Application of informative patterns in the classifier for a logical data analysis method development

IOP Conf. Series: Materials Science and Engineering **450** (2018) 052005 doi:10.1088/1757-899X/450/5/052005

Application of informative patterns in the classifier for a logical data analysis method development

R I Kuzmich[1], **A A Stupina**[1,2], **M V Karaseva**[1,2], **L N Ridel**[3] **and T V Dubrovskaya**[3]

[1] Department of Economics and Information Technology of Management, Siberian Federal University, Krasnoyarsk, Russia
[2] Department of System Analysis and Operation Research, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
[3] Department of Economics and Organization in Branches of Chemical and Forest Complex, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

E-mail: h677hm@gmail.com

Abstract. The paper proposes a new method of a classifier for a logical data analysis method development. It is based on the information content of the patterns that forms the obtained classifier. The empirical confirmation of the expediency of this method on the problem of predicting complications of myocardial infarction is given.

1. Introduction

A large number of recognition problems that attract the attention of investigators in medicine can be formulated as follows. There exists a data sample that consists of two disjoint sets Ω^+ and Ω^- of n-dimensional vectors that belong to the positive or negative class, respectively. The vector components, also called attributes, can be both numeric or nominal, and binary. The problem is that it is necessary to determine what class a vector of n variables belongs to for some new observation (which is also a vector of n variables) [1].

The method of data analysis based on the principle of the derivation of logical patterns or rules [2], is investigated for solving this problem. Each rule should cover a lot of observations of one and the same class and practically do not cover the observations of another class. Taking together a number of rules, one can get a classifier that will solve the particular task.

Thus, the main problem in solving classification problems is the development of an adequate classifier that could classify the newly arrived observation, i.e., the observation that was not taking into account in its development.

This paper proposes a method for development a classifier for the logical data analysis method based on the selection of informative patterns for its development, thereby simplifying the original classifier while preserving the quality of the classification as a whole.

2. Theoretical analysis

In the method of logical data analysis, the following procedure for the classifier development for the case of two classes (positive and negative) is proposed [3]:

1) Eliminate redundant variables in the source data sample. For this, a subset S is defined in the set of variables. It makes it possible to distinguish positive and negative observations. Then, Ω_s^+ and Ω_s^- of sets Ω^+ and Ω^- for S are used for this method operation. This procedure is used in different methods of classification and data analysis. A peculiarity of its realization in the method of logical data analysis is that not only significant features are pointed out, but also the combination of features that have a collective impact on the result [4].

2) To develop a pattern that would cover, in addition to this observation, a maximum number of observations of its class and would not cover any observations of another class, or allow the capture of a certain number of observations of another class with respect to each observation in the training sample to increase the total cover patterns.

3) As a result, we get a family of positive and negative maximum patterns combined into a classifier.

To classify a new observation one should apply the following rule:

a) If the observation satisfies the conditions of one or more positive patterns and does not satisfy the conditions of any of the negative, then it is classified as positive.

b) If the observation satisfies the conditions of one or several negative patterns and does not satisfy the conditions of any of the positive, then it is classified as negative.

c) If the observation satisfies the conditions p' of p positive patterns and q' of q negative regulations, then the observation sign is defined as p'/p- q'/q.

d) If the observation does not satisfy the conditions of any pattern, positive or negative, it remains unclassified.

A positive pattern is a sub cube of the space of Boolean variables B_2^t that intersects a set Ω_s^+ and it has some bounded number of common elements with a set Ω_s^-. A negative pattern is given similarly.

Each pattern is characterized by two factors. The first one is coverage, i.e. how many observations of the certain class it covers, and the second one is degree, i.e. how many variables are involved in its formation. As a rule, patterns with a small degree have a greater coverage, i.e. they are more informative and their interpretability improves, i.e. the smaller the variables involved in developing the pattern, the clearer and understandable to the expert is. The pattern should be informative, i.e. it should cover more observations of their class and fewer observations of another class. There exist several criteria of informativeness to measure the informative nature of the pattern [5]. This paper proposes to use the boosting criterion, to evaluate the information content of the pattern:

$$H(p,n)=\sqrt{p}-\sqrt{n}\,, \tag{1}$$

where p is a number of observations of its class that captures the constructed pattern; n is a number of observations of another class that captures the developed pattern.

As it was mentioned earlier, a classifier included all the patterns that were developed regarding to each observation of the training sample. As a result, if the volume of the training sample increases, then the size of the classifier also increases. It should be noted that the developed patterns are characterized by different informativeness. The patterns covering a small number of observations are not statistically reliable because there are too many patterns among them. They allow more errors on independent control data than on a training sample. Therefore, it is proposed to form a classifier only from informative patterns, i.e. theit informativeness is higher than a certain information threshold (H_0), provided by an investigator. As a result, this will lead to the reduction in the number of rules in the classifier without losing the quality of the classification or with a slight change in the positive or negative level.

While solving the given task a problem of the informativeness threshold selection arises. To solve this problem, the paper proposes the following iterative procedure. At the first step, the informativeness threshold should be chosen equal to zero for the positive and negative sets of patterns, thus we obtain the initial classifier consisting of the maximum number of patterns. At the next step of the procedure, it is suggested to select the informativeness threshold for positive (negative) patterns is

equal to the value of the mean informative value (H_{mean}) for all positive (negative) patterns of the classifier:

$$H_{mean} = \frac{1}{q} \cdot \sum_{i=1}^{q} H_i \ ,$$

where H_i is the informativeness of the positive i-*th* pattern calculated according to the formula (1).

For a negative set of patterns, the mean informativeness value is calculated similarly.

Having obtained a new classifier consisting of more informative patterns and calculating the mean informativeness values for negative and positive patterns of the given classifier, we will use the obtained values of the mean informativeness to develop a subsequent classifier consisting of patterns whose informativeness exceeds the mean informativeness value of the previous classifier. Thus, we construct each subsequent classifier using the mean informativeness value of the previous one. In this case, a number of patterns reduces, and the mean informativeness content increases for each subsequent classifier. A condition for its stopping is one should consider a moment of increasing a number of unclassified (uncovered) observations while classification, i.e. the patterns included in the previous classifier do not cover some of the observations included into the examining sample. Therefore, it is necessary to return either to the previous classifier by changing the value of the two informativeness thresholds to their previous values, or to try to change the value of only one informativeness threshold by positive (negative) patterns and observe how this change will affect a number of unclassified observations and the results of the classification as a whole.

3. Results

We perform a series of experiments on the problem of complicating myocardial infarction [6], considering the prediction of the following complications: ventricle fibrillation, atrium fibrillation, cardiac rupture. For this, a sample of data consisting of 1,700 observations is used; the information is contained in 116 features.

Due to the rather large volume of initial data, smaller samples were used for carrying out numerous series of experiments in an acceptable time. For each of the three solved tasks, their data samples were formed. At the same time, all positive observations on a specific task were included (according to the predicted complication), since they were initially presented significantly less, and negative observations (without complication) were selected randomly from the primary sample.

To find the rules, we used a modified optimization model. It helps the rules to cover a limited number of observations of another class to "mitigate" possible inaccuracies and errors in the data. To solve the optimization problem, optimization algorithms based on the search for boundary points of an admissible region were applied [7-8]. These algorithms were developed specifically for this class of problems and they are based on the behavior of monotonic functions of the optimization model in the space of Boolean variables. The algorithms for finding boundary points are search ones, i.e. they do not require the specification of functions in explicit form, using algebraic expressions, but they use the computation of functions at points [9].

In all the problems considered below, 20% of the original sample were used to check the classifier and did not participate in its construction.

The classification results for the three complications of myocardial infarction obtained with the help of the developed software system are given below [10].

3.1. Problem1. Ventricle fibrillation

A sample of 70 patients with complications (positive observations) and 70 patients without complications (negative observations) was used for the tests. The test results are shown in table 1.

Table 1. Classification results for the problem of ventricle fibrillation with a change in the informativeness threshold, H_0

Set of rules	Number of rules	Mean informativeness, H_{mean}	Informativeness threshold, H_0	Cover of negative patterns	Cover of positive patterns	Number of uncovered observations	Classification accuracy, %
negative	56	2,86	0	26	5	0	93
positive	56	3,47	0	5	39		86
negative	36	3,26	2,86	30	5	0	93
positive	33	4,17	3,47	5	40		86
negative	16	3,52	3,26	33	5	2	100
positive	24	4,32	4,17	5	42		71
negative	16	3,52	3,26	33	5	0	100
positive	33	4,17	3,47	5	40		86

3.2. Problem 2. Atrium fibrillation

A sample of data consisting of 169 positive observations and 169 negative observations was used for the tests. The test results are shown in table 2.

Table 2. Classification results for the problem of atrium fibrillation with a change in the informativeness threshold, H_0

Set of rules	Number of rules	Mean informativeness, H_{mean}	Informativeness threshold, H_0	Cover of negative patterns	Cover of positive patterns	Number of uncovered observations	Classification accuracy, %
negative	137	3,15	0	50	15	0	81
positive	133	2,72	0	15	43		78
negative	77	3,67	3,15	57	15	0	78
positive	69	3,23	2,72	15	50		75
negative	38	3,94	3,67	61	15	0	69
positive	26	3,61	3,23	15	55		83
negative	15	4,21	3,94	66	15	4	72
positive	10	4	3,61	15	60		69
negative	38	3,94	3,67	61	15	1	69
positive	10	4	3,61	15	60		75
negative	15	4,21	3,94	66	15	1	72
positive	26	3,61	3,23	15	55		75

3.3. Problem 3. Cardiac rupture

A sample of data consisting of 54 positive observations and 54 negative observations was used for the tests. The test results are shown in table 3.

Table 3. Classification results for the cardiac rupture problem with a change in the informativeness threshold, H_0

Set of rules	Number of rules	Mean informativeness, H_{mean}	Informativeness threshold, H_0	Cover of negative patterns	Cover of positive patterns	Number of uncovered observations	Classification accuracy, %
negative	44	3,33	0	30	5	0	70
positive	42	3,44	0	5	30		83
negative	29	3,52	3,33	32	5	0	70
positive	27	3,76	3,44	5	32		83
negative	10	3,72	3,52	33	5	0	70
positive	14	3,93	3,76	5	33		92

According to the obtained results, it can be noted that the classifier development as a composition of informative patterns makes it simpler, since the number of patterns that forms the classifier is reduced in 2-4 times regarding to the complete set of patterns for a particular task. At the same time, the accuracy of the classification either does not decrease or decreases insignificantly. In some cases, with the removal of less informative patterns from the classifier, the accuracy of the classification increases. This is due to the fact that these patterns, being statistically unreliable, covered the observations of the testing sample, i.e. they participated in the classifier along with informative patterns, thereby increasing the classification error.

References

[1] Golovenkin S E, Gulakova T K, Kuzmich R I, Masich I S and Shulman V A 2010 Model of logical analysis for solving the problem of predicting myocardial infarction *Vestnik SibSAU* **4(30)** 68-73
[2] Alexe S, Blackstone E. and Hammer P L 2003 Coronary Risk Prediction by Logical Analysis of Data *Annals of Operations Research* **119** 15-42
[3] Hammer P L and Bonates T 2005 Logical Analysis of Data: From Combinatorial Optimization to Medical Applications *RUTCOR4*
[4] *Research Report* 10-2005
[5] Hammer P L, Kogan A and Lejeune M 2004 Modeling Country Risk Ratings Using Partial Orders *RUTCOR Research Report* 24-2004
[6] Furnkranz J, Flach P A 2005 Roc 'n' rule learning-towards a better understanding of covering algorithms *Machine Learning* **58(1)** 39–77
[7] 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)
[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)