

ОБОБЩЕНИЕ КОМБИНИРОВАННОГО МЕТОДА «ФАКТОРНЫЙ АНАЛИЗ + ТАКСОНОМИЯ»

В.Д. Мазуров

Уральский федеральный университет

им. первого Президента России Б.Н. Ельцина, г. Екатеринбург

Рассматривается обобщенный метод анализа данных наблюдений, необходимый для обнаружения закономерностей, оценки важности признаков объектов и нахождения скрытых факторов. Предлагаемый метод основан на использовании дискриминантного анализа, таксономии и оценки информативности подсистем признаков. Анализ данных предполагает не только поиск глубинных факторов, но и обнаружение скрытых закономерностей. Предложена модель задачи обнаружения закономерностей, заключающаяся в ее сведении к дискриминантному анализу – задаче разделения множеств. При этом оказалось достаточным рассмотрение случая разбиения на два класса. При отсутствии аналитического описания этого разбиения работать с объектами можно только по их взаимодействию с приборами и экспертизами. Тогда восстановление классов производится по соответствующим прецедентным множествам. Конкретные приложения полученных результатов – задачи поиска закономерностей в неформализованных задачах математической экономики, математической биологии и медицины. В статье показано, что теория алгоритмов позволяет анализировать не только неформализованные задачи, но и в принципе неформализуемые.

Ключевые слова: дискриминантный анализ, таксономия, структура гиперграфа, теория алгоритмов.

Анализ данных наблюдений позволяет описывать группировки в множестве объектов и множестве их признаков. Это нужно для обнаружения закономерностей, оценки важности признаков объектов и нахождения скрытых факторов. Этот анализ можно проводить на основе дискриминантного анализа, таксономии и оценки информативности подсистем признаков [1–8]. Ранее мы предложили таксономию столбцов матрицы объект/признак для факторного анализа данных наблюдений. Факторами являются латентные признаки, линейными комбинациями которых определяются наблюдаемые признаки. Здесь предлагается некоторое обобщение этого метода. Оказывается, структура факторов (в том числе структура расстояний между ними) определяется структурой некоторых таксонов. Такую методику мы с аспирантом Д.В. Гилевым применили в задаче поиска закономерностей в медицине [9].

Пусть A – прецедентная выборка в множестве M в линейном пространстве L , $A = B \cup C$, пересечение множеств B и C пусто. Мы имеем в виду, что M – генеральная совокупность объектов из L , A – прецедентное множество описаний объектов, B и C – прецедентные подмножества двух классов.

Имеем таблицу наблюдений объектов двух классов с указанием значений предикатов:

Классы	Объекты	Предикаты	$p(x)$	$p(y)$
B	$\{x\}$	p	1	-1
C	$\{y\}$		-1	1

Преобразуем таблицу к следующему виду:

	Объекты	Предикат	$p(z)$
$z \in B \cup (-C)$	$\{z\}$		1

Предполагаем, что $B \cup (-C)$ можно считать матрицей (в общем случае бесконечной), строки которой – векторы состояния объектов в выбранном пространстве признаков. Положим

Краткие сообщения

$[B U (-C)]^* = D$, где $*$ – оператор транспонирования, D – множество столбцов массива векторов $B U (-C)$. Эти рассуждения понятны в конечномерном случае.

Применим к множеству D операцию таксономии:

$$\text{TAXON}(D) = \cup D_i \quad (i \in I).$$

Обычно требуют, чтобы при различных i и j таксоны D_i, D_j не пересекались. Но в нашем подходе (когда мы ищем факторы) они могут пересекаться.

Определим фактор

$$f_i = z(D_i),$$

где z – операция вычисления некоторого среднего элемента множества. В частности, $z(N)$ – это может быть средний арифметический вектор множества N :

$$z(N) = \sum (x \in N) / |N|,$$

где $|N|$ – число элементов множества N . Эта формула имеет простой смысл в конечномерном случае.

Вообще, вид элемента $z(N)$ определяется содержательным смыслом решаемой задачи.

Интерпретация фактора f_i есть функция смыслов факторов, входящих в таксон D_i .

При этом мы рассматриваем тот случай, когда D_i есть i -я максимальная по включению совместная подсистема системы

$$\phi(z) > 0, z \in D, \phi \in \Phi.$$

В этом случае таксоны могут пересекаться.

Это система относительно искомого преобразования ϕ , применяемого в задачах таксономии. Если она совместна, то имеется только один таксон и соответственно один фактор. Однако чаще всего в прикладных задачах она несовместна и определяет несколько факторов.

Структуру гиперграфа максимальных по включению совместных подсистем (связанная с структурой факторов) глубоко исследовал М.Ю. Хачай [10]. Эти разработки важны в методе комитетов для несовместных задач прогнозирования, классификации и распознавания.

Определим расстояние между факторами через расстояния между соответствующими таксонами. Это расстояние может быть определено как среднее значение расстояний между элементами таксонов, либо как расстояние между центрами тяжести таксонов.

Анализ данных предполагает не только поиск глубинных факторов, но и обнаружение скрытых закономерностей.

Здесь я хочу предложить одну из моделей задачи обнаружения закономерностей.

Во-первых, замечу, что эту задачу можно свести к дискриминантному анализу – задаче разделения множеств. При этом достаточно рассмотреть случай разбиения на два класса. Конкретные приложения – решённые нами задачи поиска закономерностей в неформализованных задачах математической экономики, математической биологии и медицины. Интересно, что теория алгоритмов позволяет анализировать не только неформализованные задачи, но и в принципе неформализуемые. При этом опираются на теоремы Гёделя.

Теперь предположим, что мы располагаем следующей информацией. Есть конечное множество векторов состояний объектов. Предполагается, что оно разбито на классы, но аналитическое описание этого разбиения вначале отсутствует. Более того, объекты представляют собой «вещи в себе», но работать с ними можно по их взаимодействию с приборами и экспертизами. Результат такого взаимодействия – представление объекта как вектора x в линейном пространстве L .

Моделирование разбиения пространства L на два класса X_1 и X_2 зависит от выбора класса L^* разделяющих функций. Классы разделяются выражением $y = f(x)$, f – искомая функция из L^* . При этом

X_1 содержится в множестве $\{x: f(x) > 0\}$;

X_2 – в множестве $\{x: f(x) < 0\}$;

Граница Γ между классами – в множестве $\{x: f(x) = 0\}$.

Восстановление классов производится по прецедентным множествам $A, B,$ и C из классов X_1, X_2 и Γ соответственно.

Приложения: социология, биология и медицина

Работа поддержана РФФ № 14 – 11 – 00109.

Литература

1. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Институт математики, 1999. – 267 с.
2. Браверманн, Э.М. Структурные методы обработки эмпирических данных / Э.М. Браверманн, И.Б. Мучник. – М.: Наука, 1983. – 404 с.
3. Мазуров, В.Д. Нейронные сети и экспертизы / В.Д. Мазуров, Н.И. Потанин // Тезисы конференции ИММ Уро РАН. – 1983.
4. Вапник, В.Н. Теория распознавания образов / В.Н. Вапник. – М.: Наука, 1974. – 416 с.
5. Мазуров, В.Д. Упорядочить хаос / В.Д. Мазуров, И.И.Еремин // Известия УрГУ. Серия «Общественные науки». – 2001. – 21. – С. 6–9.
6. Донской, В.И. Синтез согласованных линейных оптимизационных моделей по прецедентной информации / В.И. Донской // Ученые записки ТНУ. Сер. «Физ.-мат. науки». – 2010. – Т. 23. – № 2. – С. 56–65.
7. Калядин, Н.И. Конструктивизация моделей классификации конечных объектов / Н.И. Калядин // Изв. ИМИ УдГУ. – 2007. – № 1(38). – 231 с.
8. Мазуров, В.Д. Метод комитетов в задачах оптимизации и классификации / В.Д. Мазуров. – М.: Наука, 1990. – 248 с.
9. Мазуров, В.Д. Модель динамики объектов в противоречивых условиях / В.Д. Мазуров, Д.В. Гилев // *Sworld. Problems and ways of their solution in science, transport and so on.* – 2012. – December – P. 34–41.
10. Хачай, М.Ю. Комитетные решения несовместных систем ограничений и методы обучения распознаванию: дис. ... д-ра физ.-мат. наук / М.Ю. Хачай. – Екатеринбург, 2004. – 175 с.

Мазуров Владимир Данилович, д-р физ.-мат. наук, профессор, профессор кафедры эконометрики и статистики высшей школы экономики и менеджмента, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина», г. Екатеринбург; vldmazurov@gmail.com

Поступила в редакцию 5 апреля 2015 г.

DOI: 10.14529/ctcr150217

**SOME GENERALIZATION OF COMBINED METHOD
“FACTOR ANALYZE + TAXONOMY”**

V.D. Mazurov, Ural Federal University named after the First President of Russia Boris Yeltsin, Ekaterinburg, Russian Federation, vldmazurov@gmail.com

The article discusses the generalized method of analysis of observational data needed to detect regularities assess the importance of attributes of the objects and finding the hidden factors. The proposed method is based on the use of discriminated analysis, taxonomy and evaluation of the information subsystems signs. Analysis of the data suggests not only the search for the underlying factors, but also the discovery of hidden patterns. A model of the problem of detection of regularities is suggested, that is its reduction to the discriminanted analysis – the problem of separating sets. In this case, it was sufficient to consider the case of the partition into two classes. In the absence of an analytical description of this partition, work with objects, you can only interact with their devices and examinations. Then the classes recovery makes by the relevant case sets. Specific applications of the results are the search problem patterns in non-formalized problems of mathematical economics, mathematical biology and medicine. The article shows that the theory of algorithms allows us to analyze not only formalized tasks, but also, in principle, non-formalizable.

Keywords: discriminant analysis, taksonomy, structure of the hypergraph, the theory of algorithms.

References

1. Zagoruyko N.G. *Prikladnye metody analiza dannykh i znaniy* [Applied Methods of the Analysis of Data and Knowledge]. Novosibirsk, Institute of Mathematics, 1999, 267 p.
2. Bravermann E.M., Muchnik I.B. *Strukturnye metody obrabotki empiricheskikh dannykh* [Structural Methods of Processing Empirical Data]. Moscow, Science, 1983, 404 p.
3. Mazurov V.D., Potanin N.I. Neural Networks and Examinations. *Tezisy konferentsii IMM Uro RAN* [Theses of the IMM Conference Uro RAN], 1983.
4. Vapnik V.N. *Teoriya raspoznavaniya obrazov* [Theory of Recognition of Images]. Moscow, Science, 1974 416 p.
5. Mazurov V.D., Eremin I.I. [Organize Chaos]. *News of USU. Ser. Social Sciences*, 2001, No. 21, pp. 6–9. (in Russ.)
6. Donskoy V.I. [Synthesis of the Coordinated Linear Optimizing Models According to Case Information]. *Scientific Notes of TNU. Ser. Physical and Math Sciences*, 2010, vol. 23, no. 2, pp. 56–65. (in Russ.)
7. Kalyadin N.I. *Konstruktivizatsiya modeley klassifikatsii konechnykh ob'ektov* [The Constructivization of the Classification Model Targets]. Izhevsk, Bulletin of the Institute of Mathematics and Informative of Udmurt State University, 2014, no. 1(38), 231 p.
8. Mazurov V.I. D. *Metod komitetov v zadachakh optimizatsii i klassifikatsii* [Metod of Committees in Problems of Optimization and Classifications]. Moscow, Science, 1990, 248 p.
9. Mazurov V.I., Gilev D.V. Model of Dynamics of Objects in Contradictory Conditions. *Sworld. Problems and Ways of their Solution in Science, Transport and so on*. December, 2012, pp. 34–41.
10. Khachay M. Yu. *Komitetnye resheniya nesovmestnykh system ogranicheniy i metody obucheniya raspoznavaniyu. Avtoref. doct. dis.* [Komitetnye Decisions of Not Joint Systems of Restrictions and Methods of Training in Recognition. Abstract of doct. diss.]. Ekaterinburg, 2004. 175 p.

Received 5 April 2015

БИБЛИОГРАФИЧЕСКОЕ ОПИСАНИЕ СТАТЬИ

Мазуров, В.Д. Обобщение комбинированного метода «Факторный анализ + таксономия» / В.Д. Мазуров // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2015. – Т. 15, № 2. – С. 139–142. DOI: 10.14529/ctcr150217

REFERENCE TO ARTICLE

Mazurov V.D. Some Generalization of Combined Method “Factor Analyze + Taxonomy”. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2015, vol. 15, no. 2, pp. 139–142. (in Russ.) DOI: 10.14529/ctcr150217