An iterative feature selection procedure for a classification problem based on the method of logical analysis of data

R I Kuzmich¹, A A Stupina², I S Zhirnova³, O V Slinitsyna⁴ and I I Boubriak⁶

¹Department of Business Informatics and Business Process Modeling, Siberian Federal University, Krasnoyarsk, Russia
²Department of Digital Control Technologies, Siberian Federal University, Krasnoyarsk, Russia
³Department of System Analysis and Operation Research, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
⁴Department of International Management, Krasnoyarsk State Agrarian University, Krasnoyarsk, Russia
⁵Department of Economics and Business Process Management, Siberian Federal University, Krasnoyarsk, Russia
⁶Department of Biochemistry, Oxford University, Oxford, UK

E-mail: h677hm@gmail.com

Abstract. An iterative procedure for selecting features for classifying observations is proposed. The main principles of the proposed iterative procedure are ranking and selection of features according to the frequency of their use when constructing logical patterns based on the method of logical analysis of data. The empirical confirmation of the expediency of this procedure is given.

1. Introduction

The problem of feature selection has been well known for a long time, for a long period of time it has been the main topic of research in the field of machine learning. In [1], a broad overview of the methods used to solve this problem is given. Analysis of literature sources shows that the methods used for feature selection are: principal component method [2], genetic algorithm [3], exhaustive search of heuristics [1], artificial neural networks [4], support vector machine [5].

The presence of too many features in a dataset, as well as the presence of features that do not affect the decision to belong to a certain class, such as experimental noise, can lead to an error in any attempt to extract knowledge from the data.

The paper discusses the method of logical analysis of data, which belongs to the group of inductive learning methods that reveal knowledge from data and represent it in the form of logical patterns [6, 7]. Each logical pattern is represented as a conjunction of features or their negations. Each logical pattern has common properties: coverage (the number of observing class conditions that satisfy its conditions) and degree (the number of terms in a conjunction).

In this paper, we propose an approach for feature selection based on the application of this method. Empirical confirmation of the feasibility of using the method of logical analysis of data for the selection of features is carried out on real problems.
2. Theoretical analysis

The general formulation of the binary classification problem can be formulated as follows. The sample for the problem consists of disjoint sets $\Omega^+$ (a set of observations of a positive class) and $\Omega^-$ (a set of observations of a negative class) of $n$-dimensional vectors. For real problems, vector elements can take different types of values (binary, nominal, quantitative).

Due to the different types of feature values, a procedure for their binarization is required for the correct operation of the method [8]. The next step is the problem of constructing a reference set of features, the essence of which is to select a subset of features $S$ capable of separating with high accuracy the sets of observations of positive and negative classes. The projections $\Omega^+_S$ and $\Omega^-_S$ of the sets $\Omega^+$ and $\Omega^-$ onto $S$ are used in the further work of the method [6].

At the next stage of the method, it is necessary to build optimization models that allow the formation of logical patterns. Let us present optimization models using the example of positive logical patterns. Similar reasoning can be applied to negative logical patterns.

A positive logical $\alpha$-pattern for $\alpha \in \{0,1\}$ is a logical pattern based on observation $\alpha$. The task is to form for each observation $\alpha \in \Omega$, the maximum logical $\alpha$-pattern, that is, capturing the maximum number of observations of a positive class. For this purpose, we define a logical $\alpha$-pattern using binary variables $Y=(y_1,y_2,\ldots,y_t)$:

$$y_k = \begin{cases} 1, & \text{if } k - \text{th feature in rule;} \\ 0, & \text{otherwise.} \end{cases}$$

For pure patterns, coverage of observations of another class is unacceptable. Therefore, for each observation of the negative class $\beta \in \Omega^-$, the variable $y_k$ must be equal to 1, at least for one index $k$ for which $\beta_k \neq \alpha_k$:

$$\sum_{k=1}^t y_k \geq 1 \text{ for any } \beta \in \Omega^-_S,$$

When forming a more stringent restriction, you can change the number 1 to another positive integer number $d$.

To establish the fact of coverage by a logical $\alpha$-pattern of observation of a positive class $\sigma \in \Omega^+$, it is necessary that the value of the variable $y_k$ equals 0 only for indices $k$ that are not fixed in the pattern. Thus, the objective function of the optimization model can be represented as the sum of observations of a positive class captured by the logical $\alpha$-pattern and calculated as:

$$\sum_{\sigma \in \Omega^+} \prod_{j=1}^{t} (1-y_k) .$$

By combining the objective function and the constraint function, we obtain an optimization model for finding logical patterns [9]:

$$\sum_{\sigma \in \Omega^+, j=1}^{t} \prod_{\sigma_k \neq \alpha_k} (1-y_k) \rightarrow \max,$$

$$\sum_{k=1}^t y_k \geq d \text{ for any } \beta \in \Omega^-_S, y \in \{0,1\}^t.$$
of observations of another class) are better suited, the degree of which is less as a rule, and the coverage is greater than that of pure logical patterns.

To form partial logical patterns, the constraint function can be represented in the following form:

$$\sum_{\beta \in \Omega} z_\beta \leq D,$$

where $$D$$ is the number of observations of the opposite class that are allowed to be covered by the logical pattern (positive integer).

After constructing logical patterns, it is necessary to determine the importance of each feature in order to make a decision on the inclusion or exclusion of a feature from the original (complete) set of features for the problem being solved, i.e. implement a feature selection procedure. According to [10], the importance of a feature is the frequency of using a feature in the generated logical patterns, on the basis of which a decision on the classification of new observations is made. Consequently, the number of inclusions of a feature in the formation of logical patterns is directly proportional to its importance. The fewer times a feature is involved in the generated logical patterns, the less important it is when making a decision for a specific problem. Such a feature is a candidate for removal from the initial feature set. When removing such features from the initial set of features, we obtain new feature sets, which are called truncated.

An iterative procedure for obtaining a truncated set of features is proposed.

At the first step, it is necessary to use the initial set of features to build logical patterns, quantify the importance of each feature in the set, and rank features by importance in descending order. Each next step of the procedure for forming a truncated set of features consists in removing from the previous (parent) set of half of the features with the lowest rank.

The classification accuracy of the classifier built on the new truncated set is estimated on the test set:

- If an acceptable quality is found, i.e. if the precision obtained on the new truncated set is approximately equal to the parental precision, the constructed truncated set replaces the previous set of features and the process continues.
- If the classification accuracy is significantly reduced, then the new truncated set is improved by including 75% of the features from the parent feature set.
- If this procedure does not create a truncated set that defines a set of patterns with acceptable quality, then the procedure stops at the parent set of features as the final one.

Also, the criteria for stopping the search for the optimal truncated set of features can be: the restriction set on the minimum number of features in the truncated set established by the researcher, an increase in the number of uncovered observations of the test set.

The main idea of this paper is to determine the feasibility of using the method of logical analysis of data for the problem of feature selection.

3. Results

Experimental research of the proposed procedure for the selection of features are carried out on the problem of predicting the complications of myocardial infarction [11]. As complications, we will choose the prediction of ventricular fibrillation, pulmonary edema.

An optimization model (1,3) was used to find the logical patterns. It is a conditional pseudo-Boolean optimization problem, the solution of which is necessary in the formation of each logical pattern. Based on the fact that the functions of the optimization model (1,3) are monotonic, specially developed optimization algorithms are proposed [12].
In all the presented practical problems 20% of the sample was used to test the classifier, and 80% of the sample was used as a training sample.

3.1. Problem 1. Ventricular fibrillation
The sample includes 70 observations of a positive class and 70 observations of a negative class. The results for all sets of features are presented in table 1.

**Table 1.** Accuracy of classification for the problem of ventricular fibrillation on different sets of features.

| Set number | Number of features in a set | Set of rules | Cover of negative patterns | Cover of positive patterns | Mean degree of rules | Unclassified observation | Classification accuracy, % | Overall accuracy, % |
|------------|-----------------------------|--------------|---------------------------|---------------------------|---------------------|-------------------------|--------------------------|----------------------|
| 1          | 112                         | negative     | 26                        | 5                         | 5                   | 0                       | 90                       | 82.14                |
|            |                              | positive     | 4                         | 26                        | 2                   | 0                       | 77.78                    |                      |
| 2          | 56                          | negative     | 27                        | 5                         | 5                   | 0                       | 90                       | 82.14                |
|            |                              | positive     | 4                         | 26                        | 2                   | 0                       | 77.78                    |                      |
| 3          | 28                          | negative     | 25                        | 5                         | 4                   | 0                       | 90                       | 85.71                |
|            |                              | positive     | 4                         | 22                        | 3                   | 0                       | 83.33                    |                      |
| 4          | 14                          | negative     | 24                        | 5                         | 4                   | 2                       | 70                       | 78.57                |
|            |                              | positive     | 4                         | 20                        | 2                   | 0                       | 83.33                    |                      |
| 5          | 21                          | negative     | 23                        | 5                         | 4                   | 0                       | 80                       | 82.14                |
|            |                              | positive     | 4                         | 23                        | 2                   | 0                       | 83.33                    |                      |

According to the results obtained (table 1), it can be noted that the removal of features with a low importance value allows maintaining high classification accuracy, while reducing the complexity of the problem being solved. The truncated set no 3 should be considered the optimal set of features, since the highest value of the overall classification accuracy was obtained on it. In addition, when switching to the truncated set no 4, unclassified observations appear, which indicates the lack of features for building a classifier with a high generalizing ability.

3.2. Problem 2. Pulmonary edema
The sample includes 181 observations of a positive class and 157 observations of a negative class. The results for all sets of features are presented in table 2.

**Table 2.** Classification results for pulmonary edema on different sets of features.

| Set number | Number of features in a set | Set of rules | Cover of negative patterns | Cover of positive patterns | Mean degree of rules | Unclassified observation | Classification accuracy, % | Overall accuracy, % |
|------------|-----------------------------|--------------|---------------------------|---------------------------|---------------------|-------------------------|--------------------------|----------------------|
| 1          | 112                         | negative     | 38                        | 5                         | 8                   | 0                       | 70                       | 79.41                |
|            |                              | positive     | 5                         | 26                        | 4                   | 0                       | 92.86                    |                      |
| 2          | 56                          | negative     | 41                        | 5                         | 7                   | 0                       | 65                       | 77.94                |
|            |                              | positive     | 5                         | 31                        | 4                   | 0                       | 96.43                    |                      |
| 3          | 28                          | negative     | 34                        | 5                         | 6                   | 0                       | 60                       | 76.47                |
|            |                              | positive     | 5                         | 30                        | 4                   | 0                       | 100                      |                      |
| 4          | 14                          | negative     | 33                        | 5                         | 6                   | 1                       | 47.5                     | 64.71                |
According to the results (table 2), it can be noted that in the transition from the truncated set no 3 to the truncated set no 4, the overall classification accuracy drops sharply, and an unclassified observation appears. These facts are the basis for considering the truncated set no 3 as optimal. Note that the truncated set no 3 does not have the highest accuracy, but the number of features in this set is four times less than in the initial set.

4. Conclusion
In conclusion, it should be noted that the proposed iterative procedure for selecting features based on the method of logical analysis of data is appropriate, since it can significantly reduce the dimension of the problem being solved and at the same time maintain a high quality of classification. The main provisions of this procedure are the assessment of the importance of each feature in the initial set and the formation of truncated sets of the most important features.

In addition, such a given procedure allows you to determine features with zero and maximum importance for any problem being solved, which is undoubtedly useful for a researcher or specialist in this field. For example, such information will help an expert determine the role of a particular feature in constructing logical patterns on the basis of which decisions are made on the classification of new observations.

Acknowledgments
The research was supported financially by a grant from the President of Russian Federation (Project No. MK-1376.2020.9).

References
[1] Dash M and Lui H 1997 Feature selection for classification Intelligent data analysis 1(3) 131-56
[2] Gulin V V 2013 Methods for reducing the dimensionality of the indicative description of documents in the problem of text classification Vestnik of the Moscow Power Engineering Institute 2 115-21
[3] Chtioui Y, Bertrand D and Barba D 1998 Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision Journal of the Science of Food and Agriculture 76(1) 77-86
[4] Setiono R and Lui H 1997 Neural network Feature selection IEEE Transaction on neural networks 8(3) 654-62
[5] Bradley P S and Mangasarian O L 1998 Feature selection via concave minimization and support vector machines Proceedings of the Fifteenth International Conference on Machine Learning 82-90
[6] Hammer P L and Bonates T 2005 Logical Analysis of Data: From Combinatorial Optimization to Medical Applications RUTCOR Research Report 10-2005
[7] Kuzmich R I, Masich I S and Stupina A A 2017 Models of the formation of patterns in the method of logical data analysis Control systems and information technologies 1(67) 33-7
[8] Rastigin L A and Freimanis E E 1988 Solving multi-scale optimization problems by random search methods Problems of random search 11 9-25
[9] Masich I S, Kraeva E M, Kuzmich R I and Gulakova T K 2011 Comparative analysis of data classification methods on practical problems of forecasting and diagnostics Control systems and information technologies 1(43) 20-5
[10] 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

[11] Golovenkin S E, Gorban A N, Shulman V A and Rossiev D A 1997 Complications of myocardial infarction: a database for testing recognition and prediction Preprint 1997.6 (Krasnoyarsk: Inst. of Computational Modeling SD RAS)

[12] 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 4 111-