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*, and *Artur Nikiforov*

Siberian Federal University, 79, Svobodny Ave., Krasnoyarsk, Russia

Abstract. For efficient operation of machine learning methods, it is necessary to set their parameters up properly. Both the computational complexity and the accuracy of the method for the problem being solved can depend on the selected parameter values. The paper discusses the method of logical analysis of data, which is used to solve classification problems and consists of a number of steps. At each step, it is necessary to set the method parameters up to suit the problem being solved. When setting parameters, one should be guided by a compromise between the accuracy of the method and the computational complexity. With comparable values for classification accuracy of several variants of the method implementation, preference will always be given to the simplest of them in terms of computational complexity. Since the method works only with binary characteristics, at the first stage it is necessary to binarize quantitative characteristics. The binarization procedure is associated with the choice of the "number of binarization thresholds" parameter. This paper proposes an experimental approach to determine the best value of the specified method parameter for the problem being solved.

1 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. To solve classification problems, various machine learning methods are effectively used: naive Bayes classifier [1], support vector method [2], nearest neighbor method [3], artificial neural networks [4] and others. There is no universal classification algorithm that is ideal for any task. Each algorithm has its own advantages and disadvantages. In this paper, we will consider a general problem associated with the operation

* Corresponding author: romazmich@gmail.com

of most machine learning methods, using the example of the method of logical analysis of data [5].

2 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. An unsuccessful choice of parameter values can negatively affect the performance of the method (for example, the speed of the method work) and the results obtained (for example, classification accuracy). 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. We also note that when setting parameters, one should be guided by a compromise between the accuracy of the method and the computational complexity. With comparable accuracy values for several method implementation options, preference will always be given to the simplest of them in terms of computational complexity.

3 Research questions

This study needs to answer the following 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?

4 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.

5 Research methods

The paper explores a method of logical analysis of data designed to solve classification problems. As input, the method receives a sample of data on the binary classification problem being solved, which can consist of a large number of input and one output attributes. Subsequently, the received data is processed in order to obtain a set of logical rules, which are presented in the form of a small number of elementary expressions understandable to humans. Based on the specified set of rules, the process of making decisions about whether a new observation belongs to a particular class occurs by voting by majority. It should be noted that the process of extracting logical rules from the initial sample in the classification

method under study goes through a number of stages: binarization of quantitative attributes, search for a reference set of attributes for the problem being solved, and formation of a set of patterns. At each stage of the method there are a number of parameters that affect the quality of the resulting logical rules, so we will provide a detailed description of each stage.

The initial data is represented by two disjoint sets Ψ^+ (the set of n-dimensional observations of the positive class) and Ψ^- (the set of n-dimensional observations of the negative class). Observation attributes take on values of different natures (binary, quantitative, etc.). Since the method only works with binary characteristics, it is necessary to perform the procedure of binarization of quantitative attributes.

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:

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

After binarization of attributes, their total number for the considered problem can increase sharply, which directly affects the computational complexity of the method. An idea arises to reduce the attribute space of searching for logical rules. In the method of logical analysis of data, a similar idea is reflected at the stage of searching for the minimum support set S, i.e. a set of attributes that allows us to separate Ψ^+ and Ψ^-. At the next stages, the projections Ψ_S^+ and Ψ_S^- of the sets Ψ^+ and Ψ^- onto S will be used [6].

Let us represent the problem of finding the minimum support set as a combinatorial optimization problem. Let us associate each attribute b_i ($i = 1,\ldots,t$) of the original sample with a new binary variable r_i. Values of r_i equal to 0 indicate non-belonging to S, and values equal to 1 indicate belonging to S. The notations $Q = (q_1, q_2,\ldots,q_t)$ are introduced. This vector is associated with Ψ_S^+, and this vector $U = (u_1, u_2,\ldots,u_t)$ is associated with Ψ_S^-. A variable is entered:

$$m_i(Q,U) = \begin{cases} 1, q_1 \neq u_1, \\ 0, q_1 = u_1. \end{cases}$$

The condition of disjoint Ψ_S^+ и Ψ_S^- requires the fulfillment of the inequality $\Sigma(Q,U)r_i \geq 1$ for any $Q \in \Psi_S^+$ and $U \in \Psi_S^-$. You can also replace the number 1 on the right side of the inequality with the integer g to strengthen the constraint.

The final optimization model for searching for a reference set of attributes can be presented as follows:

$$\sum_{i=1}^{t} r_i \to \min$$

$$\sum_{i=1}^{t} m_i(Q,U) r_i \geq g \text{ for any } Q \in \Psi_S^+ \text{ и } U \in \Psi_S^-,$$

where $r \in \{0,1\}^t$.

The next stage is the construction of sets of positive and negative logical rules based on Ψ_S^+ and Ψ_S^-.

Let's consider the formation of a set of positive logical rules. The main criterion for the quality of the resulting rules will be the maximization of their coverage Ψ_S^+ under the restriction on coverage Ψ_S^-. Therefore, for each observation $\alpha \in \Psi_S^+$ we will look for an α-

rule that allows us to cover the maximum number of observations Ψ_S^+. Let's set the required logical rule using binary variables $X = (x_1, x_2, \ldots, x_t)$:

$$x_k = \begin{cases} 1, \text{if the k} - \text{th attribute is in the rule} \\ 0, \text{otherwise.} \end{cases}$$

When constructing pure logical rules, the α-rule should not cover a single element of Ψ_S^-. Therefore, the inequality must be satisfied:

$$\sum_{\beta_k \neq \alpha_k}^{t}{}_{k=1} \; x_k \geq 1 \text{ for any } \beta \in \Psi_S^-$$

To strengthen the constraint, we use another positive integer g on the right side of the inequality instead of 1.

On the contrary, the observation $\sigma \in \Psi_S^+$ is then included in the generated rule when it differs from $\alpha \in \Psi_S^+$ only in those attributes that are not included in the resulting rule. The total number of observations Ψ_S^+ for the α-rule is calculated by the formula:

$$\sum_{\sigma \in \Psi_S^+} \prod_{\substack{j=1 \\ \sigma_k \neq \alpha_k}}^{t} (1 - x_k)$$

Therefore, the optimization model for the formation of pure logical rules can be represented as follows:

$$\sum_{\sigma \in \Psi_S^+} \prod_{\beta_k \neq \alpha_k}^{t}{}_{j=1} \; (1 - x_k) \rightarrow \max \tag{1}$$

$$\sum_{\beta_k \neq \alpha_k}^{t}{}_{k=1} \; x_k \geq g \text{ for any } \beta \in \Psi_S^-, x \in \{0,1\}^t \tag{2}$$

Samples in real classification problems are characterized by the presence of missing data, outliers, which leads to the problem of separability of observations Ψ_S^+ and Ψ_S^-. In this case, a transition to the construction of partial logical rules is proposed, i.e. rules that are allowed to cover a certain number of observations of another class [7]:

$$\sum_{\beta \in \Psi_S^-} z_\beta \leq G, \tag{3}$$

$$z_\beta = \begin{cases} 0, \text{if } \sum\limits_{\substack{j=1 \\ \beta_k \neq \alpha_k}}^{t} x_k \geq g; \\ 1, \text{otherwise,} \end{cases}$$

where G is the number of observations Ψ_S^- that are allowed to be covered by the logical rule.

6 Results and discussion

Let us conduct numerical studies on the problem of complications prediction of the myocardial infarction – ventricular fibrillation [8].

To find the rules, we use a modified model with rules covering a certain limited number of observations of another class (1, 3). The experiments were implemented on a sample consisting of 70 observations of the positive class (patient with a complication) and 70 observations of the negative class (patient without a complication). The initial sample is divided into training (80%) and testing (20%).

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).

As a criterion for selecting the best value of this parameter, it is proposed to use the average coverage of logical rules formed on the basis of the obtained binary attributes.

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

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

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

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

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

7 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.

At the next step of the research, it is planned to integrate a mechanism for self-tuning of the method parameters to solve practical problems.

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

References

1. M. Burlakov, *Using optimized naïve bayes classifier in the problem of SMS classification*, in Proceedings of the Samara Scientific Center of the Russian Academy of Sciences **18**, 4-4 (2016)
2. S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Can. gen. and prot. **15(1)** (2018)
3. S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, IEEE trans. on neural netw. and lear. syst. **29(5)** (2017)
4. S.D. Shibaykin, A.I. Egunova, A.A. Abbakumov, *Analysis of the use of neural networks, gradient boosting and the nearest neighbor method for classifying normative and reference information*, in Proceedings of the Scientific and Technical Bulletin of the Volga Region **2** (2020)
5. R. Kuzmich, A. Stupina, L. Korpacheva, S. Ezhemanskaja, I. Rouiga, Informatica **29(3)** (2018)
6. P.L. Hammer, T. Bonates, Logical data analysis: From Combinatorial Optimization to Medical Applications. RUTCOR Research Report **10-2005** (2005)
7. R.I. Kuzmich, I.S. Masich, Contr. syst. and inform. tech. **2** (2014)
8. S.E. Golovenkin, A.N. Gorban, B.A. Schulman et al., Comp. center of Sib. Br. of Rus. Acad. of Scien. **6** (1997)