One-Class SVM for outliers detection in epileptic EEG

Matvey Khoymov[1], Vadim Grubov[1,2], Vladimir Maksimenko[1,2],
Nikita Utashev[3], Denis Andrikov[4], Semen Kurkin[1,2]
[1]Baltic Center for Artificial Intelligence and Neurotechnology,
Immanuel Kant Baltic Federal University, Kaliningrad, Russia
[2]Neuroscience and Cognitive Technology Laboratory, Innopolis University, Kazan, Russia
[3]National Medical and Surgical Center named after N. I. Pirogov,
Ministry of Healthcare of the Russian Federation, Moscow, Russia
[4]Research and Production Company "Immersmed", Moscow, Russia

***Abstract*—In this paper possibility of machine learning methods applications for epileptic activity detection in records of electroencephalography is studied for further application in medical decision-support system.**

***Keywords*—epilepsy, EEG, detection, machine learning.**

I. Introduction

Epilepsy is one of the most common neurological disease: about 1% of the world's population suffers from this disease. The epilepsy causes seizures that happens due to excessive hypersynchronous neuron activity of the brain and they are accompanied by uncontrollable convulsions with loss of consciousness [1].

One of the approaches to diagnose the epilepsy is based on consideration of the electroencephalography record of the subject [2]. To do this, you need to make a long record of the patient's EEG (sometimes record can exceed 24 hours), after that, the specialist manually reviews the obtained data and looks for epileptiform activity in it [3]. It is obvious that the human needs to spend a lot of time and effort to review 24 hours long records, and sometimes more, to find seizure lasting 1-2 minutes. At the same time, the human factor cannot be canceled, due to a seizure can be overlooked an incorrect diagnosis can be made. Moreover, epileptic discharges are formed randomly and are difficult to predict [4], [5], making it difficult to automatically mark up records of brain electrical activity [6].

To help medicine employee machine learning methods can be applicated, they would be trained to detect and indicate epileptiform activity on EEG record of the patient, and these marks would be checked by specialist. Such an approach should at least reduce the burden on specialists involved in the diagnosis of epilepsy.

II. Materials and methods

A. Machine learning methods

In epilepsy diagnostic it all comes down to the problem of binary classification [7]. There are normal data (class 0) and seizure data (class 1). However, there is perceptible problem of the imbalance in classes [8]. During 24 hours the patient may experience 1–3 seizures with long about 1.5 minutes. This means that the share of class 1 data is 0.32% at best. For this reason, classical models can show poor results, as they are designed to be trained on data with an almost equal distribution of sample objects between classes.

Machine learning methods are divided in three main groups: supervised learning, unsupervised learning and semi-supervised learning [9]. Supervised and unsupervised learning methods are suitable for solving the problem under consideration. Supervised learning requires ready-made data markup, thanks to which the model establishes a relationship between the features of a sample and its class [10]. Although such models have high accuracy, but they can be overfitted, thereby they demonstrate poor results when get new data. EEG is highly variable, so risk of overfitting is very high. The model that is trained by EEG data of one person and shows perfect results can fail to detect epilepsy in EEG of another patient [11]. Such approach is based on representation of epileptic seizures as the extreme events [12], [13]. Therefore, the optimal choice to address the problem is unsupervised machine learning methods [14]. When using these methods, it is not necessary to mark up the data in advance, since the data is clustered according to its features and each cluster is assigned a class. And, since the epileptic data is much smaller than normal, this problem may be related to the problem of anomaly detection [15].

B. Data preprocessing

Database for models training was provided by the National Medical and Surgical Center named after N. I. Pirogov of the Russian Healthcare Ministry. All patients gave their consent to participate in the experiment. The recording was made during the daily activities of patients with purpose to establish epileptic symptoms and ways of further treatment. The recording duration varied from 8 to 84 hours according to patient conditions and number of detected seizures needed to confirm the diagnosis. None of the seizures were deliberately induced, they all occurred spontaneously. The data was pre-labeled by a specialist. Database contains EEG data of 80 patients with diagnosed focal epilepsy at all.

Data recording was carried out from 25 channels arranged according to the “10-20” technique. Only waves with a frequency of 1-30 Hz were considered, since they are recognized as more demonstrative for the epilepsy detection on the EEG. The epilepsy seizures are well displayed in the form of emissions on the energy of continuous wavelet transform (CWT) [16], [17]. In this work CWT was employed with averaged frequency range of 1–30 Hz and 60-second time intervals [11].

III. Results

The code was written in the Python programming language using the Scikit-learn library. One-class support vector machine was chosen as the first machine learning method [18], since it has already showed itself well in biological data analysis [19].

The support vector machine (SVM) takes the data to a higher dimensional space and then finds a hyperplane that separates the vectors into two classes. One-class SVM learns to separate outliers from normal data. It is worth noting that an individual model was trained for each patient, since the EEG is very variable.

The main task was to select such model parameters in order to obtain the best result. The effectiveness of the model was estimated according to two criteria: recall (TPR) and precision (PPV):

$$TPR = \frac{TP}{TP + FN} * 100, \quad (1)$$

$$PPV = \frac{TP}{TP + FP} * 100, \quad (2)$$

where TP – amount of the correctly detected seizures; FP – amount of the falsely detected seizures; FN – amount of the missed seizures.

The parameters were iterated manually with subsequent visualization of the results (further mean values of the group of models are presented). First, the optimal *kernel* parameter was being selecting, which was used in the algorithm. There are 4 standard *kernel* types at all: “linear”, “rbf”, “sigmoid” and “poly”. *Kernel* “linear” does not work well with large amounts of data, so it was excluded from the selection. The “rbf” *kernel* showed the best result of three remaining *kernels* (Table I).

TABLE I
Mean values of models results when changing the kernel.

kernel	TPR, %	PPV, %
”rbf”	88.25	2.10
”poly”	26.75	2.13
”sigmoid”	100.00	0.89

The remaining parameters: *nu*, *gamma*, *tol*, *degree*, *coef0*. The last two parameters only work with “sigmoid” and “poly” *kernels*, so they were not considered. Parameter *nu* is a value of bound that indicates the share of outliers in the data. *Gamma* – coefficient of the kernel, that determines the degree of fit by the hyperplane of the vectors. *Tol* – stopping criterion: algorithm stops learning when new value of loss exceeds last loss value minus *tol*.

The heatmaps (Fig. 1) clearly shows that the model with parameters *gamma* – “scale” has the best effectiveness. Also, they demonstrate that optimal range of *tol* values is from 0.0001 to 0.00001.

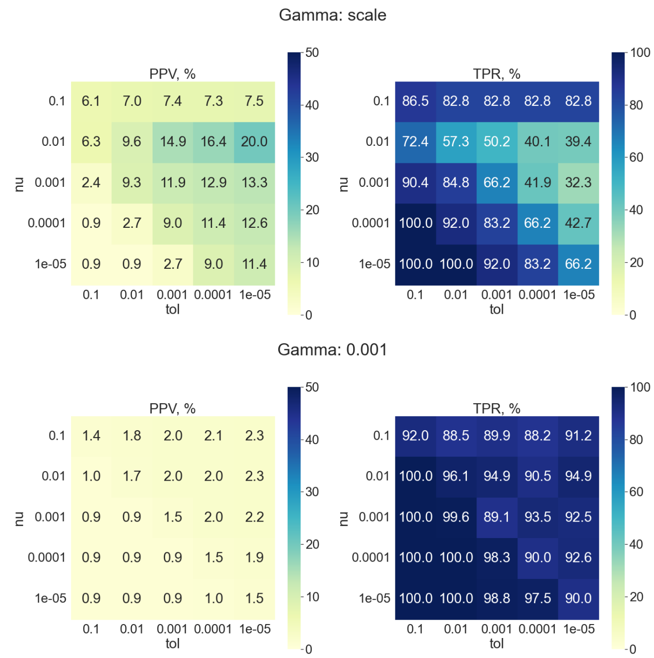

Fig. 1. Dependence of PPV and TPR values on parameters *gamma*, *nu* and *tol* (averaged values of all models).

After that, the models were trained with the following parameters: *kernel* – “rbf”, *gamma* – “scale”, *nu* – $[10^{-i}$, $i \in [-1, -5]]$, *tol* – [0.0001, 0.00001]. After training, raw scores of the samples were extracted, the value corresponding to the percentile from 0.00001 to 50 was chosen and the data was reclassified: if the sample score of was higher than the calculated value, then the sample was designated as an outlier. This parameter was named *threshold*. The optimal value of *threshold* turned out to be the range from 1 to 0.5 (Fig. 2).

IV. Conclusion

The best result (TPR: 51.58%, PPV: 17.13%) was showed by model with parameters as follows: *kernel* – “rbf”, *gamma* – “scale”, *nu* – 0.1, *tol* – 0.0001.

Although the considered one-class SVM models cannot fully automate the process of diagnosing epilepsy, they can be used to narrow the search area for epileptic seizures on the subject’s EEG, which will facilitate the work of specialists.

In future work, it is planned to consider other unsupervised machine learning methods for anomaly detection, such as: k-Nearest Neighbors, IsolatedForest, etc.

ACKNOWLEDGEMENTS

This work was supported by the President Program, grant MD-590.2022.1.2 and grant NSH-589.2022.1.2.

REFERENCES

[1] Carl E Stafstrom and Lionel Carmant. Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harbor perspectives in medicine*, 5(6):a022426, 2015.

[2] Lauren Harris and Heather Angus-Leppan. Epilepsy: diagnosis, classification and management. *Medicine*, 48(8):522–528, 2020.

[3] Shelagh JM Smith. Eeg in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005.

[4] Evgenia Sitnikova, Alexander E Hramov, Vadim V Grubov, Alexey A Ovchinnkov, and Alexey A Koronovsky. On–off intermittency of thalamo-cortical oscillations in the electroencephalogram of rats with genetic predisposition to absence epilepsy. *Brain research*, 1436:147–156, 2012.

[5] Alexey A Koronovskii, Alexander E Hramov, Vadim V Grubov, Olga I Moskalenko, Evgenia Sitnikova, and Alexey N Pavlov. Coexistence of intermittencies in the neuronal network of the epileptic brain. *Physical Review E*, 93(3):032220, 2016.

[6] Gilles van Luijtelaar, Annika Lüttjohann, Vladimir V Makarov, Vladimir A Maksimenko, Alexei A Koronovskii, and Alexander E Hramov. Methods of automated absence seizure detection, interference by stimulation, and possibilities for prediction in genetic absence models. *Journal of neuroscience methods*, 260:144–158, 2016.

[7] Roshan Kumari and Saurabh Kr Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7), 2017.

[8] Richmond Addo Danquah. Handling imbalanced data: A case study for binary class problems. *arXiv preprint arXiv:2010.04326*, 2020.

[9] Shagan Sah. Machine learning: a review of learning types. 2020.

[10] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer, 2008.

[11] Oleg E Karpov, Vadim V Grubov, Vladimir A Maksimenko, Semen A Kurkin, Nikita M Smirnov, Nikita P Utyashev, Denis A Andrikov, Natalia N Shusharina, and Alexander E Hramov. Extreme value theory inspires explainable machine learning approach for seizure detection. *Scientific Reports*, 12(1):1–14, 2022.

[12] Nikita S Frolov, Vadim V Grubov, Vladimir A Maksimenko, Annika Lüttjohann, Vladimir V Makarov, Alexey N Pavlov, Evgenia Sitnikova, Alexander N Pisarchik, Jürgen Kurths, and Alexander E Hramov. Statistical properties and predictability of extreme epileptic events. *Scientific reports*, 9(1):1–8, 2019.

[13] Oleg E Karpov, Vadim V Grubov, Vladimir A Maksimenko, Nikita Utaschev, Viachaslav E Semerikov, Denis A Andrikov, and Alexander E Hramov. Noise amplification precedes extreme epileptic events on human eeg. *Physical Review E*, 103(2):022310, 2021.

[14] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[15] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[16] Evgenia Sitnikova, Alexander E Hramov, Alexey A Koronovsky, and Gilles van Luijtelaar. Sleep spindles and spike–wave discharges in eeg: their generic features, similarities and distinctions disclosed with fourier transform and continuous wavelet analysis. *Journal of neuroscience methods*, 180(2):304–316, 2009.

[17] Alexander E Hramov, Alexey A Koronovskii, Valeri A Makarov, Alexey N Pavlov, and Evgenia Sitnikova. *Wavelets in neuroscience*. Springer, 2015.

[18] Evgeny Burnaev and Dmitry Smolyakov. One-class svm with privileged information and its application to malware detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 273–280. IEEE, 2016.

[19] J. Zhou, K.L. Chan, V.F.H. Chong, and S.M. Krishnan. Extraction of brain tumor from mr images using one-class support vector machine. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6411–6414, 2005.

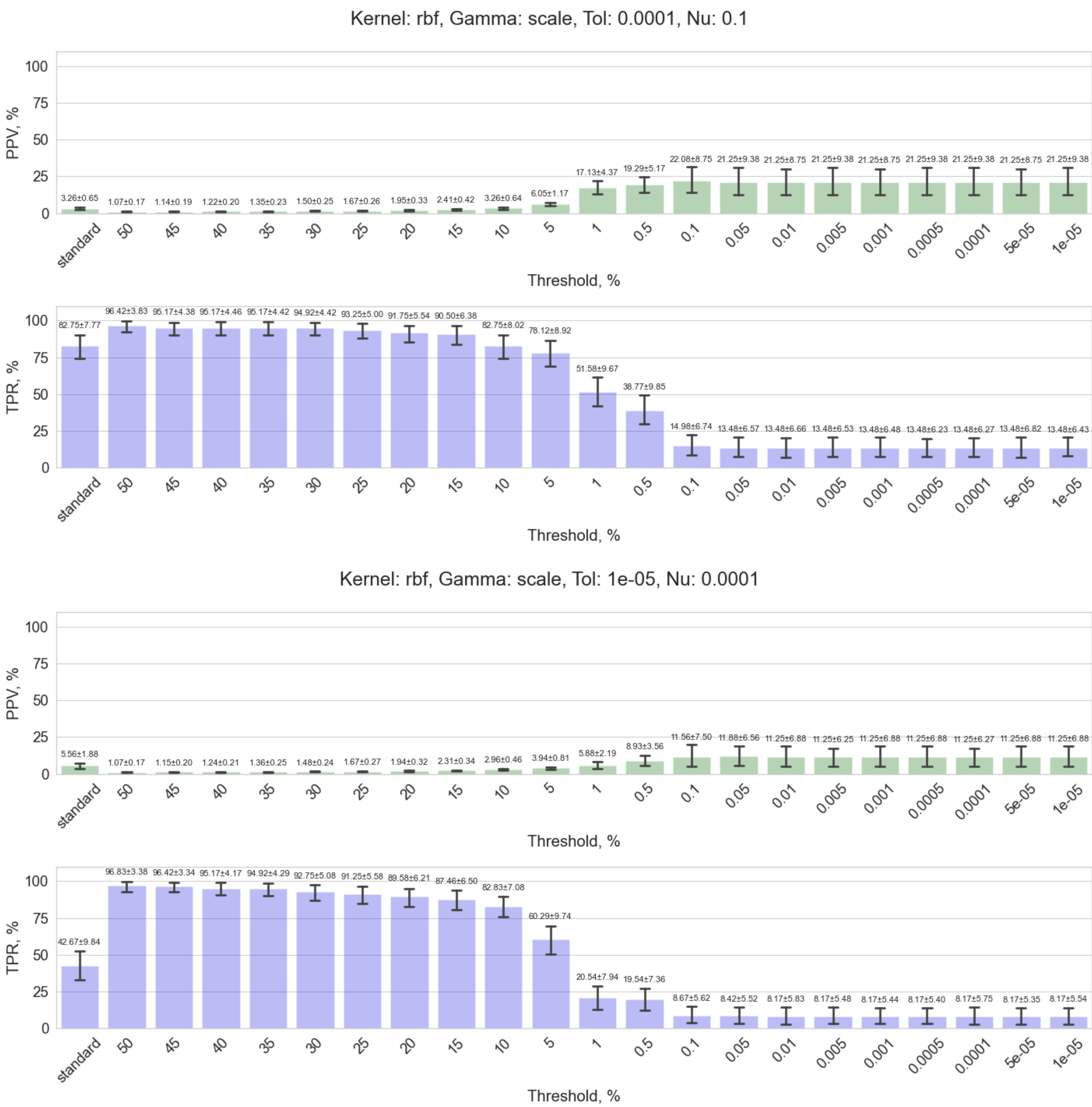

Fig. 2. Dependence of PPV and TPR values on parameters *nu*, *tol* and *threshold* (averaged values of all models).