CHOOSING AN EFFICIENT ALGORITHM FOR SOLVING THE CLASSIFICATION PROBLEM

Egamberdiev N.A.[1], Xolmuminov O.T.[2], Ochilov X.R.[3]

https://doi.org/10.5281/zenodo.7156156

Abstract. In this work, the selection of an effective algorithm for solving the classification problem was considered. The general mathematical formulation of the problem of classification is given and the available algorithms for solving it are analyzed. In addition, a taxonomy of the most used classification algorithms based on their ideological origins is presented. Classification errors were compared using K-Fold Cross-Validation testing. The coefficient of error in classification was taken into account as an efficiency indicator. The obtained results were checked on the basis of several data sets and the reliability of the algorithm was compared.

Keywords: K-Fold Cross-Validation testing method, Naive Bayes, Logistic regression, KStar, SVM, Random forest, Perceptron neural network.

Let us be given a set of n objects: $I = \{i_1, i_2, \dots, i_j, \dots, i_n\}$, here i_j object under consideration. Each object under considerationi_j is characterized as follows $i_j = \{x_1, x_2, \dots, x_h, \dots, x_m, y\}$. Each x_h variable depends on y and accepts values from different categorical sets. Properties of objects given to us $(x_1, x_2, \dots, x_h, \dots, x_m)$ Expression of laws of relevance based on y is called classification problem. A classifier model is called a classifier. If the set of values of y is finite, i.e $Y = \{y_1, y_2, \dots, y_r, \dots, y_k\}$ consists of, then this problem is a problem of classification, if the range of values of y consists of R real numbers, it is called a regression problem [1].

Depending on the classes, this is the most frequently solved problem of data analysis. The correction of this problem consists of correcting the unknown (new) object according to a certain rule to any of the previously given classes according to the set of tables [1]. To date, several algorithms have been developed to solve this problem. In particular, they can be cited as:

- with probability: 1R [2] algorithm, Naive Bayes [3, 4] algorithm and its various modifications: selective Naive Bayes [5], semiNaive Bayes [6], single dependence Bayesian classifiers [7, 8], K-dependence Bayesian classifiers [9], Naive Bayes extended with Bayesian network [10], unconstrained Bayesian classifiers [11] and Bayesian multinets;
- to the court tree method: overlay algorithm, ID3 (Iterative Dichotomiser 3), S4.5, CART (Classification and Regression Trees) [12], CHAID (chisquared automatic interaction detector) [13];

- by the nearest neighbors method: KNN algorithm [14], Wavelet Based K-Nearest Neighbor Partial Distance Search (WKPDS) algorithm [15], Equal-Average Nearest Neighbor Search (ENNS) algorithm, EqualAverage Equal-Norm Nearest Neighbor codeword search (EENNS) algorithm, Equal-Mean Equal-variance Equal-Norm Nearest Neighbor Search (EEENNS) algorithm [16];
- by the mathematical function method: least squares [17], SVM and its modifications GSVM (granular support vector) machines) [18,19], FSVM (fuzzy support vector machines) [20], TWSVMs (twin support vector machines).

Setting the issue. When solving the problem of classification into classes, the use of methods such as properly distributed neural network, Logistic regression, Naive Bayes, Base vectors, Random forest, and nearest neighbor gives good results. Many classification algorithms have been developed based on these methods. Each of them works well on different types of datasets. Therefore, it is important to determine which classification method is effective in solving the classification problem for the data set we have chosen. As an indicator of efficiency, it is necessary to look at the coefficient of error in classification.

The most important process in the issue of classification is the process of training the model. The model is trained to reduce errors, and then the model is tested. Objects not used during training are used for model testing.

But in many cases, dividing the educational samples into two sets in such a traditional way can cause a number of problems in the reliability of the model. That is, if the objects involved in testing are close to each other, and there are few objects among the objects involved in the training process, then the error in dividing the model into classes during testing becomes very large and obvious does not give results.

Therefore, it is advisable to use a more complicated, but more reliable testing method. One such method is the K-Fold Cross-Validation testing method for assessing model reliability. In this way, we can describe the testing process as follows. The main aspect of this method is that all the subjects participate in both the teaching process and the testing process.

In the K-Fold Cross-Validation testing method, the set of educational samples is divided into k parts. The model is then trained and tested k times. Each time it is trained, the i-th set is used only for testing, and the rest are used for training. Accordingly, the error is calculated for each test, and the average error is determined by the following formula.

$$E_o = \frac{1}{K}\sum_{i=1}^{k} E_i$$

Analysis of experimental test results. The Glass, Diabetes and German Credit datasets provided by www.kaggle.com were taken as an experiment. The characteristics of these data sets are given below (Table 1).

Table 1

Characteristics of data sets

№	Data sets	Number of objects	Number of characters	Number of classes
1	Glass	214	9	7
2	Diabetes	768	8	2
3	German Credit	1000	20	2

The problem of classification of the above data sets was solved using Naive Bayes, Logistic, MLP, Kstar, Decision Table, J48, PNN algorithms in Weka and KNIME programs. The same testing conditions were used for all algorithms, that is, the data was divided into 10 equal parts and 9 were used for training and one for testing. The average value was taken as the results (Table 2).

Table 2

Results from classification algorithms

№	The name of the algorithm	Data sets		
		Glass	Diabetes	German Credit
	Naive Bayes	48,5%	76,3%	75,4%
	Logistic	64,4%	77,2%	75,2%
	MLP	67,7%	75,3%	72,1%
	KStar	75,2%	69,1%	69,4%
	Decision Table	68,2%	71,2%	71,0%
	J48	66,8%	73,8%	70,5%
	PNN	63,1%	74,5%	66,6%

An examination of the results shows that Kstar performed well for the Glass dataset, Logistic for the Diabetes dataset, and Naive Bayes for the German Credit dataset.

Conclusion. In conclusion, it can be said that large amounts of information are currently being used in various areas of society. Based on these data, when choosing algorithms to solve the problem of classification, not only the error rate, but also the time taken to train the algorithm is becoming one of the criteria for evaluating the reliability of the algorithm. In such conditions,

choosing the classification algorithm based on the above-mentioned algorithm will give good results.

References:

1. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Xolod I.I. Texnologii analiza dannыx: Data Mining, Visual Mining, Text Mining, OLAP. 2-ye izdanie. Sankt–Peterburg, «BXV-Peterburg», 2007. -375 s.
2. Holte R.C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets// Machine Learning. – 1993. – № 11. – P. 63-90.
3. Maron, M. E.On relevance, probabilistic indexing and information retrieval // Journal of the ACM. – 1960. – V. 7, N. 3. – P. 216-244
4. Minsky M. Steps toward artificial intelligence // Proceedings of the IRE. - 1961. – V. 49. – P. 8-30
5. Lei Y. Visual tracker using sequential Bayesian learning: discriminative, generative, and hybrid // IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). – 2008. – V. 38, N. 6. – P. 1578-1591
6. Kononenko I. Semi-I Bayesian classifier // Machine Learning–EWSL-91. – 1991.- V. 482. – P. 206-219
7. Sahami M. Learning limited dependence Bayesian classifiers // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. – Portland, Ore, USA, 1996. – P. 335–338
8. Friedman N. Learning belief networks in the presence of missing values and hidden variables // Proceedings of the 14th International Conference on Machine Learning. – 1997. – P. 125-133.
9. Lei Y. Visual tracker using sequential Bayesian learning: discriminative, generative, and hybrid // IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). – 2008. – V. 38, N. 6. – P. 1578-1591
10. Ranka S. CLOUDS: A decision tree classifier for large datasets // Proceedings of the 4th Knowledge Discovery and Data Mining Conference, 1998.– P. 2-8.
11. Larose D. T. k-Nearest Neighbor Algorithm // Discovering Knowledge in Data: An Introduction to Data Mining. – 2005. – P. 90-106.
12. Hwang W. Fast kNN classification algorithm based on partial distance search // Electronics letters. – 1998. – V. 34, N. 21. – P. 2062-2063
13. Jeng-Shyang Q. Yu-Long, S. Sheng-He. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. – 2004. – V. 87, N. 4. – P. 961-963
14. Tang Y. [et al.]. Granular support vector machines for medical binary classification problems// Proceedings of the IEEE Symposium on Computational

Intelligence in Bioinformatics and Computational Biology (CIBCB '04), 2004. – P. 73-78.
15. Guo, H. A novel learning model-kernel granular support vector // International Conference on Machine Learning and Cybernetics. – 2009. – V. 2. – P. 930-935.
16. Lian K. [et al.] Study on a GA-based SVM Decision-tree Multi-Classification Strategy/ // Acta Electronica Sinica. – 2008. - V. 36, N. 8. – P. 1502–1507.
17. Lin C. F. Fuzzy support vector machines // IEEE Transactions on neural networks. – 2002. – V. 13, N. 2. – P. 464–471.
18. Savasere, A. An efficient algorithm for mining association rules in large databases // Proceedings of the 21th International Conference on Very Large Data Bases (VLDB '95). – Zurich, Swizerland, 1995. –P. 432-444
19. Baranov A.V. i dr. Data Mining. Teoriya i praktika (Pod red. I.N.Baryanseva). –M.: Izdatelskaya gruppa «BDS-press», 2006, – 208 s