GALAXY INTERNATIONAL INTERDISCIPLINARY RESEARCH JOURNAL (GIIRJ)
ISSN (E): 2347-6915
Vol. 10, Issue 5, May. (2022)

RAPID ASSIMILATION OF THE DECISION TREE LEARNING ALGORITHM

Sultangaliev Amirkhan Manasovich
2nd Year of Master's Degree Tashkent University of Information Technologies Named After Muhammad Al-Khwarizmi,
+998999513315 amir_xan@list.ru

ABSTRACT

We know that it will take a very long time to learn for a decision tree, and to solve this problem, we offer several ways of learning that will not take much time.

Keywords: Decision tree, algorithms, classification, differential evolution, differences.

INTRODUCTION

To date, much attention is paid to data analysis. One of the most important areas of data analysis is the task of classification. To solve this problem, there are many different methods of decision support. Decision trees have proven themselves well in this area. A decision tree is a decision–making method based on the application of various separation functions of the original data set, in particular simple threshold rules. A decision tree is a binary tree in which: a function is assigned to each inner vertex, and a forecast is assigned to each leaf vertex.

METHODS

Decision trees, like any other machine learning algorithm, have their defining parameters. So, for example, to choose the partition of a data set in some node, it is necessary to optimize some function. This function in trees is called the criterion of informativeness. The criteria of informativeness can be different: loss function, Gini criterion, entropy criterion, etc. There are also various stop criteria. The variation of these and other parameters provides a variety of algorithms for learning the decision tree.

In this paper, two main algorithms for learning the decision tree were implemented and compared with each other: ID3 and CART. These algorithms terminate if there are observations of one class left in the sheet or if a limit on the depth of the tree is set, for the problems solved in this work, no depth limit was imposed.

The main difference between these algorithms lies in the different criteria of informativeness. The ID3 algorithm uses the entropy criterion:

$$H(R)=-\sum_{k=1}^{k} p_k \log p_k ,$$

Where p_k is the proportion of objects of class k that hit the vertex R.

The CART algorithm uses the Gini criterion:

$$H(R)=\sum_{k=1}^{k} p_k (1-p_k) .$$

Optimization of the presented information content criteria in the decision tree learning algorithm is performed by a complete search over the original data set, but this is a very time-consuming process, since it is necessary to calculate the values of the information content criterion for all attribute values for all observations of the training sample. To reduce the running time of the algorithm, in this paper, an optimization of the information content criterion is proposed for an attribute selected using the Separation Measure algorithm, which considers the attribute more important, for which the sample averages by class are the most distant. Optimization was performed by the method of differential evolution.

RESULTS OF THE EXPERIMENT

Issue number	ID3	CART	ID3 (DE)	CART (DE)
1	0.85; 4422; 18	0.85; 4741; 22	0.85 (0.86); 55; 16	0.85 (0.86); 60; 17
2	0.71; 8; 15	0.65; 8; 16	0.7 (0.75); 2; 12	0.68 (0.73); 2; 13
3	0.88; 259; 13	0.89; 390; 16	0.91 (0.94); 10; 13	0.9 (0.92); 11; 15
4	0.79; 196; 10	0.77; 218; 12	0.77 (0.79); 21; 12	0.76 (0.8); 21; 12

RESULTS AND DISCUSSION

The algorithmic complex uses an algorithm of differential evolution with self-tuning of the mutation strategy and with the adaptation of parameters according to the Success History Adaptation algorithm. Differential evolution performed 300 calculations of the objective function, i.e. the criterion of informativeness.

Databases from the repository were taken as classification tasks. In accordance with the content of the databases, the following tasks were determined:

1) determination of soil type from satellite image;
2) determining the type of car;
3) object type recognition by its segment;
4) recognition of the urban landscape.

The results of comparing the algorithms are presented in the table. In each cell of the table, the classification accuracy on the test dataset, the training time of the decision tree (in seconds) and the depth of the resulting tree are indicated. For algorithms with optimization of the criterion of informativeness by the method of differential evolution, the results averaged over 100 runs are presented due to the stochasticity of the algorithm. Also, for trees optimized by the differential evolution algorithm, the best found classification accuracy on test data is indicated in parentheses. The best results of accuracy and time for each of the tasks are highlighted in bold.

In the presented paper, a comparison of some of the main algorithms for learning a decision tree is carried out. A possible way to optimize the learning process is presented. As you can see, the proposed modification not only speeds up the learning process of the decision tree, but also allowed us to find a better solution on the tested problems than in the standard way. It should also be noted that the depth of the trees did not change significantly when using the modification.

CONCLUSION

Decision trees are a powerful tool for solving problems, including in the rocket and space industry. In the future, in order to increase the efficiency of this method, it is proposed to automate the process of forming decision trees by evolutionary algorithms.

REFERENCES

1. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C.T. Stone // Wadsworth. Belmont. California. 1984.
2. Mitrofanov S. A. Primenenie neyroevolyucionnogo podxoda pri reshenii zadach klassifikacii // Reshetnevskie chteniya: Materials of the XXI International Scientific Conference (November 11-14, 2014, Krasnoyarsk) : at 3 o'clock / under the general editorship of Yu. Yu. Loginov; SibGU named after M. F. Reshetnev. Krasnoyarsk, 2017. No. 2 (21). pp. 226-227.
3. Kushnareva T. V., Lipinsky L. V. Algoritm geneticheskogo programmirovaniya dlya avtomatizirovannogo formirovaniya derevyev prinyatiya resheniya // Reshetnevskie chteniya : Materials of the XVIII International Scientific Conference (November 11-14, 2014, Krasnoyarsk) : at 3 o'clock / under the general editorship of Yu. Yu. Loginov ; Sib. gos. aerospace. un-T. Krasnoyarsk, 2014. No. 1 (18). pp. 84-86