Specifics of the tasks by adjusting the parameters of the logical data analysis method

R I Kuzmich[1], A I Vinogradova[2,3]

[1]Department of Economics and Information Technology of Management, Siberian Federal University, Krasnoyarsk, Russia
[2]Department of Advertising and Cultural Studies, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

E-mail: vinogradova50@gmail.com

Abstract. One of the advantages of the data analysis method is the ability to take into account the specifics of the problem. When this advantage is realized in practice, the method parameters are adjusted to a specific task. The more parameters a method needs to be configured, the greater the possible number of implementations. However, in this case it becomes more difficult to configure this method for a specific task. Note also that the correct setting of the method parameters allows you to find a compromise between the criteria set by the customer to the results of the method. The paper presents the possibility of adjusting the parameters of the method of logical data analysis in order to take into account the specifics of the problem being solved, and adjusts the parameters of the method in solving the problem of controlling the landing of a spacecraft.

1. Introduction

The method of logical data analysis refers to the logical classification algorithms, the principle of which is to identify patterns in the data and formalize them in the form of a set of rules described by a simple formula. This method has been successfully used to solve a number of problems from different areas [1-3]. The main idea of the method is the joint use of actions on "differentiation" and "integration", produced on the area of the space of initial features, containing the given positive and negative observations. At the step of "differentiation" (pattern formation) a family of small subsets with characteristic positive and negative features are determined. At the "integration" step (construction of the classifier), the unions of these subsets formed in a certain way are considered as approximations of the regions of the feature space containing positive and, accordingly, negative observations [4].

The method of logical data analysis is a flexible tool enough for data analysis, allowing to take into account the specifics of a particular classification task and the requirements of the customer when solving it. At the stages of construction of the reference set (a set of features that allows to separate with high accuracy positive observations from negative ones), formation of regularities, construction of the classifier, there are parameters of the method, which by means of purposeful adjustment allow to maintain a balance between different criteria for comparison of classification algorithms. The following is a description of the method parameters for each of these steps and how to configure them.

2. The possibilities of setting the parameters of the logical data analysis method

At the stage of construction of the reference set, the researcher establishes the minimum number of differences between the observations of the two classes, i.e. the number of features on which they should differ. We obtain different sets of features that are used in the future in the construction of patterns, by varying this parameter. Changes in this parameter affect the accuracy of the classification and the complexity of the rules. The smaller the set of features used for separation, the lower the complexity of building rules, as the search space is reduced. However, with a significant reduction in the search space, it is not possible to build rules and a composition of these rules that correctly classifies the observations of the test sample.

At the stage of regularities formation, the use of an optimization model that allows the rule to cover a small number of observations of another class, allows you to find patterns with a higher coverage, from which a more accurate classifier is built. This approach is effective in solving problems with the emissions and noise presence and a large number of gaps in the sample data.

When using this optimization model, the researcher sets the number of observations of another class, which can capture each rule. With the help of regulation of this parameter, a compromise is established between the recognizing and generalizing classifier abilities. When the parameter value is low, the retraining effect occurs because the percentage of correctly classified observations from the training sample exceeds the percentage of correctly classified observations from the test sample. By increasing the value of the parameter, we achieve a balance between the recognizing and generalizing abilities of the classifier.

Also in the search for regularities in the use of optimization models to generate rules that produce significantly different subsets of observations of the sample, there is a parameter that indicates the maximum number of patterns that covers the observation of the training samples in the classifier [5]. This parameter for each class is set in the range from 1 to the maximum number of built regularities for this class. If the parameter takes a value close to or equal to the maximum number of regularities for this class, the new classifier works similarly to the optimization model with the maximum coverage. If the argument is to aim for 1 reduces the number of rules with the value of the objective function is greater than 0, which make up the classifier, as it captures the observations and their weights are set to zero. In the new classifier there is an insufficient number of regularities, which in the end are not able to classify the newly incoming observations, ie. the quality of classification is reduced. In this case, there is a high percentage of refusals from classification. It is empirically verified that the parameter value should be selected in the range from 5 to the average coverage of the regularities constructed using the optimization model with the maximum coverage, and the lower the parameter value, the less the number of rules in the classifier, i.e., its interpretability increases.

At the stage of classifier constructing in the implementation of the algorithmic procedure of building a classifier as a composition of informative patterns the threshold of informativeness acts as the parameter governing the number of patterns in the classifier [6]. With its gradual increase, the interpretability of the classifier increases, as the number of rules in it decreases, but, starting with a certain parameter value, there is an increase in refusals from the classification, therefore, a decrease in the accuracy of the classification as a whole. The increase in failures is due to the removal of all rules that previously covered certain observations of the test sample, i.e. the appearance of uncovered observations during the test. Therefore, it is necessary to set the correct value of the information threshold in order to maintain a balance between the interpretability of the classifier and the accuracy of the classification.

3. Results

We will adjust the parameters of the method in solving the problem of controlling the landing of the spacecraft. Note that the sample size for this problem is 15. Table 1 shows a sample for this task, consisting of 6 observations that belong to the class with manual control of the ship (class 0), and 9 observations belonging to the class with automatic landing of the ship (class 1). Each object in the

sample is characterized by seven features: stability, error, sign, and, magnitude, visibility, class. As you can see, there are missing values in the sample, which are marked "*" in table 1.

Table 1. Initial sample for the spacecraft landing control problem

class	stability	error	sign	wind	magnitude	visibility
1	*	*	*	*	*	1
0	1	*	*	*	*	0
0	0	2	*	*	*	0
0	0	1	*	*	*	0
0	0	3	1	1	*	0
0	*	*	*	*	4	0
1	0	4	*	*	1	0
1	0	4	*	*	2	0
1	0	4	*	*	3	0
1	0	3	0	0	1	0
1	0	3	0	0	2	0
1	0	3	0	1	1	0
1	0	3	0	1	2	0
0	0	3	0	0	3	0
1	0	3	0	1	3	0

The task is to extract the rules on the basis of the available data sample, which can be used to classify observations.

The feature of the configuration method of the logical analysis of data for this task is the choice of the testing method. As a rule, for classification tasks the percentage division is used – a method of testing, in which the initial sample is divided into two parts: training and test. But since the sample of observations consists of only 15 observations, cross-checking is used as a method of testing in this case.

The most frequently used method of cross-validation – *k*-regional method of statistics. This method consists in random division of the sample into *k* approximately identical subsets, one of these subsets is marked as a test subset, the model is built on *k*-1 subsets, and then tested on *k*-volume. This process is repeated *k* times, each time a new test subset is selected, then the average accuracy is displayed as a measure of the quality of the method used.

The case of *k*-regions is called the method of a penknife or alternate skipping if the number *k* is taken equal to the number of observations in the sample, i.e. the test subset always consists of only one observation [7].

Since there are missing values in the sample, a modified optimization model was used to find the rules, allowing the rules to cover a limited number of observations of another class. To solve the optimization problem, optimization algorithms based on the search for the boundary points of the admissible area [8-9] were used. These algorithms were developed specifically for this class of problems and are based on the behavior of monotone functions of the optimization model in the space of Boolean variables. The boundary point search algorithms are search algorithms, i.e. they do not require functions to be specified explicitly, using algebraic expressions, but use function calculations at points [10].

Examples of the rules that make up the classifier for the method of logical data analysis are given in table 2. The rules are obtained using a software application implemented by the authors [11]

Table 2. Examples of rules for a spacecraft landing control problem

class	stability	error	sign	wind	magnitude	visibility
0	1					
0		<3				
0			1			
1						1
1					<4	
1		≥3				

According to the results, the classification accuracy was 80 %, i.e. 12 out of 15 observations were classified correctly. Each constructed rule consists of one variable. When building all the rules, only the "wind" variable is not involved. The obtained rules allow us to answer the main question: why does a particular observation belong to this class?

To compare the results of the proposed method by accuracy, this problem is solved in the WEKA data analysis system using algorithms C4.5 [12], RIPPER [12], Adaboost [13]. Number of correctly classified observations for these algorithms: C4.5-9, RIPPER-9, Adaboost-11. Thus, the method of logical data analysis as a whole showed the best result in terms of classification accuracy, in addition, it has the ability to maintain a balance between different criteria for comparison of classification algorithms.

It should also be noted that the peculiarity of the proposed method is that instead of simply answering the question to which class the new observation belongs, it builds an approximation of the regions of the feature space containing observations of the corresponding classes. The most important advantages of this approach are the ability to provide an explanation for any solution obtained by the method, the ability to identify new classes of observations, the ability to analyze the role and nature of signs.

References

[1] Alexe S, Blackstone E and Hammer P L 2003 Coronary Risk Prediction by Logical Analysis of Data *Annals of Operations Research* **119** 15-42

[2] Hammer P L, Kogan A and Lejeune M 2004 Modeling Country Risk Ratings Using Partial Orders *RUTCOR Research Report* 24-2004

[3] Herrera J F A and Subasi M M 2013 Logical Analysis of Multi-Class Data *RUTCOR Technical Report* 05-2013

[4] Brauner M W, Brauner D, Hammer P L , Lozina I and Valeyre D 2004 Logical analysis of computer tomography data to differentiate entities of idiopathic interstitial pneumonias *RUTCOR Research Report* 30-2004

[5] Kuzmich R and Masich I 2014 Modification to an objective function for building patterns aimed at increasing the distinction between the rules of the classification model *Management Systems and Information Technologies* **2(56)** 14-18

[6] Kuzmich R and Masich I 2012 Building a classification model as a composition of informative patterns *Management Systems and Information Technologies* **2(48)** 18-22

[7] Refaeilzadeh P, Tang L and Liu H 2007 On comparison of feature selection algorithms *AAAI Workshop - Technical Report* 05-2007

[8] Masich I S 2006 Approximate algorithms for finding the end-points for the problem of constrained pseudo-Boolean optimization *Vestnik SibSAU* **1(8)** 39-43

[9] Antamoshkin A N and Masich I S 2006 Heuristic search algorithms for monotone pseudo-boolean function conditional optimization *Engineering and automation problems* **1(5)** 55-61

[10] Antamoshkin A N and Masich I S 2007 Pseudo-Boolean optimization in case of unconnected feasible sets *Models and Algorithms for Global Optimization. Series: Springer Optimization and Its Applications vol 4, ed A Törn and J Žilinskas* (New York: Springer) 111-122

[11] Antamoshkin A N, Kuzmich R I and Masich I S 2016 Modified method of logical analysis of data 2016619162 (Moscow: RosPatent)
[12] Vijayarani S and Divya M 2011 An efficient algorithm for generating classification rules *International Journal of Computer Science and Technology* **2(4)** 512-515
[13] Sun B, Chen S, Wang J and Chen H 2016 A robust multi-class AdaBoost algorithm for mislabeled noisy data *Knowledge-Based Systems* **102** 87-102