Agnieszka Duraj[1], Łukasz Chomątek[1]

Institute of Information Technology, Lodz University of Technology(1)

doi:10.15199/48.2023.01.48

Outlier detection in EEG signals

Abstract. *In this paper, the topic of detection of outliers in EEG signals was discussed, which facilitates making decisions about the diagnosis of a patient based on this study. We used two methods to detect outliers: the support vector machine and the k nearest neighbors method. The experiments were performed on a publicly available dataset containing EEG test results for 500 patients. The obtained results showed that the methods we used allow for the outlier detection efficiency at the level of 93%*

Streszczenie. *W niniejszej pracy podjęto temat detekcji wyjątków w sygnałach EEG, co pozwala na ułatwienie podejmowania decyzji co do diagnozy pacjenta na podstawie tego badania. Do detekcji wyjątków wykorzystaliśmy dwie metody: maszynę wektorów nośnych i metodę k najbliższych sąsiadów. Eksperymenty zostały przeprowadzone na ogólnodostępnym zbiorze danych zawierającym wyniki badania EEG dla 500 pacjentów. Uzyskane wyniki pokazały, że użyte przez nas metody pozwalają na uzyskanie skuteczności detekcji wyjątków na poziomie 93%* ***(Wykrywanie wyjątków w sygnałach EEG)***

Keywords: outlier detection, EEG signal, SVM
Słowa kluczowe: wykrywanie wyjątków, EEG, maszyna wektorów nośnych

Introduction

Data mining as a part of the Knowledge Discovery in Databases (KDD) process is currently a very popular field of scientific research. The idea of this field is to quickly find regularities or irregularities in a data set. The authors are interested in detecting atypical features of the examined set, called outliers. The methods for detecting outliers are diverse. In scientific research, one can find outlier detection methods based on statistical methods, distance measures or the functions of similarity and dissimilarity.

A comprehensive overview of this field is given in the works [1, 3, 6]. Detecting outliers using innovative methods such as linguistic summaries and genetic algorithms was proposed in the works of Duraj and co-authors [7, 10, 11, 8, 9, 12, 4].

In this article, the authors deal with the detection of outliers in medical signals, in particular in the EEG signal. Electrocardiographic and electroencephalographic tests are one of the basic tests assessing human health. The first ECG examination assesses the condition of the heart and detects possible disorders. The second EEG assesses brain function leading to the recognition of not only emotions, but most of all it recognizes disease states, including epilepsy, Parkinson's disease, Alzheimer's disease, sleep apnea, insomnia and many others. In this study, EEG signals will be examined in order to detect atypical, exceptional changes that may indicate brain disorders in the examined patient. These can be epilepsy attacks, but also changes that precede these attacks. More on the classification of the EEG signal and the detection of emotions is given, for example, in [17, 5, 14].

The EEG signal is generated by impulses transmitted between neurons. This signal is recorded by electrodes placed on the head. These electrodes, arranged according to the 10-20 system, define the distances between adjacent electrodes. The arrangement of standard electrodes is shown in Fig.1.

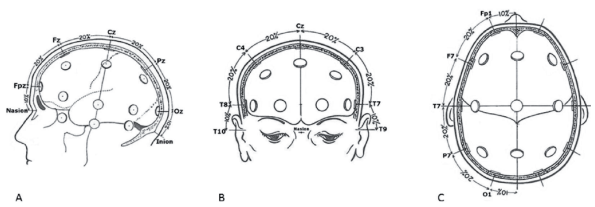

Fig. 1. An example of the arrangement of electrodes during the recording of the EEG signal, A: side view, B: front view, C: top view. [15]

The purpose of this article is to detect outliers in the EEG signal. In our case, such outliers will be symptoms of epilepsy. The data selected for analysis are medical EEG signaling data collected to identify seizure symptoms. Visualizations of sample data are shown in Fig. 2.

In order to detect outliers, the Support Vector Machine method and the k-Nearest Neighbors method were applied and compared. It is considered to define a new complex classifier supporting the EEG signal analysis by detecting outliers in brain dysfunction.

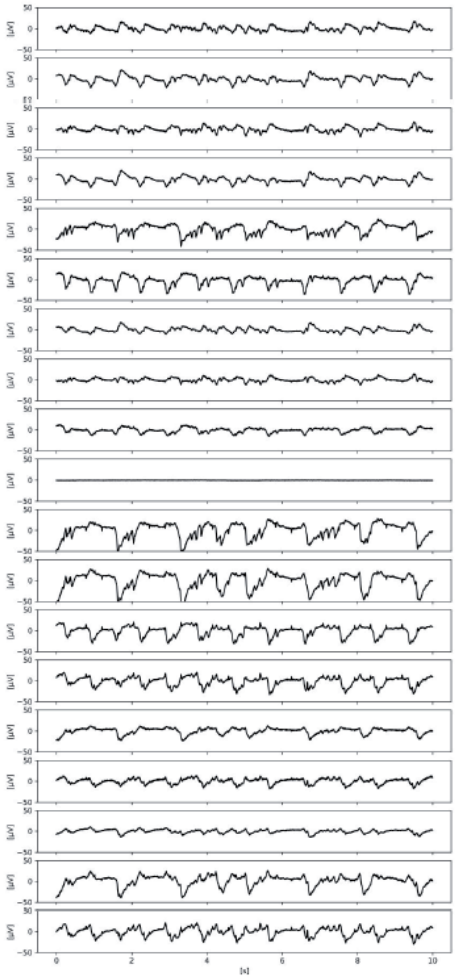

Fig. 2. EEG signal

The work is organised as follows:
Methods are described in sections Support Vector Machines, Linear SVM model and k-Nearest Neighbor Method. Then we present the dataset characteristics and the results of conducted experiments. The conclusions from our research are presented in the final section.

Support Vector Machines

The Support Vector Machines (SVM) method introduced by Vladimir N. Vapnik [16] creates decision spaces. These spaces are determined by dividing the entire space according to the created boundaries separating objects. In its simplest form, it divides a space into two subspaces - two classes, and separates them with a boundary line. An unknown object will be classified depending on which space it belongs to. This simplest case with two hyperplanes becomes intuitive and simple. Similar to the problem of linear regression. There are several types of support vectors, with different basis functions. These are, for example: linear, polynomial basis function, RBF (radial basis functions), sigmoid basis function, Gaussian basis function, etc. The author's research focused on the Gaussian basis function.

The SVM algorithm is often used to classify or predict data. It mainly consists in selecting the best of the considered discriminatory hyper-planes. Therefore, it is important to maximize the margin of separation between the two classes, while maintaining the smallest classification error. SVM is used in many different fields, from data analysis to speech, text or classification recognition, financial and medical data analysis. The main advantage of the support vector machine is the ability to process non-numeric data, data streams. In this case, an appropriate selection of the nucleus functions as well as the construction and finding of the hyperplane separating points belonging to two or more classes are important. The margin between the two datasets should be set to the maximum (largest value) [13, 16].

The description of how to create hyperplanes is easiest for a linear model. The non-linear model was not used by us in our research, therefore it is not discussed in this paper.

Linear SVM model

Let x_i be an input vector and y_i be a class label with possible values $\{-1, 1\}$. Let us consider the training set as a pair (x_i, y_i) for $i = 1, 2, ..., p$, $x_i \in R^d$. Let us assume that classes y_i are linearly separable. Then, function $g(x)$ defined by (1) will be a hyperplane separating the classes, where $w = [w_1, w_2, ..., w_N]^T$, $x = [x_1, x_2, ..., x_N]^T$.

$$g(x) = w^T x + b = 0 \tag{1}$$

If the assumptions (2) are met, the optimal hyperplane that maximizes the margin of separation can be written by the equation (3). In addition, the distance $dist$ of the selected x object from the optimal hyperplane is given by the equation (4).

$$\begin{cases} w^T x + b > 0 for y_i = 1 \\ w^T x + b < 0 for y_i = -1 \end{cases} \tag{2}$$

$$g(x) = w_0^T x + b_0 = 0 \tag{3}$$

$$dist(x) = \frac{g(x)}{||w_0||} \tag{4}$$

The graphic interpretation of the created hyperplanes by the support vector method is shown in Fig. 3.

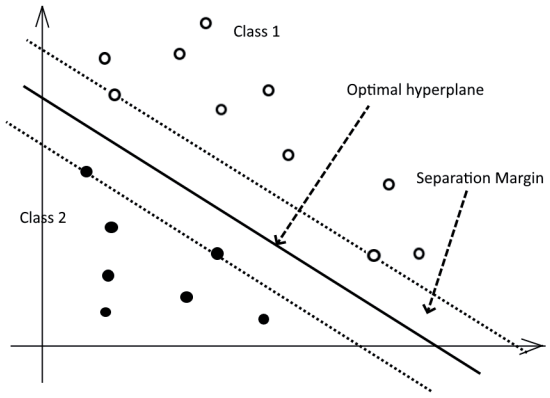

Fig. 3. Illustration of hyperplane in SVM

The point which is closest to the optimal hyperplane is a pair (x_i, y_i) for which $w^T x + b = 1$ for $y_i = 1$ and $y_i = -1$. Separation margin is obtained according to (5).

$$\rho = \frac{2}{||w_0||} \tag{5}$$

Solving the problem of maximizing ρ is tantamount to minimizing the Euclidean norm of the weight vector $||w||$ with the condition $\min \frac{1}{2} w^T w$ and the constraints $y_i(w^T x + b) \geq 1$.

Lagrange multipliers are used to solve the optimization problem. Then the primary Lagrange function L_P is minimized or the dual Lagrange function L_D is maximized, described respectively by the equations (6) and (7), where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_p]$ is a vector of Lagrange multipliers.

$$L_P(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{p} \alpha_i y_i (x_i \ldots w + b) + \sum_{i=1}^{p} \alpha_i \tag{6}$$

$$L_D = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{7}$$

The equation (7) for nonlinearly separable data can be written as the equation (8) where ξ is a non-negative complementary variable and ϕ is the weight selected by the user specifying the treatment testing errors in relation to the designated margin.

$$\min(\frac{1}{2} w^T w + \varphi \sum_{i=1}^{p} \xi_i) \tag{8}$$

For $\xi_i \geq 0$ we have a limitation given by (9).

$$y_i(w^T x + b) \geq 1 - \xi_i \tag{9}$$

The maximum upper estimate is defined as $\sum_{i=1}^{p} \xi_i$ is the limit, the maximum upper estimate.For L_D, constraint (9) becomes (10).

$$0 \leq \alpha_i \leq \varphi \tag{10}$$

Non-zero values of the Lagrange multipliers α_i with constraint functions equal to zero, marked as M_v, create for L_D of the dual task optimal weights of the hyperplane in the form (11).

$$qaw_0 = \sum_{i=1}^{M_v} \alpha_i y_i x_i \tag{11}$$

By subjecting the data to appropriate transformations, it is possible to use the support vector method for non-linearly separable patterns, this case is not discussed in this paper.

k-Nearest Neighbor Method

The k-nearest neighbors (kNN) method consists in finding k observations distributed in space, which are (will be) the closest (according to the selected metric) neighbors for the object in question x. The values are averaged and the x object is assigned to the most numerous class in the group. The distance is determined on the basis of a measure (metric). The most common measure for comparing quantitative features is the Euclidean distance measure defined by the equation (12).

$$d_e = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{12}$$

Other distance measures include, for example, the Minkowski metric defined by the equation (13) and its specific generalizations such as the city distance, called a taxi measure, or the Manhattan distance defined by the equation (14), Chebychev distance metric and many others.

$$d_{Minkowski} = \sqrt[m]{\sum_{i=1}^{n}|x_i - y_i|^m} \tag{13}$$

$$d_{Manhattan}(x, y) = \sum_{i=1}^{n}|(x_i - y_i)| \tag{14}$$

Two difficult situations can arise with these distances. First, as dimensionality increases, the difference between near and far points in a given space becomes invisible. Second, the value of Minkowski's distance contains the values dominated by the features with the largest span.

Dataset Characteristics

A set of EEG signals data taken from [2] was selected for the research.

This set contains 500 files with a record of brain activity for 23.6 seconds each. The time series are sampled at 4097 data points. Each data point is an EEG record value at a different point in time. Each 4097 data points were shuffled into 23 chunks containing 178 data points.

The last column of data contained classification labels $\{1, 2, 3, 4, 5\}$.

- 1 - outlier detection which is a registration of seizure activities;
- 2 - localized tumor;
- 3 - recording of a healthy area of the brain;
- 4 - eyes closed, while recording EEG signal the patient had eyes;
- 5 - eyes open, it means that the patient had his eyes open while recording the EEG signal of the brain

Experiments

Two loss functions were used for outlier detection with SVM: hinge loss and squared hinge. Changes in the number of epochs from 1,000 to 50,000 were made. The results are given in Table 1.

Table 1. Accuracy of outlier detection using SVM

number of epochs	loss function	accuracy
1000	hinge loss	0.846
1000	squared hinge	0.890
2000	hinge loss	0.934
2000	squared hinge	0.857
5000	hinge loss	0.931
5000	squared hinge	0.925
10000	hinge loss	0.842
10000	squared hinge	0.855

The results of the experiment show that both the type of loss function and the number of epochs have an impact on the accuracy of the classification. For a small number of epochs, the classification efficiency was 84-89%, depending on the choice of loss function. We observed the greater effectiveness at 1000 epochs for the squared hinge function. As the number of epochs increased, the classification efficiency increased to around 93%. At 5,000 epochs, we achieved better results for the hinge loss function.

The results of classifying outliers by the method of k Nearest Neighbors for $k = \{3, 5, 10, 15, 20, 25\}$ and the Euclidean, Manhattan, and Minkowski metrics are given in Tables 2, 3 and 4 respectively.

Table 2. Accuracy of outlier detection with k-NN, Euclidean metric

k	accuracy
3	0.625
5	0.736
10	0.748
15	0.699
20	0. 672
25	0.637

It can be noticed that the effectiveness of the kNN classifier with the Euclidean metric is very much influenced by the selection of the k coefficient. The results for $k = 3$ and $k = 10$ differ by 16%. As the value of k increases above 10, the classification accuracy decreases. This is due to the fact that subsequent points, often located at a considerable distance from the classified sample, negatively affect the correctness of its assignment.

Table 3. Accuracy of outlier detection with k-NN, Manhattan metric

k	accuracy
3	0.699
5	0.839
10	0.868
15	0.812
20	0.772
25	0.767

Similar conclusions can be drawn after analyzing the results obtained for the Manhattan metric. The highest classification accuracy was observed for $k = 10$, and the difference between the best and the weakest accuracy was 19%. The comparison of the results obtained for the Manhattan metric and the Euclidean metric shows that the first of these metrics allows for greater classification accuracy. The difference for $k = 10$ is about 14%.

Table 4. Accuracy of outlier detection with k-NN, Minkowski metric

k	accuracy
3	0.625
5	0.736
10	0.748
15	0.699
20	0. 672
25	0.637

The results collected in Table 4 were obtained for the kNN classifier with Minkowski's metric. As in the case of other metrics, the best results were obtained for k = 10. The use of Minkowski's metric does not give better results than those obtained for the Manhattan metric.

Conclusion

In this work, studies were carried out to assess the effectiveness of various methods of outlier detection in the EEG examination. For this problem, the SVM classifier and the kNN method were used. Both methods were run multiple times with different parameter sets. For the SVM method, the influence of the number of epochs and the loss function on the accuracy of outlier identification was examined. For the kNN method, the influence of the number of neighbors and the metric were examined.

Regarding the SVM method, the best results were obtained for 5000 epochs and for the hinge loss function. With this number of epochs, the classification accuracy for the squared hinge function differed by approximately 1% in favor of the hinge loss function. Therefore, it is worth noting that the appropriate selection of the number of epochs has a significant impact on the accuracy of the classification, because in the remaining cases we obtained worse results for the hinge loss function than for the squared hinge function.

By analyzing the results of the conducted experiments, it can be concluded that for the number of neighbors $k = 10$ of the k-NN algorithm, the best accuracy results were obtained for each type of distance measure used. Additionally. in all studies, the kNN classification using the Manhattan metric gave slightly better results than the classification using the Euclidean or Minkowski distance.

The comparison of the results obtained for the SVM method and the kNN method with the Manhattan metric shows that the SVM method allows for a greater accuracy of the classification. The difference for the best parameter sets is around 6.7% in favor of the SVM method.

Summing up, both tested methods allow to obtain high accuracy in the task of identifying outliers in the signals from EEG examinations, but better results can be obtained by the SVM method. For the kNN method, it's a good idea to use the Manhattan metric and not set the k parameter too high. The obtained results showed that the undertaken research direction was good. In the future, we plan to conduct research related to the nonlinear model for the SVM method.

***Authors**: PhD DSc Agnieszka Duraj, PhD Łukasz Chomątek Institute of Information Technology, Lodz University of Technology al. Politechniki 8, 93-590 Lodz, Poland email: agnieszka.duraj@p.lodz.pl : lukasz.chomatek@p.lodz.pl*

REFERENCES

[1] Aggarwal, C.C.: Outlier Analysis. Springer Science and Business Media (2013)
[2] Andrzejak, R.G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.E.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E **64**(6), 061907 (2001)
[3] Barnett, V., Lewis, T.: Outliers in statistical data, vol. 3. Wiley New York (1994)
[4] Chomatek, L., Duraj, A.: Multiobjective genetic algorithm for outliers detection. In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on. pp. 379–384. IEEE (2017)
[5] Dura, A., Wosiak, A., Stasiak, B., Wojciechowski, A., Rogowski, J.: Reversed correlation-based pairwised eeg channel selection in emotional state recognition. In: International Conference on Computational Science. pp. 528–541. Springer (2021)
[6] Duraj, A.: Metody analizy danych z detekcją wyjątków. Akademicka Oficyna Wydawnicza EXIT (2019)
[7] Duraj, A.: Outlier detection in medical data using linguistic summaries. In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on. pp. 385–390. IEEE (2017)
[8] Duraj, A., Chomatek, L.: Outlier detection using the multiobjective genetic algorithm. Journal of Applied Computer Science **25**(1), 29–42 (2017)
[9] Duraj, A., Chomatek, L.: Supporting breast cancer diagnosis with multi-objective genetic algorithm for outlier detection. In: International Conference on Diagnostics of Processes and Systems. pp. 304–315. Springer (2017)
[10] Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Outlier detection using linguistically quantified statements. International Journal of Intelligent Systems **33**(9), 1858–1868 (2018)
[11] Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Detection of outlier information by the use of linguistic summaries based on classic and interval-valued fuzzy sets. International Journal of Intelligent Systems **34**(3), 415–438 (2019)
[12] Duraj, A., Szczepaniak, P.S., Chomatek, L.: Intelligent detection of information outliers using linguistic summaries with non-monotonic quantifiers. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 787–799. Springer (2020)
[13] Kumar, M.A., Gopal, M.: A hybrid svm based decision tree. Pattern Recognition **43**(12), 3977–3987 (2010)
[14] Opałka, S., Stasiak, B., Szajerman, D., Wojciechowski, A.: Multi-channel convolutional neural networks architecture feeding for effective eeg mental tasks classification. Sensors **18**(10), 3451 (2018)
[15] Seeck, M., Koessler, L., Bast, T., Leijten, F., Michel, C., Baumgartner, C., He, B., Beniczky, S.: The standardized eeg electrode array of the ifcn. Clinical neurophysiology **128**(10), 2070–2077 (2017)
[16] Vapnik, V.: The nature of statistical learning theory. Springer science & business media (1999)
[17] Wosiak, A., Dura, A.: Hybrid method of automated eeg signals' selection using reversed correlation algorithm for improved classification of emotions. Sensors **20**(24), 7083 (2020)