ITM Web of Conferences **59**, 01018 (2024) https://doi.org/10.1051/itmconf/20245901018
HMMOCS-II 2023

Two-level optimization by differential evolution in decision tree learning algorithm

Sergei Mitrofanov*, and *Eugene* Semenkin

Reshetnev Siberian State University of Science and Technology, Department of System Analysis and Operations Research, Institute of Informatics and Telecommunications, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation

Abstract. Decision tree learning algorithms have a long history, but they are still popular due to their efficiency. Tree construction starts at the root and continues to each leaf node, creating a "near-optimal" tree. One of the key steps in creating a decision tree is the selection of a feature to split at each root node, which affects the classification accuracy. This process can be quite labor intensive. The article proposes a new approach to constructing decision trees based on the use of differential evolution for two-level tree optimization. Differential evolution works in two stages: at the first level, a feature is selected for separation, and at the second level, the threshold value for this feature is optimized. The results of the work were tested on several examples of classification problems.

1 Introduction

The use of artificial intelligence has reached such a level that it is being implemented everywhere, in information technology, in household appliances, in cars, in education, and so on. Due to the widespread use of artificial intelligence, its developers need to constantly be focused on improving the efficiency of its work on applied tasks, since user loyalty depends on this. As a result, there is a need to develop additional algorithms and approaches to improve basic artificial intelligence methods [1–2]. Decision trees are one of the most well-known basic machine learning algorithms [3], which allows to present the results of itself work in a form convenient for analysis by non-information technology specialists [4-5]. Decision trees have many different implementations, and there are many collective algorithms based on them. Decision trees are capable of solving classification, clustering and forecasting problems [5].

Decision tree learning is a recursive process that involves splitting a set of objects using simple rules until a vertex is declared to be a leaf vertex. This process is repeated until all leaf nodes are recognized as leaves [6]. During the process of creating a decision tree, various problems arise, each of which must be solved. One of the main problems is the choice of an attribute [7] by which division will be made at a given node (partition attribute).

* Corresponding author: sergeimitrofanov95@gmail.com

The article examines the effectiveness of an approach based on the use of differential evolution for two-level optimization of a decision tree, using the example of solving practical classification problems.

2 Two-level optimization approach

Various filtering methods have been previously studied and tested to select a separation feature [8]. The results of practical tests showed that the most suitable methods may differ depending on the specific application. However, instead of using a specific feature selection method, optimization can be applied to select the most suitable one.

We use differential evolution for two-level optimization. In the first step, we optimize the feature number for the threshold value, where we impose restrictions on the number of features and only allow integer values. In the second step, we optimize the threshold value. We optimize both values at the same time.

Figure 1 shows a diagram of the proposed approach.

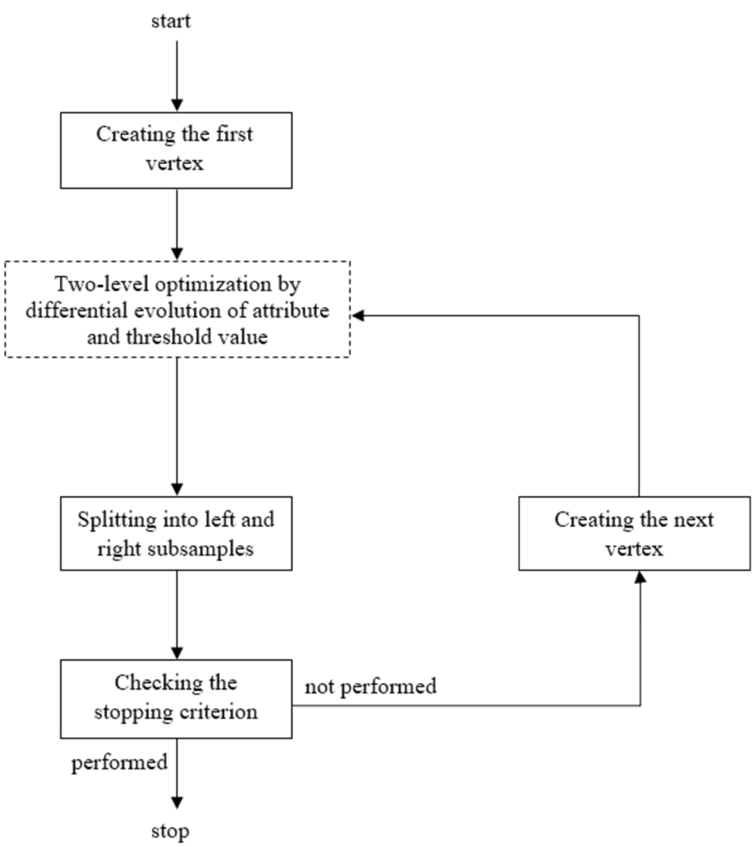

Fig. 1. Decision tree learning algorithm with two-level optimization.

3 Classification tasks

To test the effectiveness of the proposed approach, the following classification problems in Table 1 were used [9].

Table 1. Classification tasks.

Task number	Name	Sample size	Number of attributes	Number of target classes
1	Type of car determining	470	18	4
2	Type of cityscape determination	675	147	9
3	Iris variety recognition	150	4	3
4	Parkinson's disease recognition	756	754	2
5	Defining images by segments	2310	19	7
6	Heart defect recognition	270	13	2
7	Soil type recognition from satellite imagery	6435	36	6
8	Biodegradable chemicals recognition	1055	41	2

To compare the proposed approach with other methods for training decision trees, a standard algorithm and a modified algorithm were used using the Separation Measure feature selection method and threshold optimization using differential evolution, which was studied in the authors' article [6].

It is important to note that the results for the modified algorithms using differential evolution are presented as the average of 100 runs due to the randomness of the algorithm. We will consider the efficiency as the ratio of the number of correctly classified objects to the number of all objects in the test sample.

The results are presented in Table 2.

Table 2. Tests results.

Task number	Efficiency of the standard algorithm	Efficiency of the Algorithm with Separation Measure	Efficiency of the algorithm with two-level optimization
1	0.711	0.718	**0.781**
2	0.794	0.784	**0.829**
3	0.978	**1**	0.969
4	0.454	**0.74**	**0.74**
5	0.885	0.922	**0.96**
6	0.778	0.84	**0.903**
7	0.842	0.854	**0.882**
8	0.808	0.798	**0.839**

4 Conclusion

The method, based on the use of two-level optimization when constructing a decision tree, has demonstrated its effectiveness in comparison with standard and modified decision tree learning algorithms. The resulting model turned out to be more accurate and reliable when solving classification problems. This method allows you to automatically select the most appropriate feature for dividing data at each tree node, which provides a more flexible and adaptive model structure. This is especially useful in situations where the data has a complex structure and many relationships between features.

It is important to note that the use of differential evolution as an optimizer allows one to find optimal values of model parameters in conditions of multidimensionality and nonlinearity of the search space. This ensures faster and more reliable convergence of the algorithm to the optimal solution, which in turn increases the accuracy of the model.

Thus, the proposed method is a promising direction in the development of machine learning algorithms and can be used in various applied tasks that require data analysis and classification.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant № 075-15-2022-1121).

References

1. O.I. Zakharova, E.S. Artyushkina, S.V. Kholopov, Eurasian Scientific Association **4-2(62)**, 97-99 (2020)
2. S.A. Mitrofanov, T.S. Karaseva, Proceedings of the International Workshop “Hybrid methods of modeling and optimization in complex systems”, 111-116 (2022)
3. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and regression trees* (1984)
4. A.N. Vinogradov, A.N. Lebedev, M.V. Tereshonok, T-Comm – Telecommunications and Transport, 28-30 (2009)
5. A. Volkov, N. Yatskov, V. Grinev, Journal of the Belarusian State University, Maths, Computer Science **1**, 77-89 (2019)
6. S.A. Mitrofanov, E.S. Semenkin, Siberian Journal of Science and Technology **20(3)**, 312-319 (2019)
7. S.A. Mitrofanov, E.S. Semenkin, IOP Conf. Ser.: Materials Science and Engineering **1047**, 12082 (2021)
8. S. Mitrofanov, E. Semenkin, AIP Conf. Proc. **2700(1)**, 030008 (2023)
9. Machine Learning Repository [Electronic resource]. URL: https://archive.ics.uci.edu/ml/index.php (accessed: 06.10.2023)