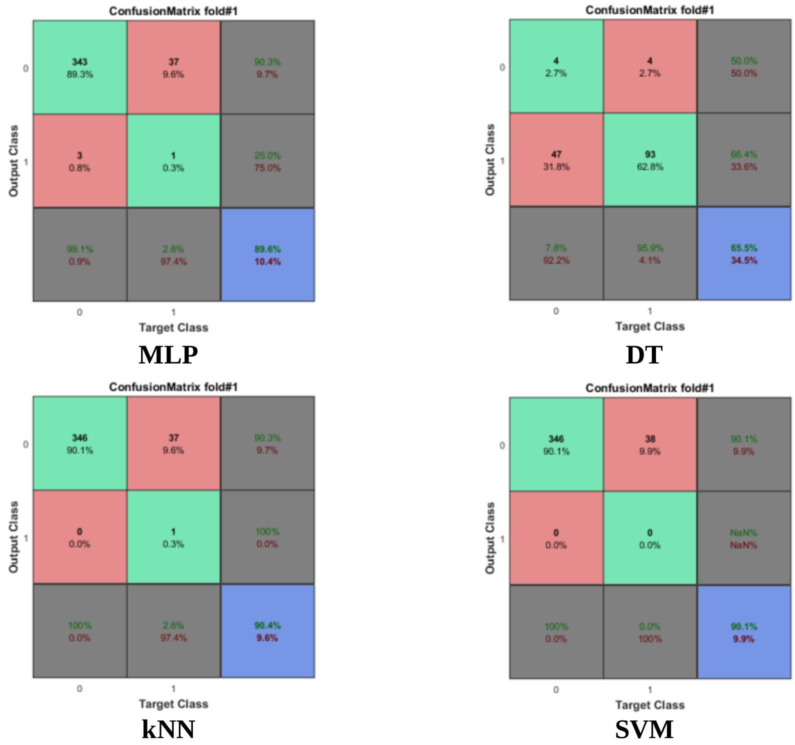

Fig. 3. Confusion matrices results over the original data (Fold #1)

The confusion matrices obtained over the original data have better values due to the imbalanced data. However, the balancing makes the accuracy of classification less as compared with the original data, though allows to change the ratio of classes samples, so the data are more balanced than the original one.

4. CONCLUSIONS

EEG signal of two healthy subjects was studied using the methods of machine learning, namely, decision trees (DT), multilayer perceptron (MLP), K-nearest neighbours (kNN), and support vector machines (SVM). Since the data were imbalanced, the appropriate balancing was performed by Kmeans clustering algorithm. The original and balanced dataset was classified by means of the mentioned above 4 methods. The obtained classification results were compared. It was found, that SVM showed the best result for the both datasets in terms of accuracy. This method gave 87.47% accuracy for the original data and 66.95% accuracy for the balanced data. MLP and kNN produce the comparable results which are almost the same. DT accuracies are the lowest for the given dataset, with 83.82% for the original data and 61.48% for the balanced data. The respective confusion matrices were also built.

In general, the methods of machine learning allow classifying the EEG signals and obtaining rather accurate results.

REFERENCES

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.

Amin, H. U., Mumtaz, W., Subhani, A. R., Saad, M. N. M., & Malik, A. S. (2017). Classification of EEG Signals Based on Pattern Recognition Approach. *Frontiers in Computational Neuroscience*, *11*(103), 1–12.

Bryant, R. A., & Sindicich, N. (2007). Hypnosis and Thought Suppression – More Data: A Brief Communication. *International Journal of Clinical and Experimental Hypnosis*, *56*(1), 37–46.

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dvey-Aharon, Z., Fogelson, N., Peled, A, & Intrator, N. (2015). Schizophrenia Detection and Classification by Advanced Analysis of EEG Recordings Using a Single Electrode Approach. *PLoS ONE*, *10*(4), 1–12.

Haykin, S. (Ed.). (2009). *Neural Networks and Learning Machines (3rd Edition).* New Jersey, Prentice Hall.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Lawhern, V., Hairston, W. D., McDowell, K., Westerfield, M., & Robbins, K. (2012). Detection and classification of subject-generated artifacts in EEG signals using autoregressive models. *Journal of Neuroscience Methods*, *208*(2), 181–189.

Li, J., Struzik, Z., Zhang, L., & Cichocki, A. (2015). Feature learning from incomplete EEG with denoising autoencoder. *Neurocomputing*, *165*, 23–31.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability – Volume 1: Statistics*, 281–297.

Parvinnia, E., Sabeti, M., Zolghadri Jahromi, M., & Boostani, R. (2014). Classification of EEG Signals using Adaptive Weighted Distance Nearest Neighbor Algorithm. *Journal of King Saud University – Computer and Information Sciences, 26*(1), 1–6.

Podgorelec, V. (2012). Analyzing EEG signals with machine learning for diagnosing Alzheimer’s disease. *Elektronika i Elektrotechnika*, *18*(8), 61–64.

Provençal, S. C., Bond, S., Rizkallah, E., & El-Baalbaki, G. (2018). Hypnosis for burn wound care pain and anxiety: A systematic review and meta-analysis. *Burns*, *44*(8), 1870–1881.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Real, R. G. L., & Kübler, A. (2014). Auditory oddball paradigm during hypnosis. *Institute of Psychology, University of Würzburg*.

Sanei, S., & Chambers, J. A. (Eds.). (2007). *EEG Signal processing*. Great Britain, Chippenham, John Wiley & Sons.

Satapathy, S. K., Jagadev, A. K., & Dehuri, S. (2017). Weighted majority voting based ensemble of classifiers using different machine learning techniques for classification of EEG signal to detect epileptic seizure. *Informatica*, *41*(1), 99–110.

Sun, L., Jin, B., Yang, B., Tong, J., Liu, C., & Xiong, H. (2019). Unsupervised EEG Feature Extraction Based on Echo State Network. *Information Sciences*, 475, 1–17.

Terhune, D. B., Cleeremans, A., Raz, A., & Lynn, S. J. (2017). Hypnosis and top-down regulation of consciousness. *Neuroscience and Biobehavioral Reviews*, *81*(A), 59–74.

Thilakvathi, B., Shenbaga, Devi, S., Bhanu, K., & Malaippan, M. (2017). EEG signal complexity analysis for schizophrenia during rest and mental activity. *Biomedical Research*, *28*(1): 1–9.

Wood, C., & Bioy, A. (2008). Hypnosis and Pain in Children. *Journal of Pain and Symptom Management*, *35*(4), 437–446.