HMMOCS-II 2023

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ | SIBERIAN FEDERAL UNIVERSITY

IASA

Science & Technology City Hall

HMMOCS-II 2023:
Hybrid methods of modeling and optimization in complex systems

«Influence of the number of attribute binarization thresholds on the quality of the generated set of rules for the method of logical analysis of data»

Roman Kuzmich, Katerina Ponomareva, Artur Nikiforov

Krasnoyarsk 2023

Introduction

The key idea of the classification problem is to train the classification algorithm on observations of the training sample and then apply it to observations of the testing sample in order to determine the class for each of its observations. The training sample contains, in addition to the input attribute values, the values of the output attribute (class) for each observation belonging to it. The exam sample contains only the input attribute values for each observation belonging to it. Thus, it is necessary, having a trained classification algorithm and knowing the input attribute values for each observation of the examining sample, to determine the output attribute values for them, i.e. the class to which each observation of the examining sample belongs.

Each algorithm has its own advantages and disadvantages. In this paper, we will consider a general problem associated with the operation of most machine learning methods, using the example of the method of logical analysis of data.

Problem statement

One of the problems with most machine learning methods is choosing the values for all the parameters it needs to work. The choice of method parameter values is implemented by the researcher based on his experience in solving similar problems or any subjective ideas.

Therefore, it is necessary to propose methods that allow us to correctly set the parameter values for the method. One way is to obtain information to set the correct parameters by conducting repeated experiments on the problem being solved. In this case, it is necessary to establish a quantitative criterion or several criteria by which we will compare the parameter values proposed for the method.

Research Questions

- How to organize the procedure for selecting the number of binarization thresholds for quantitative attributes?
- What criterion(s) should be used to compare the values proposed by the researcher for the parameter «number of thresholds for binarization of quantitative attributes»?
- What effect does the number of attribute binarization thresholds have on the generalizing abilities of the resulting logical rules?

Purpose of the Study

Answers to these questions will allow us to determine a strategy when setting up the parameters of the method of logical analysis of data at the stages of binarization of quantitative attributes and searching for a reference set of binary attributes. It should also be noted that a compromise can be found between the classification accuracy and the computational complexity of the problem by correctly setting the parameters of the method under study.

Research Methods

The paper proposes the following method for binarizing quantitative characteristics:

- the number of binarization thresholds for attribute k is set;
- a procedure is performed for ordering in ascending order the values of the training sample for each quantitative attribute $j = 1,\ldots,d$: $f_1^j <= f_2^j <= \ldots <= f_n^j$;
- the number of measured values for each attribute is divided into $k+1$ equal groups, the number of elements z in each group is determined;
- the threshold values F_m^j are set equal to the values f_{z*m+1}^j, где $m = 1,\ldots,k$.
- binary attributes $b_1^j, b_2^j ,\ldots, b_{k+1}^j$ are assigned for each quantitative characteristic f^j.
- the values of binary variables $b_0^j, b_1^j ,\ldots, b_k^j$ are determined based on the obtained measurement of the quantitative characteristic f^j using the following formula:

$$b_{m-1}^j = \begin{cases} 1, f^j > F_m^j, \\ 0, \text{otherwise.} \end{cases}$$

Findings

Let us conduct numerical studies on the problem of complications prediction of the myocardial infarction – ventricular fibrillation.

It is proposed to experimentally determine the best value of the parameter «Number of attribute binarization thresholds»proposed by the researchers. In this case, 4 options for conducting the experiment are considered:

- a complete set of attributes, a complete set of rules (Table 1);
- a complete set of attributes, a partial set of rules (Table 2);
- a truncated set of attributes, a complete set of rules (Table 3);
- a truncated set of attributes, a partial set of rules (Table 4).

Number of attribute binarization thresholds	Set of rules	Average coverage of negative observations	Average coverage of positive observations	Average pattern degree	Classification accuracy, %
2	Neg.	28	5	5	80
	Pos.	4	21	3	67
3	Neg.	28	5	5	90
	Pos.	4	26	3	83
4	Neg.	26	5	5	80
	Pos.	4	24	3	83
5	Neg.	26	5	5	80
	Pos.	4	26	3	78
6	Neg.	25	5	5	90
	Pos.	4	30	2	78

Table 1. Classification accuracy and indicators of rules set with complete sets of attributes and rules for the problem of ventricular fibrillation

Findings

Number of attribute binarization thresholds	Set of rules	Average coverage of negative observations	Average coverage of positive observations	Average pattern degree	Classification accuracy, %
2	Neg.	27	5	5	70
	Pos.	4	20	4	61
3	Neg.	27	5	4	80
	Pos.	4	27	3	78
4	Neg.	23	5	5	70
	Pos.	3	25	4	72
5	Neg.	22	5	4	80
	Pos.	4	26	3	72
6	Neg.	23	5	4	80
	Pos.	4	30	3	72

Table 2. Classification accuracy and indicators of rules set with complete set of attributes and partial set of rules for the problem of ventricular fibrillation

Findings

Number of attribute binarization thresholds	Set of rules	Average coverage of negative observations	Average coverage of positive observations	Average pattern degree	Classification accuracy, %
2	Neg.	28	5	4	70
	Pos.	5	18	4	56
3	Neg.	27	5	5	90
	Pos.	5	23	4	78
4	Neg.	25	5	4	70
	Pos.	4	24	3	61
5	Neg.	25	5	4	70
	Pos.	4	24	3	61
6	Neg.	22	5	4	90
	Pos.	4	29	2	78

Table 3. Classification accuracy and indicators of rules set with truncated set of attributes and complete set of rules for the problem of ventricular fibrillation

Findings

Number of attribute binarization thresholds	Set of rules	Average coverage of negative observations	Average coverage of positive observations	Average pattern degree	Classification accuracy, %
2	Neg.	27	5	4	60
	Pos.	5	17	4	56
3	Neg.	27	4	4	80
	Pos.	4	23	3	83
4	Neg.	27	5	4	70
	Pos.	4	22	3	72
5	Neg.	24	5	4	70
	Pos.	4	22	3	61
6	Neg.	23	5	4	70
	Pos.	4	25	3	72

Table 4. Classification accuracy and indicators of rules set with truncated set of attributes and partial set of rules for the problem of ventricular fibrillation

Conclusion

In this paper, we investigated the influence of the parameter «Number of attribute binarization thresholds» on the quality of the generated set of rules for the method of logical analysis of data. An experimental approach is proposed to determine the best value of the specified method parameter for the problem being solved.

According to the obtained results, it can be noted that the best value of the parameter «Number of thresholds for binarization of attributes» for the problem being solved is the value 3. It is this number of thresholds for binarization of quantitative attributes that allows us to obtain the highest quality logical rules with the highest average coverage of observations. A classifier built on the basis of such rules has good generalizing abilities. Also in favor of the parameter value obtained experimentally is the reduction of the attribute space for searching for logical rules, i.e. reducing the computational complexity of the problem being solved.

Acknowledgments

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

Contacts

Roman Kuzmich, Katerina Ponomareva, Artur Nikiforov

Siberian Federal University

E-mail: romazmich@gmail.com

HMMOCS-II 2023: Hybrid methods of modeling and optimization in complex systems