УДК 519.854.33

Вестник СибГАУ
2014. № 5(57). С. 20–25

ВЫБОР ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ ДЛЯ ПОСТРОЕНИЯ РЕШАЮЩЕГО ПРАВИЛА РАСПОЗНАВАНИЯ

А. Н. Антамошкин, И. С. Масич

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: i-masich@yandex.ru

Исследуется один из аспектов построения логических алгоритмов распознавания – отбор закономерностей из множества найденных закономерностей в данных.

Рассматривается задача распознавания объектов, описываемых бинарными признаками и разделенных на два класса. В результате выполнения процедуры поиска закономерностей по обучающей выборке (набору исходных данных) найден ряд закономерностей. Встает вопрос отбора закономерностей из общего их числа для формирования решающего правила, что способно не только уменьшить его размер, но и повысить качество распознавания.

Один из способов произвести отбор закономерностей – выделить подмножество закономерностей, которые необходимы для покрытия всех объектов обучающей выборки. Эта задача формулируется в виде задачи оптимизации. Полученная оптимизационная модель представляет собой задачу условной псевдобулевой оптимизации, в которой целевая функция и функции в ограничениях являются унимодальными монотонными псевдобулевыми функциями.

Другой способ заключается в том, чтобы произвести отбор таких закономерностей, которые при совместном использовании увеличат разделяющую способность решающего правила. В качестве критерия при формировании решающего правила рассматривается ширина «разделяющей полосы». Еще один способ заключается в отборе опорных объектов, на основе которых формируются правила.

Отбор логических закономерностей, произведенный в соответствии с предлагаемым подходом, позволяет значительно снизить их число и упростить решающее правило, практически не снижая точность распознавания. Это делает решающее правило прозрачным, а результаты более интерпретируемыми, что необходимо для поддержки принятия решений при распознавании.

Ключевые слова: анализ данных, классификация, логический алгоритм, распознавание.

Vestnik SibGAU
2014, No. 5(57), P. 20–25

THE SELECTION OF LOGICAL PATTERNS FOR CONSTRUCTING A DECISION RULE OF RECOGNITION

A. N. Antamoshkin, I. S. Masich

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: i-masich@yandex.ru

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. In consequence of performance the procedure of searching patterns on the training set (a set of input data) a number of patterns has been found. The question is to select some patterns from their total number to form a decision rule. That can not only reduce the size of the decision rule, but also improve recognition.

One way to make a selection of patterns is to select a subset of patterns that is 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 to select supporting objects and on their basis to form the rules.

The 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.

Keywords: analysis of data, classification, logical algorithm, recognition.

Постановка задачи отбора закономерностей. К настоящему времени разработаны довольно эффективные алгоритмы классификации для решения задач диагностики и прогнозирования, которые при умелой настройке решают задачи с большой точностью. Но при практическом применении таких алгоритмов зачастую встает вопрос об интерпретируемости и доказательности результатов. Для принятия решений требуется модель в явном виде, такая модель, в которой вычисляемые решения обоснованы и опираются на имеющиеся данные. В данной работе строится модель принятия решений, состоящая из набора логических правил, которые описывают закономерности в исследуемом явлении или системе. Основная задача – выявить эти закономерности и привести к виду, в котором они будут использованы для построения модели принятия решений. Выявление закономерностей на основе имеющегося набора данных является сложной вычислительной задачей, требующей эффективного алгоритмического обеспечения и его программной реализации. Но получаемые классификаторы способны эффективно решать практические задачи [1; 2], в том числе в аэрокосмической отрасли [3; 4].

Процесс формирования решающих правил сопровождается решением задач выбора наилучших альтернатив в соответствии с некоторым критерием. Формализация этого процесса в виде ряда задач комбинаторной оптимизации формирует гибкий и эффективный алгоритм логического анализа для классификации данных.

Рассмотрим задачу распознавания объектов, описываемых бинарными признаками и разделенных на два класса $K = K^+ \cup K^- \subset \{0,1\}^n$. Объект $X \in K$ описывается бинарным вектором $X = (x_1, x_2, \ldots, x_n)$ и может быть представлен как точка в гиперкубе пространства бинарных признаков B_2^n.

Под *закономерностью* P (или *правилом*) понимается терм, который покрывает хотя бы один объект некоторого класса и не покрывает ни одного объекта другого класса. То есть закономерность соответствует подкубу, имеющему непустое пересечение с одним из множеств (K^+ или K^-) и пустое пересечение с другим множеством (K^- или K^+ соответственно). Закономерность P, которая не пересекается с K^-, будем называть положительной, а закономерность P', которая не пересекается с K^+ – отрицательной. Закономерности являются элементарными блоками для построения логических алгоритмов распознавания.

Предположим, что в результате выполнения процедуры поиска закономерностей по обучающей выборке найден ряд положительных закономерностей P_i, $i = 1, \ldots, p$, и отрицательных закономерностей N_j, $j = 1, \ldots, n$.

Решающая функция может быть задана выражением

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

для некоторого объекта a, где $P_i(a) = 1$, если закономерность P_i покрывает объект a, и $P_i(a) = 0$ в противном случае. То же самое для $N_j(a)$.

В [5] описаны алгоритмы поиска закономерностей. В частности, это алгоритмы, которые ведут поиск закономерности, опираясь на некоторый объект обучающей выборки. Поэтому в результате их работы может быть записано большое число закономерностей, вплоть до числа объектов обучающей выборки, некоторые из которых, впрочем, могут повторяться. При решении многих задач встает вопрос отбора закономерностей из общего их числа для формирования решающего правила, что способно не только уменьшить его размер, но в некоторых случаях и улучшить распознавание.

Основное преимущество, которое предоставляют логические алгоритмы распознавания при решении практических задач, – это прозрачность процесса распознавания новых объектов по полученной модели. Все выявленные закономерности представлены в явном виде. Но если этих закономерностей много, более, допустим, 10–15, то алгоритм распознавания становится трудно интерпретируемым. В связи с этим исследуем некоторые способы отбора из общего числа найденных закономерностей.

Минимизация числа закономерностей. Введем переменные, определяющие, будет ли закономерность присутствовать в решающей функции:

$$x_i = \begin{cases} 1, P_i \text{ присутствует в решающей функции,} \\ 0, \text{ в противном случае;} \end{cases}$$

$$y_j = \begin{cases} 1, N_j \text{ присутствует в решающей функции,} \\ 0, \text{ в противном случае.} \end{cases}$$

Один из способов произвести отбор закономерностей – выделить подмножество закономерностей, которые необходимы для покрытия всех объектов обучающей выборки [6]. Каждый объект обучающей выборки должен при этом покрываться хотя бы одной закономерностью. Используя введенные переменные, это условие можно записать в виде

$$\sum_{i=1}^{p} x_i P_i(a) \ge 1 \text{ для любого } a \in K^+,$$

$$\sum_{j=1}^{n} y_j N_j(a) \ge 1 \text{ для любого } a \in K^-.$$

Для повышения робастности алгоритма число 1 в правой части неравенств следует заменить целым положительным числом *d*. В таком случае каждый объект обучающей выборки должен подчиняться заданному количеству *d* закономерностей.

Таким образом, имеем следующую задачу минимизации числа используемых в решающем правиле закономерностей:

$$\sum_{i=1}^{p} x_i + \sum_{j=1}^{q} y_j \to \min$$

при ограничениях на переменные:

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

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

Полученная оптимизационная модель представляет собой задачу условной псевдобулевой оптимизации, в которой целевая функция и функции в ограничениях являются унимодальными монотонными псевдобулевыми функциями [7]. Для решения задачи использовались приближенные алгоритмы условной псевдобулевой оптимизации, основанные на поиске оптимального решения среди граничных точек допустимой области [8].

Для того чтобы оценить, как влияет уменьшение числа закономерностей в решающем правиле на точность распознавания, была проведена серия экспериментов на задачах распознавания и прогнозирования. Поиск закономерностей производился на основе оптимизационной модели [9], позволяющей находить максимальные закономерности, т. е. закономерности с наибольшим покрытием объектов некоторого класса. Каждая выборка данных была разделена на две части – обучающую и тестовую. На основе каждого объекта обучающей выборки производился поиск закономерности. Проводилось сравнение качества распознавания решающих правил, построенных из полного набора закономерностей и из уменьшенного набора, полученного путем решения описанной выше оптимизационной задачи.

При проведении экспериментов использовались следующие задачи распознавания [10]:

– breast-cancer – задача диагностики рака молочной железы, объем выборки – 699 объектов, описываемых 9 разнотипными признаками (в результате бинаризации получено 80 бинарных признаков);

– wdbc – задача диагностики рака молочной железы, объем выборки – 569 объектов, описываемых 30 разнотипными признаками (120 бинарных признаков);

– hepatitis – задача диагностики наследственного гепатита, объем выборки – 155 объектов, описываемых 19 разнотипными признаками (37 бинарных признаков);

– spect – данные по сердечной компьютерной томографии объем выборки – 80 объектов, описываемых 22 бинарными признаками.

Результаты экспериментов приведены в табл. 1. Как видно, применение решающего правила, основанного на уменьшенном наборе закономерностей, в некоторых задачах приводит к незначительному снижению качества распознавания, но в то же время сопровождается значительным снижением числа закономерностей, которые необходимо использовать для принятия решения, что положительно сказывается на прозрачности получаемых решений.

Максимизация разделяющей полосы. Еще один способ заключается в том, чтобы произвести отбор таких закономерностей, которые при совместном использовании увеличат разделяющую способность решающего правила.

Таблица 1

Результаты распознавания

Задача распознавания	Набор закономерностей	Число положительных закономерностей	Число отрицательных закономерностей	Точность распознавания положительных объектов	Точность распознавания отрицательных объектов
Breast-cancer	Полный набор	419	209	0,97	0,91
	Уменьшенный набор	12	14	0,97	0,88
Wdbc	Полный набор	291	163	0,94	0,98
	Уменьшенный набор	9	11	0,92	0,96
Hepatitus	Полный набор	27	97	0,8	0,85
	Уменьшенный набор	7	7	0,8	0,81
Spect	Полный набор	38	34	1	0,83
	Уменьшенный набор	7	8	1	0,83

В качестве критерия при формировании решающего правила рассмотрим ширину «разделяющей полосы»:

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

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

Учтем наличие выбросов, которые могут присутствовать в реальных задачах. Для этого введем переменную

$$z^a = \begin{cases} 1, a \text{ принимается за выброс,} \\ 0, \text{ в противном случае.} \end{cases}$$

Тогда задачу отбора закономерностей можно записать в следующем виде:

$$v^+ + v^- - C\sum_{a \in K} z^a \cdot \left|b^a\right| \to \max,$$

где $v^+ = \min\{D'(a): a \in K^+, z^a = 0\}$,

$$v^- = \min\{-D'(a): a \in K^-, z^a = 0\},$$

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

$$b^a = \begin{cases} v^+ - D'(a), a \in K^+, \\ v^- + D'(a), a \in K^-. \end{cases}$$

Алгоритмы для решения таких задач оптимизации приведены в [11].

Декомпозиция обучающей выборки при выявлении закономерностей. Рассматриваемые в работах [12; 13] способы поиска закономерностей предполагают использование в качестве «опорной точки» объект обучающей выборки (прецедент), частичное повторение свойств которого может быть обнаружено в других объектах этого же класса. Описанный выше способ предписывает использовать большое число таких опорных объектов (возможно, всех объектов обучающей выборки) для получения закономерностей, а затем проводить отбор из найденных.

Рассмотрим другой способ, заключающийся в отборе самих этих опорных объектов. Всё множество объектов обучающей выборки некоторого класса, скажем K^+, можно разбить на группы объектов так, чтобы объекты были схожи внутри каждой группы:

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

Для этого можно использовать алгоритм *k*-средних, в результате работы которого получаем набор центроидов $c_1, c_2, ..., c_k$ так, что будет выполняться правило

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

для всех $i = 1, 2, ..., k$, $i \neq j$,

где K_j^+ – множество объектов, входящих в кластер с центроидом c_j.

Эти центроиды можно использовать в качестве опорных объектов для выявления логических закономерностей.

Описанный подход позволяет существенно снизить трудоемкость работы логического алгоритма распознавания, производя отбор объектов, используемых в качестве опорных при поиске закономерностей.

Рассмотрим результаты использования этого подхода применительно к задаче прогнозирования осложнений инфаркта миокарда: фибрилляции предсердий (ФП) и фибрилляции желудочков (ФЖ) [14]. Для нахождения центроидов использовался алгоритм *k*-средних программного приложения Weka [15], для поиска закономерностей и оценки точности построенного решающего правила использовалось авторское программное обеспечение.

Выборка для задачи ФП состояла из 184 положительных и 184 отрицательных объектов, описываемых 112 разнотипными признаками. Число бинаризованных признаков составило 215. Для каждого класса выделено по 15 центроидов, которые использовались для поиска закономерностей.

Выборка для задачи ФЖ состояла из 80 положительных и 80 отрицательных объектов, описываемых 112 разнотипными признаками. Число бинаризованных признаков составило 200. Для каждого класса выделено по 10 центроидов, которые использовались для поиска закономерностей.

10 % объектов выборки было выделено для тестирования полученного решающего правила. Результаты распознавания приведены в табл. 2.

В результате использования декомпозиции объектов обучающей выборки и соответствующего отбора объектов, используемых в качестве опорных для поиска закономерностей, получаем упрощение решающего правила – число используемых в решающем правиле закономерностей уменьшается в 7–10 раз. При этом для некоторых задач наблюдается даже увеличение точности распознавания тестовых объектов.

Заключение. Подводя итог, следует заключить, что отбор логических закономерностей, произведенный в соответствии с некоторым критерием, позволяет значительно снизить их число и упростить решающее правило, лишь немного снижая точность распознавания. При решении ряда практических задач распознавания и прогнозирования большое значение имеет интерпретируемость получаемых решений и возможность их обосновать, опираясь на правила и закономерности, которые, в свою очередь, основаны на прецедентах в виде объектов выборки данных. Поэтому использование описанных в этой работе подходов представляется полезным для решения таких задач.

Таблица 2

Сравнение результатов распознавания

Задача распознавания	Набор закономерностей	Число положительных закономерностей	Число отрицательных закономерностей	Точность распознавания положительных объектов	Точность распознавания отрицательных объектов
ФП	Полный набор	165	165	0,7	0,79
	Уменьшенный набор	15	15	0,68	0,77
ФЖ	Полный набор	72	72	0,87	0,71
	Уменьшенный набор	10	10	0,9	0,88

Библиографические ссылки

1. Ovarian Cancer Detection by Logical Analysis of Proteomic Data / G. Alexe [et al.] // Proteomics. 2004. No. 4(3). P. 766–783.

2. From Diagnosis to Therapy via LAD / D. Axelrod [et al.] // Invited Lecture at INFORMS Annual Meeting. Denver, CO. October, 2004.

3. Dupuis C., Gamache M., Páge J. F. Logical analysis of data for estimating passenger show rates in the airline industry // Journal of Air Transport Management. 2012. 18. P. 78–81.

4. Esmaeili S. Development of equipment failure prognostic model based on logical analysis of data // Master of Applied Science Thesis / Dalhousie University. Halifax, Nova Scotia, 2012.

5. Масич И. С. Комбинаторная оптимизация в задаче классификации // Системы управления и информационные технологии. 2009. № 1.2(35). С. 283–288.

6. Hammer P. L., Bonates T. O. Logical analysis of data – An overview: From combinatorial optimization to medical applications // Annals of Operations Research. 2006. 148.

7. Antamoshkin A. N., Masich I. S. Pseudo-Boolean optimization in case of unconnected feasible sets // Models and Algorithms for Global Optimization. Series: Springer Optimization and Its Applications. Springer. 2007. Vol. 4. P. 111–122.

8. Антамошкин А. Н., Масич И. С. Эффективные алгоритмы условной оптимизации монотонных псевдобулевых функций // Вестник СибГАУ. 2003. Вып. 4. С. 60–67.

9. Модель логического анализа для решения задачи прогнозирования осложнений инфаркта миокарда / С. Е. Головенкин [и др.] // Вестник СибГАУ. 2010. Вып. 4(30). С. 68–73.

10. Bache K., Lichman M. UCI Machine Learning Repository. Irvine, CA : University of California, School of Information and Computer Science, 2013. URL: http://archive.ics.uci.edu/ml.

11. Масич И. С. Приближенные алгоритмы поиска граничных точек для задачи условной псевдобулевой оптимизации // Вестник СибГАУ. 2006. 1(8). С. 39–43.

12. Alexe G., Hammer P. L. Spanned patterns for the logical analysis of data // Discrete Appl. Math. 154. 2006. P. 1039–1049.

13. Guoa C., Ryoo H. S. Compact MILP models for optimal and Pareto-optimal LAD patterns // Discrete Applied Mathematics. 160. 2012. P. 2339–2348.

14. Осложнения инфаркта миокарда: база данных для апробации систем распознавания и прогноза / С. Е. Головенкин [и др.] // Препринт № 6. 1997. Красноярск : Вычислительный центр СО РАН.

15. Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java. URL: http://www.cs. waikato.ac.nz/ml/weka.

References

1. Alexe G., Alexe S., Hammer P. L., Liotta L., Petricoin E., Reiss M. Ovarian CancerDetection by Logical Analysis of Proteomic Data. *Proteomics,* 2004, no. 4(3), P. 766–783.

2. Axelrod D., Bonates T., Hammer P. L., Lozina I. From Diagnosis to Therapy via LAD.Invited Lecture at INFORMS Annual Meeting, Denver, CO, October, 2004.

3. Dupuis C., Gamache M., Páge J. F. Logical analysis of data for estimating passenger show rates in the airline industry, Journal of Air Transport Management 18, 2012. P. 78–81.

4. Esmaeili S. 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. [Combinatorial optimization problem in the classification] *Sistemy upravleniya i informatsionnye tekhnologii*, 2009, no. 1.2(35), p. 283–288 (In Russ.).

6. Hammer P. L., Bonates T. O. Logical analysis of data – An overview: From combinatorial optimization to medical applications. Annals of Operations Research 148, 2006.

7. Antamoshkin A. N., Masich I. S. Pseudo-Boolean optimization in case of unconnected feasible sets. *Models and Algorithms for Global Optimization*. Series: Springer Optimization and Its Applications. Springer. 2007, vol. 4, p. 111–122.

8. Antamoshkin A. N., Masich I. S. [Efficient algorithms for constrained optimization pseudo-monotone functions]. *Vestnik SibGAU*. 2003, no. 4, p. 60–67 (In Russ.).

9. Golovenkin S. E., Masich I. S., Shulman V. A. Et al. [Model of logical analysis for solving problem of prognosis of myocardial infarction complication]. *Vestnik SibGAU*. 2010, no. 4(30), p. 68–73 (In Russ.).

10. Bache K., Lichman M. UCI Machine Learning Repository. Irvine, CA : University of California, School of Information and Computer Science, 2013. Available at: http://archive.ics.uci.edu/ml.

11. Masich I. S. [The heuristic algorithms of boundary points search for an constraint pseudo-Boolean optimization problem.] *Vestnik SibGAU*. 2006, no. 1(8), p. 39–43 (In Russ.).

12. Alexe G., Hammer P.L. Spanned patterns for the logical analysis of data, Discrete Appl. Math. 154, 2006. P. 1039–1049.

13. Guoa C., Ryoo H. S. Compact MILP models for optimal and Pareto-optimal LAD patterns, Discrete Applied Mathematics 160, 2012. P. 2339–2348.

14. Golovenkin S. E., Gorban A. N., Shulman V. A. et. al. *Oslozhneniya infarkta miokarda: baza dannykh dlya aprobatsii sistem raspoznavaniya i prognoza*. [Complications of myocardial infarction: a database for testing recognition systems and forecasting]. *Vychislitel'nyy tsentr SO RAN*: *Preprint.* 1997, no. 6, 14 p. (In Russ.).

15. Waikato Environment for Knowledge Analysis. Weka 3: Data Mining Software in Java. Available at: http://www.cs.waikato.ac.nz/ml/weka.