

Selection of logical patterns for constructing a decision rule of recognition

A N Antamoshkin^{1,2} and I S Masich^{1,3}

¹ Department of System Analysis and Operation Research, Siberian State Aerospace University named after academician M. F. Reshetnev, Krasnoyarsk, Russia

² Department of Economics and Information Technology of Management, Siberian Federal University, Krasnoyarsk, Russia

³ Department of Informatics and Mathematical Modeling, Krasnoyarsk State Agrarian University, Krasnoyarsk, Russia

E-mail: i-masich@yandex.ru

Abstract. We investigate an aspect of the construction of logical recognition algorithms – selection of patterns in the set of found patterns in the data.

We consider the recognition problem for objects described by binary attributes and divided into two classes. A result of performance the procedure of searching patterns on the training set (a set of input data) is a number of patterns found. The question is to select some patterns from their total number to form a decision rule. That can not only reduce size of the decision rule, but also to improve recognition.

One way to make a selection of patterns is select a subset of patterns that are needed to cover all objects of the training sample. This problem is formulated as an optimization problem. The resulting optimization model represents a problem of conditional pseudo-Boolean optimization, in which the objective function and the constraints functions are unimodal monotone pseudo-Boolean functions.

Another way is to make the selection of such patterns, which when used together will increase separating capacity of the decision rule. As a criterion for the formation of the decision rule is considered the width of the separation margin. One more way is the selection supporting objects, rules are formed on the basis of which.

Selection of logical patterns, which is made in accordance with the proposed approach, can significantly reduce the number of patterns and simplify the decision rule, almost without compromising the accuracy of recognition. This makes the decision rule clearer, and the results more interpretable. It is necessary to support decision making for recognition.

Statement of the pattern selection problem

Currently pretty effective classification algorithms are developed for the solution of problems of diagnostics and prediction, and ably tuning these algorithms solve the problems with high accuracy. But in the practical application of such algorithms, we often are interested in the question of evidence and interpretability of results. For decision making a model in



explicit form is required, such a model in which the calculated decisions are justified and based on available data. In this paper we construct a model for decision making which consists of a set of logical rules that describe patterns in a studied phenomenon or a system. The main task is to identify these patterns and lead to a form in which they will be used to construct a decision making model. Identification of pattern on the basis the available set of data is a complex computational task that requires efficient algorithmic support and software implementation. Even so constructed classifiers can effectively solve practical problems [1, 2], including those in the aerospace industry [3, 4].

The process of formation of decision rules is accompanied by solving problems of selection of the best alternatives in accordance with some criterion. Formalization of this process in a number of problems in combinatorial optimization generates flexible and efficient algorithm for logical analysis for data classification.

Let us consider the problem of recognition of objects described by binary attributes and divided into two classes $K = K^+ \cup K^- \subset \{0,1\}^n$. An object (or observation) $X \in K$ is describe by binary vector $X = (x_1, x_2, \dots, x_n)$ and can be can be represented as a point in the hypercube of binary attribute space B_2^n .

A term which covers at least one observation of a certain class and does not cover any observation of another class is called a pattern or a rule. That is, the pattern corresponds to the cube having non-empty intersection with one of the sets (K^+ or K^-) and an empty intersection with another set (K^- or K^+ respectively). A pattern P , which is disjoint with K^- , will be called positive, and a pattern P' , which is disjoint with K^+ , will be called negative. Patterns are the elementary building blocks for rule-based algorithms.

Assume that as a result of the patterns search procedures on the training set we found some positive patterns $P_i, i=1, \dots, p$, and some negative patterns $N_j, j=1, \dots, n$.

The discriminant (decision) function can be given by the expression

$$D(a) = \frac{1}{p} \sum_{i=1}^p P_i(a) - \frac{1}{n} \sum_{j=1}^n N_j(a)$$

for some observation a ,

where $P_i(a)=1$ if the pattern P_i cover the observation a and $P_i(a)=0$ in another case. The same goes for $N_j(a)$.

The algorithms for finding patterns are described in [5]. In particular, that is the algorithms that search patterns by basing on some observation of the training sample. Therefore, as a result of such search a large number of patterns can be recorded, up to the number of training observations. Some of these patterns, however, may coincide. In solving many problems there is a question of selection patterns from the set of all patterns to form a decision rule. That can not only reduce the size of classifier, in some cases that can improve recognition.

The main advantage that the rule-based algorithms provide for solving real-world problems is clarity of the process of recognition on the resulting model for new observations. All identified patterns are presented explicitly. However if the number of these patterns is large (more than 10-15) then the recognition algorithm becomes difficult to interpret. In this regard, let us explore some methods of selection from the total number of patterns discovered.

Minimizing the number of patterns

Let us introduce the variables that determine whether the pattern presents in the decision function.

$$x_i = \begin{cases} 1, P_i \text{ presents in the decision function,} \\ 0, \text{ in another case.} \end{cases}$$

$$y_j = \begin{cases} 1, N_j \text{ presents in the decision function,} \\ 0, \text{ in another case.} \end{cases}$$

One way to make a selection of patterns is select a subset of patterns, which are needed to cover all the training observations [6]. Thus each object of the training sample must be covered by at least one pattern. Using the variables introduced above, this condition can be written as

$$\sum_{i=1}^p x_i P_i(a) \geq 1 \text{ for any } a \in K^+,$$

$$\sum_{j=1}^n y_j N_j(a) \geq 1 \text{ for any } a \in K^-.$$

To improve the robustness of the algorithm the number 1 in the right side of inequalities should be replaced by a positive integer d . In this case, each object of the training sample must be covered by d patterns.

Thus, we have the following problem of minimizing the number of patterns used in the decision rule

$$\sum_{i=1}^p x_i + \sum_{j=1}^n y_j \rightarrow \min$$

subject to

$$\sum_{i=1}^p x_i P_i(a) \geq d \text{ for any } a \in K^+,$$

$$\sum_{j=1}^n y_j N_j(a) \geq d \text{ for any } a \in K^-.$$

The obtained optimization model represents a problem of conditional pseudo-Boolean optimization, in which the objective function and the constraint functions are unimodal and monotone pseudo-Boolean functions [7]. To solve the problem we use the approximation algorithms for the conditional pseudo-Boolean optimization, based on search for the optimal solution among the boundary points of the feasible region [8].

In order to evaluate the effect of reducing the number of patterns in the final rule on recognition accuracy, a series of experiments was held on recognition and prediction

problems. Patterns search was made on the basis of the optimization model [9], which allows to find the maximum patterns, that is patterns with the highest coverage of observation of a certain class. Each data sample has been divided into two parts: a learning sample and a test sample. On the basis of each observation of the training sample a pattern was found. A comparison of quality recognition decision rules is made. One decision rule is built from the complete set of patterns and another rule from a reduced set obtained by solving the optimization problem described above.

In the experiments we used the following recognition problems [10]:

breast-cancer – the problem of breast cancer diagnosis, sample size: 699 observation described by 9 multi-type attributes (80 binary attributes obtained by binarization);

wdbc – the problem of breast cancer diagnosis, sample size: 569 observation described by 30 multi-type attributes (120 binary attributes);

hepatitis – the problem of hepatitis diagnosis, sample size: 155 observation described by 19 multi-type attributes (37 binary attributes);

spect – data on cardiac computed tomography, sample size: 80 observation described by 22 binary attributes.

The results are shown in Table 1.

Table 1. Recognition results

Recognition problem	Set of patterns	Number of positive patterns	Number of negative patterns	Recognition accuracy for positive observations	Recognition accuracy for negative observations
breast-cancer	complete set	419	209	0.97	0.91
	reduced set	12	14	0.97	0.88
wdbc	complete set	291	163	0.94	0.98
	reduced set	9	11	0.92	0.96
hepatitis	complete set	27	97	0.8	0.85
	reduced set	7	7	0.8	0.81
spect	complete set	38	34	1	0.83
	reduced set	7	8	1	0.83

As can be seen from the results, the application of a decision rule based on a reduced set of patterns in some problems leads to a slight decrease in recognition quality. But at the same time this accompanied by a significant decrease in the number of patterns to be used for decision making, and that positively affects to clarity of the decision obtained.

Maximizing the separation margin

Another method is to make a selection of such patterns, which when used together will increase the separating ability of the decision rule.

As a criterion for the formation of the decision rule, consider the width of the "separating margin".

$$\min\{D(a) : a \in K^+\} - \max\{D(a) : a \in K^-\},$$

$$\text{where } D(a) = \frac{1}{p} \sum_{i=1}^p P_i(a) - \frac{1}{n} \sum_{j=1}^n N_j(a) \text{ for some observation } a.$$

We take into account the presence of outliers that may be present in real-world problems. For this purpose we introduce a variable

$$z^a = \begin{cases} 1, & a \text{ is taken as an outlier,} \\ 0, & \text{in another case.} \end{cases}$$

Then the problem of pattern selection can be written as follows.

$$\begin{aligned} v^+ + v^- - C \sum_{a \in K} z^a \cdot |b^a| &\rightarrow \max, \\ \text{where } v^+ &= \min\{D'(a) : a \in K^+, z^a = 0\}, \\ v^- &= \min\{-D'(a) : a \in K^-, z^a = 0\}, \\ D'(a) &= \frac{\sum_{i=1}^p x_i P_i(a)}{\sum_{i=1}^p x_i} - \frac{\sum_{j=1}^n y_j N_j(a)}{\sum_{j=1}^n y_j}, \\ b^a &= \begin{cases} v^+ - D'(a), & a \in K^+, \\ v^- + D'(a), & a \in K^-. \end{cases} \end{aligned}$$

Algorithms for solving such optimization problems are given in [11].

Decomposition of training sample in identifying patterns

Considered in [12, 13] methods to search for patterns suggest use as a "support" point an observation of the training sample (precedent), a partial overlapping of the properties of which can be found in other observations of the same class. The method described above requires to use a large number of support observations (possibly all observations of the training sample) to obtain patterns, and then to carry out the selection from the found patterns.

Consider another method comprising selection of these support objects. The entire set of observations of training sample for a certain class, for example K^+ , can be divided into groups of observations so that the observations were quite similar in each group:

$$K^+ = K_1^+ \cup K_2^+ \cup \dots \cup K_k^+$$

For this purpose, it can be used the k-means algorithm, as a result of which we obtain a set of centroids c_1, c_2, \dots, c_k , so that the following rule holds:

$$a \in K_j^+, \text{ if } \|a - c_j\| < \|a - c_i\|$$

For all $i = 1, 2, \dots, k, i \neq j$,

where K_j^+ is the set of observations included in the cluster with the centroid c_j .

These centroids can be used as support observations to identify logical patterns.

Described approach can significantly reduce the complexity of the logic recognition algorithm, producing a selection of observations used as support observations when searching for patterns.

Let us consider the results of using this approach in relation to the problem of predicting myocardial infarction complications: atrial fibrillation (AF) and ventricular fibrillation (VF) [14]. To find centroids the k-means algorithm was used in software applications Weka [15]. To search for patterns and estimate the accuracy of the constructed decision rule the software of authors was used.

The sample for the AF problems consisted of 184 positive and 184 negative observation described by 112 multi-type attributes. The number of binarized attributes was 215. For each class 15 centroids was allocated, which were used to search for patterns.

The sample for the VF problems consisted of 80 positive and 80 negative observation described by 112 multi-type attributes. The number of binarized attributes was 200. For each class 10 centroids was allocated, which were used to search for patterns.

10% of observations of the sample were allocated for testing the resulting decision rule. Recognition results are shown in Table 2.

Table 2. Comparison of results of recognition

Recognition problem	Set or patterns	Number of positive patterns	Number of negative patterns	Recognition accuracy for positive observations	Recognition accuracy for negative observations
AF	complete set	165	165	0.7	0.79
	reduced set	15	15	0.68	0.77

VF	complete set	72	72	0.87	0.71
	reduced set	10	10	0.9	0.88

As a result of the decomposition of observations of training sample and the corresponding selection of observations used as support observations for finding patterns, we obtain simplification of the decision rule - the number of patterns used in the final rule decreases in 7-10 times. At the same time for some problems we even observe increasing in accuracy of recognition of test observations.

Conclusion

To summarize, it must be concluded that the selection of logical patterns, produced in accordance with certain criteria, can significantly reduce the number of pattern used in the classifier and simplify the decision rule, only slightly reducing the accuracy of recognition. In solving a number of real-world problems of recognition and prediction, great importance is the interpretability of the obtained solutions and the ability to justify them on the basis of rules and patterns, which, in its turn, are based on precedents in the form of observations in the data sample. Therefore, the use approaches described in this paper is useful for solving such problems.

Acknowledgments

The results were obtained within the project No. 346 of the state task No. 2014/211 of the Ministry of Education and Science of the Russian Federation.

References

- [1] Alexe G, Alexe S, Hammer P L, Liotta L, Petricoin E and Reiss M 2004 Ovarian Cancer Detection by Logical Analysis of Proteomic Data *Proteomics* **4(3)** 766–783
- [2] Axelrod D, Bonates T, Hammer P L and Lozina I 2004 From Diagnosis to Therapy via LAD *Invited Lecture at INFORMS Annual Meeting* (Denver: CO)
- [3] Dupuis C, Gamache M and P  ge J F 2012 Logical analysis of data for estimating passenger show rates in the airline industry *Journal of Air Transport Management* **18** 78–81
- [4] Esmacili S 2012 Development of equipment failure prognostic model based on logical analysis of data *Master of Applied Science Thesis, Dalhousie University, Halifax, Nova Scotia, July 2012*
- [5] Masich I S 2009 Combinatorial Optimization in Classification Problem *Control Systems and Information Technology* **1.2(35)** 283-288
- [6] Hammer P L and Bonates T O 2006 Logical analysis of data - An overview: From combinatorial optimization to medical applications *Annals of Operations Research* **148**
- [7] Antamoshkin A N and Masich I S 2007 Pseudo-Boolean optimization in case of unconnected feasible sets *Models and Algorithms for Global Optimization* vol 4 of the series *Optimization and Its Applications* ed A Torn and J Zilinskas (Springer) pp 111-122
- [8] Antamoshkin A N and Masich I S 2003 Efficient algorithms for constrained optimization of monotone pseudo-Boolean functions *Vestnik SibSAU* **4** 60-67
- [9] 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

- [10] Lichman M 2013 UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>] Irvine, CA: University of California, School of Information and Computer Science
- [11] Masich I S 2006 The heuristic algorithms of boundary points search for an constraint pseudo-boolean optimization problem *Vestnik SibSAU* **1(8)** 39-43
- [12] Alexe G and Hammer P L 2006 Spanned patterns for the logical analysis of data *Discrete Applied Mathematics* **154** 1039–1049
- [13] Guoa C and Ryoo H S 2012 Compact MILP models for optimal and Pareto-optimal LAD patterns *Discrete Applied Mathematics* **160** 2339–2348
- [14] Golovenkin S E, Gorban A N and Shulman V A 1997 Complications of myocardial infarction: a database for testing recognition and prediction *Preprint* 1997.6 (Krasnoyarsk: Inst. of Computational Modeling SD RAS)
- [15] Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java <http://www.cs.waikato.ac.nz/ml/weka>