The application of machine learning methods in searching for statistical patterns for diagnosing obsessive-compulsive disorder

© V.A. Yunusov, S.A. Demin

Kazan Federal University, Kazan, Tatarstan, Russia
E-mail: valentin.yunusov@gmail.com

Received May 5, 2023
Revised June 29, 2023
Accepted October, 30, 2023

One of the urgent tasks of modern data sciences is defining diagnostic criteria for mental disorders. This task is complicated by the existence of many biophysical parameters, some of which may be redundant. In this paper, we apply techniques for selecting features necessary to diagnose the obsessive-compulsive disorder. With the help of machine learning methods, the classification problem was solved for the initial set of features at the first stage of work; at the second stage, subsets of the most significant diagnostic features were selected for volunteers exhibiting significant symptoms of this disorder as well as for representatives of the reference group.

Keywords: living systems, obsessive-compulsive disorder, biomedical data, machine learning methods, feature attribution.

DOI: 10.61011/TPL.2023.12.57570.13A

Defining diagnostic criteria for pathological changes in the human brain functioning, for instance, in neurological diseases and mental disorders, is an important problem of modern data sciences and biophysics. One of the common mental disorders is obsessive-compulsive disorder (OCD). OCD is characterized by the presence of obsessions and compulsions. Obsessions are obsessive, repetitive and horrid thoughts, as well as urges that cause anxiety. Compulsions are repetitive behaviors or mental rituals intended to relieve stresses caused by obsessions.

To determine the diagnostic criteria for this disease, methods of statistical analysis of electroencephalogram (EEG) and/or magnetoencephalogram (MEG) signals and the like are widely used [1,2]. Recording of a large number of experimental data on the human brain functional activity promoted intense development of machine learning methods for solving neurophysiological and biophysical problems [3–5]. The machine learning methods make it possible to reveal hidden patterns, automate and speed up the processes of feature classification and selection in raw biomedical data. To solve such problems, software packages and libraries in various programming languages are being developed.

In this study we used the Weka code designed for data preprocessing and analysis by using, among others, machine learning methods [6]. In selecting significant features, this code uses a combination of search methods and means for assessing the significance of attributes (features). The search method is used in the feature space to find a suitable feature subset. A feature estimator is a method by which each feature is evaluated in the context of the target variable.

In this study we use methods for selecting a subset of attributes, namely, CfsSubsetEval and CorrelationAttributeEval. Method CfsSubsetEval evaluates the significance of a subset of attributes by considering individual predictive ability of each feature, as well as the degree of its redundancy. As a result, there gets obtained only a subset of features strongly correlated with the target class [7]. In order to obtain a meaningful subset of characteristics, CorrelationAttributeEval estimates the Pearson correlation coefficient relative to the target class for each variable [7].

Experimental data were obtained earlier as a result of international collaboration with Goldsmiths College, University of London. The data files represented EEG signals for two groups of subjects: 15 test subjects exhibiting significant symptoms of obsessive-compulsive disorder, and 15 people demonstrating these symptoms only slightly (a conditionally reference group). In addition, all the subjects were assessed according to the OCI-R (Obsessive-Compulsive Inventory-Revised questionnaire) that is a self-administered questionnaire assessing OCD scores across six symptom domains (given in short below: for example, hand washing that is exaggerated fear of contamination, etc.: washing, checking, ordering, obsessing, hoarding and mental neutralizing [8]. EEGs were recorded under three conditions: reading phase, visualization phase and suppression phase. In the second phase, the study participants visualized the said sentence during 1 min. In the last phase, the subjects had to think for 1 min about anything other than the described event. Bioelectrical activity from different areas of the cerebral cortex was recorded with electrodes arranged in accordance with the extended International Placement Scheme „10−20%“ [9].

Our research was performed in two stages. At the first stage, a set of statistical indicators was calculated for each EEG record: Hjort parameters (activity, complexity and mobility), power of α-, β-, θ- and δ-activities of the cerebral cortex, detrended fluctuation analysis (DFA), Higuchi fractal dimension, Lempel−Ziv complexity, Petrosian fractal dimension, and sample entropy. Using five

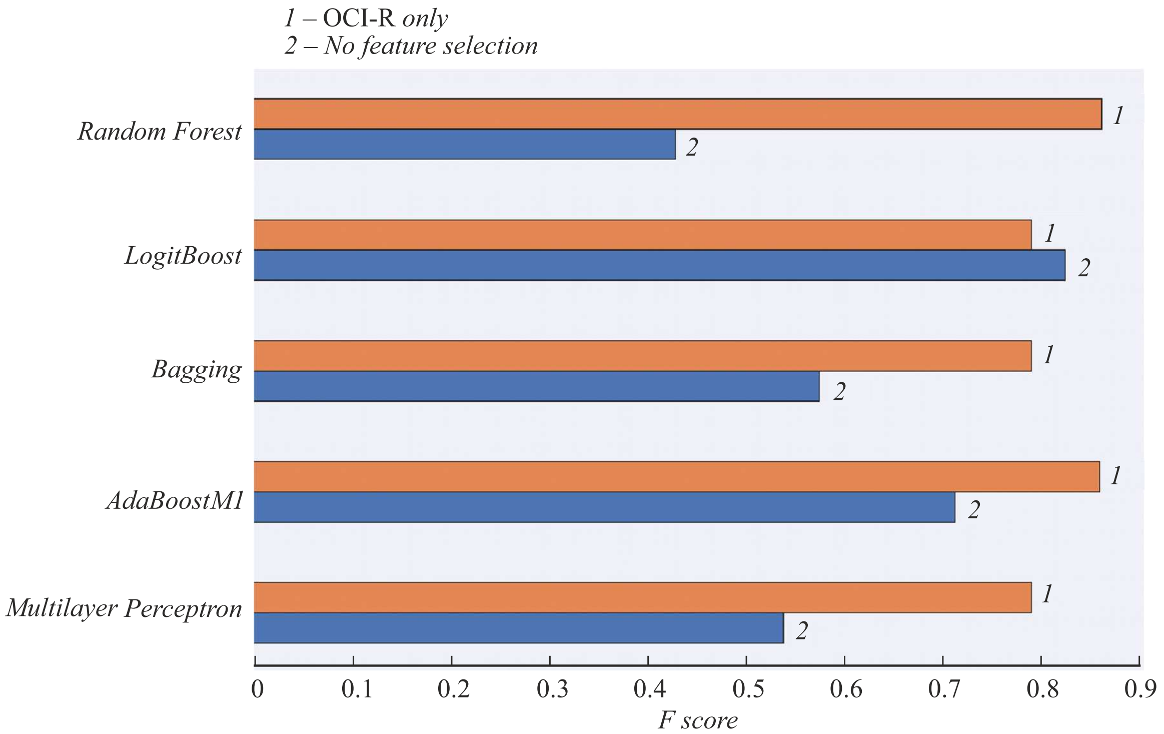

Figure 1. The F-measure of classifiers free of using feature attribution methods for a full set of statistical parameters and a subset of OCI-R factors.

machine learning methods realized in the Weka code, for the obtained parameters there was solved the problem of classifying the subjects′ EEG records into groups with low and high levels of exhibited OCD symptoms. Efficiencies of the machine learning methods were compared for dividing the subjects with different OCD manifestations into two groups. The accuracy of classification was assessed using the F-measure (the harmonic mean between the accuracy and completeness of the classifier) and AUC ROC that is a parameter describing the relationship between the model sensitivity (the proportion of true positive examples) and its specificity (described in terms of the proportion of false-positive results). We have found out that, for most methods, the F-measure and AUC ROC possess high values only when OCI-R parameters are taken into account (Fig. 1). The maximal F-measure (0.856) and AUC ROC (0.946) were achieved using the Random Forest method. Due to the small size of the data set under consideration, the sample was divided using stratified cross-validation with division into five blocks, which emulates the presence of a test sample.

At the second stage, the characteristics that made the most significant contribution to the classification were determined using the CorrelationAttributeEval and CfsSubsetEval feature selection methods. The selection of subsets of significant features was performed to possibly improve the accuracy of the classifiers, since the redundant (or noise) variables could distort the value to be predicted.

In the case of the CorrelationAttributeEval method, the subset of parameters included those of them for which the Pearson correlation coefficient calculated for the target class was quite high (Table 1). As for the CfsSubsetEval method, the previously mentioned mechanisms for assessing the significance of attributes were taken into account and, hence, the required subset included the Hjort complexity for the O_2 electrode, δ-activity for the PO_8 electrode, θ-activity for the O_z electrode, β-activity for the CP_3 and AF_8 electrodes, and OCI-R indicators: checking, ordering, mental neutralizing and obsession.

Table 1. Target-class features with the highest Pearson correlation coefficients (CorrelationAttributeEval estimator)

Characteristic	Pearson coefficient
Checking	0.733
Obsession	0.685
Ordering	0.628
δ-activity for the PO_8 electrode	0.585
β-activity for the O_2 electrode	0.541
Mental neutralizing	0.536
Washing	0.533
DFA for the O_z electrode	0.504

Upon the completion of feature selection, the obtained subsets were reclassified (Fig. 2). A comparison of the F metric and AUC ROC metric of classifiers with and without feature selection showed that for some methods the accuracy increases, while for others it decreases (Tables 2, 3). In general, the F-measure and AUC ROC

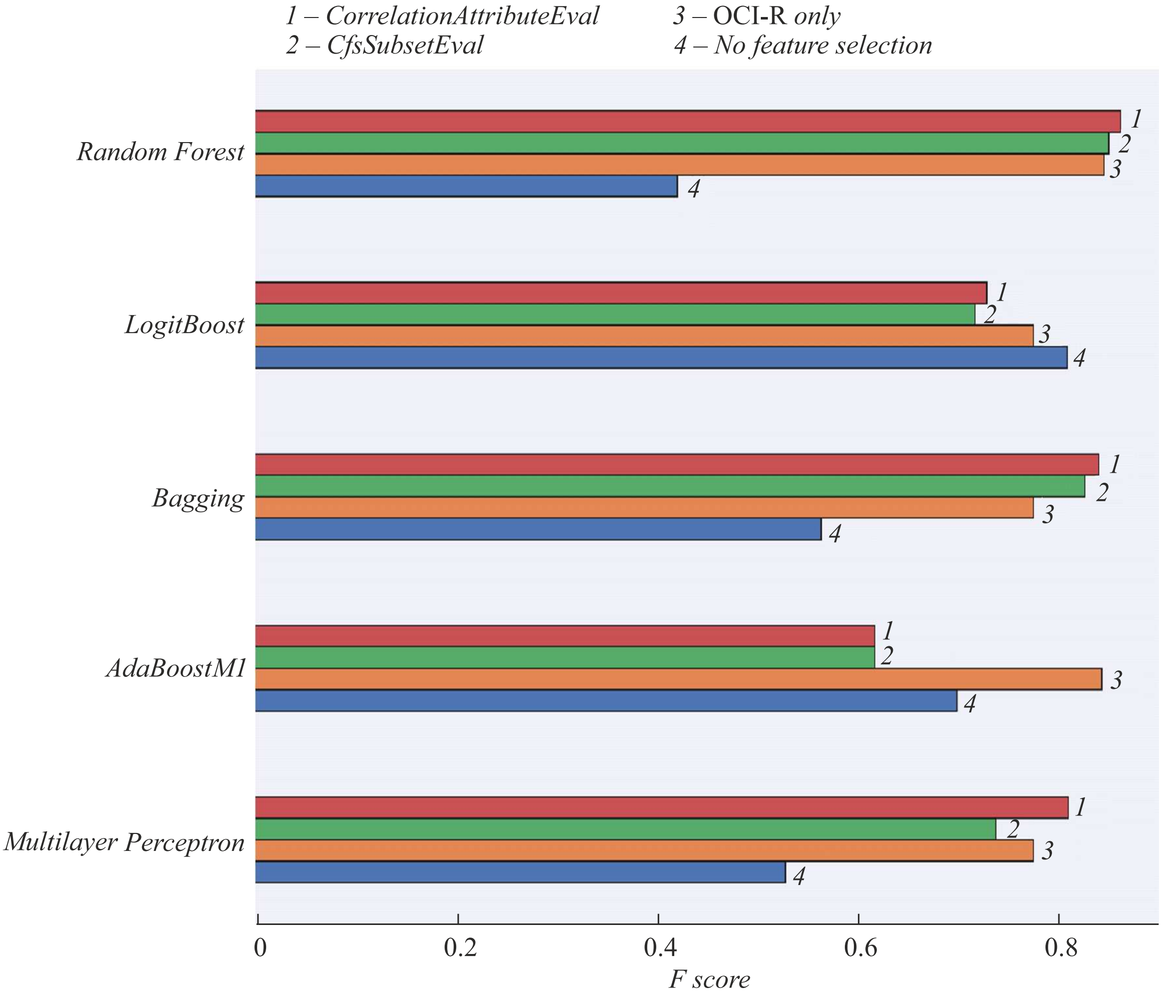

Figure 2. The F-measure of classifiers involving different feature selection methods.

Table 2. The classifierś F-measures for different feature selection methods

Classifier	F-measure			
	Without feature selection	Only OCI-R characteristics	CorrelationAttributeEval	CfsSubsetEval
Random Forest	0.426	0.856	0.873	0.861
LogitBoost	0.819	0.785	0.738	0.726
Bagging	0.571	0.785	0.851	0.837
AdaBoostM1	0.708	0.854	0.625	0.625
Multilayer Perceptron	0.535	0.785	0.82	0.747

metric without using feature selection methods for the full set of statistical parameters are lower than for subsets of selectable features. In the Random Forest method, the classifierś F-measure and AUC ROC measure are maximal for all the subsets; they reach the highest values of 0.873 and 0.969, respectively, when the CorrelationAttributeEval attribution method is used. Notice that the Random Forest method accuracy increases with decreasing number of parameters. However, in this case it is significant that, when using feature selection methods, the accuracy of the method has increased not only relative to that of the full set of characteristics, but, in addition, relative to the set of OCI-R characteristics containing a smaller number of features.

In this study, we applied machine learning methods jointly with the feature selection methods included in the

Table 3. The classifierś AUC ROC-metrics for different feature selection methods

Classifier	AUC ROC			
	Without feature selection	Only OCI-R characteristics	CorrelationAttributeEval	CfsSubsetEval
Random Forest	0.617	0.946	0.969	0.935
LogitBoost	0.811	0.809	0.936	0.8
Bagging	0.648	0.924	0.926	0.909
AdaBoostM1	0.77	0.688	0.849	0.688
Multilayer Perceptron	0.577	0.883	0.898	0.803

Weka code. We have calculated a set of statistical features for the EEG signals from people with OCD and reference panel members. The task of feature classification, as well as of selecting the most significant features for two groups of subjects, has been fulfilled.

The full initial space of characteristics was redundant for the problem under study. It was found out that indicators defined in the OCI-R method contribute significantly to the performance of the machine learning models. In addition, statistical parameters of bioelectric signals in the cerebral cortex occipital lobe also affect significantly the feature selection process. The best result was achieved using the CorrelationAttributeEval feature selection method and Random Forest classifier, whose value of the F metric was 0.873 while the AUC ROC metric value was 0.969.

Further verification of the obtained results implies involvement of a larger number of volunteers exhibiting different levels of OCD symptoms, as well as expansion of questionnaires, meters and scales for the OCD identification and monitoring. In further studies, an increase in the data set under consideration will allow applying the Random Mixing method in order to make the analysis more objective. The use of machine learning methods supported by feature selection methods in the process of statistic processing of signals of the human brain bioelectrical activity will promote an automated search for diagnostic criteria for psychiatric disorders, neurodegenerative and neurological diseases [10], and also an increase in the diagnostics accuracy and speed.

Funding

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program („PRIORITY–2030“).

Compliance with Ethical Standards

All the works included in the research involving human participants were carried out in accordance to the ethical standards of the national Scientific Ethics Committee and the 1964 Declaration of Helsinki with its subsequent supplements, as well as to similar ethical standards. The Informed Voluntary Consent was obtained from each participant of the study.

Conflict of interests

The authors declare that they have no conflict of interests.

References

[1] S.A. Demin, R.M. Yulmetyev, O.Yu. Panischev, P. Hänggi, PhysicA.A, **387** (8-9), 2100 (2008). DOI: 10.1016/j.physa.2007.12.003

[2] V.A. Yunusov, S.A. Demin, O.Yu. Panischev, N.Y. Demina, J. Phys.: Conf. Ser., **2103** (1), 012044 (2022). DOI: 10.1088/1742-6596/2103/1/012044

[3] K. Hilbert, T. Jacobi, S.L. Kunas, B. Elsner, B. Reuter, U. Leuken, N. Kathmann, Psychother. Res., **31** (1), 52 (2021). DOI: 10.1080/10503307.2020.1839140

[4] F. Ferreri, A. Bourla, C.S. Peretti, T. Segawa, N. Jaafari, S. Mouchabac, J. Med. Internet Res., **6** (12), e11643 (2019). DOI: 10.2196/11643

[5] M. Hoexter, E. Miguel, J. Diniz, R. Shavitt, G. Busatto, J. Sato, J. Affect. Disord., **150** (3), 1213 (2013). DOI: 10.1016/j.jad.2013.05.041

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, ACM SIGKDD Explor. Newsl., **11** (1), 10 (2008). DOI: 10.1145/1656274.1656278

[7] M. Hall, *Correlation-based feature subset selection for machine learning*, Ph.D. thesis (University of Waikato, Hamilton, New Zealand, 1999).

[8] E.B. Foa, J.D. Huppert, S. Leiberg, R. Langner, R. Kichic, G. Hajcak, P.M. Salkovskis, Psychol. Assess., **14** (4), 485 (2002). DOI: 10.1037/1040-3590.14.4.485

[9] R. Jones, J. Bhattacharya, J. Behav. Addict., **1** (3), 96 (2012). DOI: 10.1556/JBA.1.2012.005

[10] S.A. Demin, O.Yu. Panischev, S.F. Timashev, R.R. Latypov, Bull. Russ. Acad. Sci. Phys., **84** (11), 1349 (2020). DOI: 10.3103/S1062873820110088.

Translated by Ego Translating