The feature descriptions generating method based on distances to standards in biomedical research

Senko Oleg[1,2] senkoov@mail.ru
Salmanov Mahir[1]⋆ sy.mahir@gmail.com
Brusow Oleg[3] oleg.brusow@yandex.ru
Matveev Ivan[2] matveev@ccas.ru
Kuznetsova Anna[4] azfor@yandex.ru

[1]Moscow, M.V.Lomonosov Moscow State University
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[3]Moscow, Federal State Budgetary scientific institution "Scientific Center for Mental Health" RAS
[4]Moscow, N.M. Emanuel Institute of Biochemical Physics of the Russian Academy of Sciences

The existing machine learning techniques are largely based on the principle of compactness, that assumes proximity of descriptions of objects with similar Y. An approach based on the use of metrics that are given on a set of feature descriptions can be distinguished among the various ways of implementation compactness principle. As example k-nearest neighbors and support vector machines algorithms can be given. In the first case, the prediction Y at the point $\boldsymbol{x}$ is calculated from the values of Y on the k objects of the training sample closest to the point $\boldsymbol{x}$. In the most popular version of the support vector machine, the forecast of the value of Y is calculated from a linear combination of kernel functions, each of which depends on the distance between $\boldsymbol{x}$ and one of the support objects of the training sample. The high efficiency of machine learning methods based on the use of distance to reference objects in pattern recognition problems has been confirmed by numerous experiments. It should also be noted that methods based on assessing the general similarity of the object under consideration with previously encountered cases are often used in decision-making, for example, in medicine. The methods of using distances to reference objects cannot be reduced only to the use of linear combinations of monotonically transformed distances. An alternative approach is to use a set of distances between an arbitrary object $\boldsymbol{x}$ and reference objects as a new vector description $\boldsymbol{z(x)}$, which is then used to predict thevalue of Y using an algorithm

that was trained on a sample consisting of new vector descriptions of objects obtained in the same way the original training sample. At the same time, training can be performed using a variety of techniques, including random forest and gradient boosting. Using ensembles of decision trees allows not only to describe complex nonlinear dependencies, but also to achieve higher learning stability inherent in ensemble methods.

This approach was used to solve the problem of diagnostics by the characteristics of thrombodynamics, that is, by the dynamics of the formation of spontaneous fibrin clots.

Among the existing approaches to working with such curves, one can single out a method based on calculating several parameters characterizing it for each curve. The disadvantage of this approach is a certain loss of information. An alternative way is to use features corresponding to individual time points of reference. Each of these features takes on the values of the intensity of the reflected light at the corresponding point. The disadvantages of the approach are high dimensionality with a large number of reference points and high correlation of features. Our studies compared the effectiveness of methods based on distances to reference objects and a method based on the use of measurements at all possible reference points as signs when recognizing a group of patients with schizophrenia and a control group of subjects. Within the framework of the approach based on distances to standards, the efficiency of using three metrics was investigated: the Euclidean metric, the cosine metric, and the Minkowski metric (p=1). The two methods for selecting reference objects were used:

1. support vectors were selected from all objects based on the SVC method (Support Vector Classifier),
2. all objects of the training sample were used as reference objects.

The cross-validation method with testing at each step on one object (Leave One Out) was used to compare the effectiveness of logistic regression (LR), random decision forest (RDF) and gradient boosting over decision trees (GB). The estimation of the confidence interval of the classification result ROC AUC based on the bootstrap showed that the results obtained are stable. To ensure the clarity of the results and the possibility of their interpretation, along with multifactorial recognition methods, the data mining method was also used,

based on the optimal valid partitioning of the feature space (OVP method), within the framework of the approach presented in [1].

The ROC AUC values for the best-performing Minkowski metric (p = 1) are presented in the table. All objects of the training sample and objects selected by SVC were used as reference objects:

	LR	RDF	GB
with objects selection	0.737	0.778	0.706
without objects selection	0.723	0.778	0.762

The conducted research revealed the existence of a pronounced internal structure in the data generated using the distances to the reference objects. The internal structure also characterizes the differences between the group of patients with schizophrenia and the control group.

This research is funded by RFBR, grant 20-01-00609

[1] *Dorovskih I.V., Senko O.V., Chuchupal V.Ya., Dokukin A.A., Kuznetsova A.V.* On Possibility of Machine Learning Application for Diagnosing Dementia by Eeg Signals. // Mathematical biology and bioinformatics, 2019;14(2):543-553