Вестник БГУ. Сер. 1. 2016. № 3. С. 120–124

УДК 004.82:004.89

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ СКРЫТЫХ ЗАКОНОМЕРНОСТЕЙ НА ОСНОВЕ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ ОБУЧАЮЩЕЙ ВЫБОРКИ

В. В. КРАСНОПРОШИН [1)]***, В. Г. РОДЧЕНКО*** [1)]

[1)] *Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Республика Беларусь*

Рассматриваются метод и средства решения задачи обнаружения скрытых закономерностей в интеллектуальном анализе данных. Обоснована актуальность поиска новых решений, обеспечивающих автоматическое обнаружение ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации закономерностей. Предложен оригинальный метод автоматического обнаружения скрытых закономерностей, который базируется на применении гипотезы компактности и статистического анализа данных обучающей выборки. Описан алгоритм построения пространства решений, в котором образы классов представляют взаимно разделенные компактные сгустки. Приведена обобщенная формулировка для выявляемых закономерностей и показана возможность их интерпретации в рамках предметной области. Продемонстрирован механизм применения разработанного метода для решения в автоматическом режиме задач распознавания образов с обучением.

Ключевые слова: интеллектуальный анализ данных; обучающая выборка; машинное обучение.

AUTOMATIC DETECTION OF HIDDEN PATTERNS BASED ON STATISTICAL ANALYSIS OF THE TRAINING SAMPLE DATA

V. V. KRASNOPROSHIN [a]***, V. R. RODCHANKA*** [a]

[a] *Belarusian State University, Nezavisimosti avenue, 4, 220030, Minsk, Republic of Belarus*

The method and the existing tools for solving the problem of finding hidden patterns in data mining are considered. Substantiates the relevance of the search for new solutions to ensure automatic detection of unknown, non-trivial, practically useful and affordable interpreting hidden patterns. The original method of automatically detecting hidden patterns, which is based on the use of the hypothesis of compactness and statistical analysis of the training sample data, has been proposed. The algorithm for constructing the solution space in which patterns of classes are mutually separated by compact clusters has been described. Generalized formulation to identify patterns presented. The ability to interpret them as part of the domain is shown. The mechanism of the application of this method to solve the pattern recognition problems with learning in the automatic mode was demonstrated.

Key words: data mining; training sample; machine learning.

Образец цитирования:

Краснопрошин В. В., Родченко В. Г. Автоматическое обнаружение скрытых закономерностей на основе статистического анализа данных обучающей выборки // Вестн. БГУ. Сер. 1, Физика. Математика. Информатика. 2016. № 3. С. 120–124.

For citation:

Krasnoproshin V. V., Rodchanka V. R. Automatic detection of hidden patterns based on statistical analysis of the training sample data. *Vestnik BGU. Ser. 1, Fiz. Mat. Inform*. 2016. No. 3. P. 120–124 (in Russ.).

Авторы:

Виктор Владимирович Краснопрошин – доктор технических наук, профессор; заведующий кафедрой информационных систем управления факультета прикладной математики и информатики.

Вадим Григорьевич Родченко – кандидат технических наук, доцент; докторант кафедры информационных систем управления факультета прикладной математики и информатики.

Authors:

Victor Krasnoproshin, doctor of science (technics), full professor; head of the department of management information systems, faculty of applied mathematics and computer science.
krasnoproshin@bsu.by

Vadzim Rodchanka, PhD (technics), docent; doctoral student at the department of information management systems, faculty of applied mathematics and computer science.
rovar@grsu.by

Интенсивное развитие и использование информационных технологий практически во всех областях человеческой деятельности привело к накоплению огромных объемов данных, которые первоначально размещались в базах данных (*Database*), а в настоящее время широкое распространение получили хранилища данных (*Data Warehouse*) и витрины данных (*Data Mart*). Опыт применения технологий обработки информации показывает, что возможности языка структурированных запросов SQL позволяют реализовать «поверхностный» анализ данных с точки зрения выявления существующих внутри их закономерностей. Оперативная аналитическая обработка на основе технологии OLAP (*online analytical processing* – аналитическая обработка в реальном времени) ориентирована на извлечение из данных знаний, которые условно можно отнести к «неглубокому» уровню залегания. Наибольший же практический интерес представляют «глубинные», или скрытые, закономерности, на обнаружение которых ориентирован аппарат интеллектуального анализа данных (*Data Mining*) [1].

Под *интеллектуальным анализом данных* (ИАД) принято понимать совокупность методов для поиска и выявления в массивах данных ранее неизвестных, нетривиальных, полезных в практическом плане и доступных интерпретации знаний, которые необходимы для принятия решений в различных сферах человеческой деятельности [2].

Для повышения эффективности обнаружения и дальнейшего практического использования скрытых закономерностей актуальной является разработка методов, которые обеспечивают автоматическое извлечение ранее неизвестных и поддающихся интерпретации знаний.

В настоящей статье предлагается метод обнаружения скрытых и интерпретируемых в терминах предметной области закономерностей на основе статистического анализа данных обучающей выборки, который позволяет в автоматическом режиме решать задачи распознавания образов с учителем.

Автоматическое обнаружение закономерностей в ИАД

В информатике под знанием принято понимать такую новую информацию, которую получает решатель (человек, компьютер) вследствие применения алгоритма к результирующей информации для решения задачи и которая может быть полезной для решения других задач. Перевод информации в категорию знаний осуществляется только человеком или специализированной программой на основе анализа результатов решений с помощью соответствующих критериев оценки [3].

В случае автоматического обнаружения скрытых закономерностей к процессу извлечения новых знаний предъявляются более жесткие требования, поскольку в этом случае компьютерная программа выступает и в качестве решателя и осуществляет перевод информации в категорию знаний без участия человека.

Процесс извлечения знаний непосредственно связан с обучением. В области искусственного интеллекта различают два типа обучения: дедуктивное и по прецедентам.

Использование дедуктивного обучения связано с развитием направления, которое принято называть инженерией знаний. Оно объединяет задачи получения знаний из относительно простой информации, их систематизацию и практическое использование в виде экспертных систем. Данный тип обучения предполагает формализацию знаний экспертов и их представление в компьютере в виде базы знаний [4].

Обучение по прецедентам, или *индуктивное*, основано на выявлении закономерностей в эмпирических данных. Проблематика обучения по прецедентам касается процесса *самостоятельного* получения знаний интеллектуальной системой в процессе ее работы [5].

Поскольку «производство» знаний из огромных накопленных массивов данных является одной из базовых проблем интеллектуального анализа данных, то для эффективного обеспечения соответствующего «производственного процесса» методы ИАД должны в первую очередь обеспечивать его автоматизацию. В настоящее время автоматическое извлечение скрытых закономерностей удается реализовать только в рамках использования технологий искусственных нейронных сетей и машинного обучения (обучение по прецедентам) [6, 7].

Обучение – одно из характерных свойств искусственных нейронных сетей, а его результатом является нахождение значений коэффициентов связей между нейронами сети. При этом автоматически выявляются скрытые закономерности сложной зависимости между входными и выходными данными. Обученная нейронная сеть фактически обнаруживает скрытые закономерности, но формально представляет их в виде «черного ящика». В итоге успешно обученная искусственная нейронная сеть может выполнять обобщение, т. е. генерировать верный результат на основании данных, отсутствующих в обучающей выборке, но не дает возможности проинтерпретировать выявленные скрытые закономерности.

В случае обучения по прецедентам изначально известно конечное множество прецедентов, которое называется *обучающей выборкой*. На основе данных этой выборки требуется восстановить зависимость и построить алгоритм, правильно классифицирующий объекты. Реализацию такого алгоритма можно

осуществить в рамках использования гипотезы компактности, суть которой заключается в том, что представление одного и того же образа обычно отражается в признаковом пространстве в геометрически близких точках, образуя «компактные» сгустки [8]. Из практики известно, что гипотеза компактности не всегда подтверждается. Например, если среди признаков имеются случайные – *малоинформативные или неинформативные*, то такой случай относится к разряду плохого представления образов и экземпляры одного и того же образа не формируют явно выделяемых компактных сгустков [9].

В рамках развития методов *обучения по прецедентам* предлагается на основе статистического анализа данных обучающей выборки реализовать в автоматическом режиме процедуру построения такого пространства решений, в котором выполняется гипотеза компактности.

Метод построения пространства решений

Пусть имеется множество объектов, каждый из которых принадлежит к одному из заранее заданных классов и формально описывается вектор-столбцом вида $z^T = (z_1, z_2, \ldots, z_n)$, где $z_i \in R$ – значение i-го признака. Объединение всех объектов множества задает так называемую *обучающую выборку*, которая формально является матрицей $Z_{n \times m}$, где $m = m_1 + m_2 + \ldots + m_k$; m_i – количество объектов i-го класса; k – количество классов.

Формальный образ i-го класса первоначально предлагается представить в виде матрицы, построенной на основе всех объектов класса. Проведем нормировку обучающей выборки $Z_{n \times m}$, получим $X_{n \times m}$, где $x_{ij} = \dfrac{(z_{ij} - z_{\min})}{(z_{\max} - z_{\min})}$, и объединим все векторы i-го класса в отдельную матрицу вида

$$X^i_{n \times m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \ldots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \ldots & x^i_{2m_i} \\ \ldots & \ldots & \ldots & \ldots \\ x^i_{n1} & x^i_{n2} & \ldots & x^i_{nm_i} \end{pmatrix},$$

где $i = \overline{1, k}$; $j = \overline{1, m_i}$.

При этом объекты класса в пространстве R^n представляются вектором с координатами вершины $(x_1, x_2, \ldots, x_n)$, где $x_i \in R$ – значение i-го признака.

Воспользуемся гипотезой компактности, которая гласит, что образам классов соответствуют компактные множества в пространстве признаков, т. е. классы образуют локализованные подмножества в пространстве признаков, а граница между классами имеет достаточно четко выраженную форму [10].

Для простоты рассмотрим гипотезу компактности в предположении, что исследуются объекты двух классов, формальные образы которых заданы матрицами $X^1_{n \times m_1}$ и $X^2_{n \times m_2}$. Очевидно, что каждая i-я строка содержит выборку нормированных значений i-го признака. Используя непараметрический статистический критерий, проводим попарно сравнение значений всех признаков на основе данных соответствующих строк матриц. Здесь фактически определяется, достаточно ли мала зона перекрещивающихся значений между двумя рядами (ранжированным рядом значений в первой выборке и таким же во второй выборке).

Для построения пространства решений отбираем только те признаки, для которых непараметрический статистический критерий выявил наличие существенного различия в рассматриваемых выборках.

Поскольку в общем случае имеется k классов, то для построения пространства решений воспользуемся следующим критерием отбора признаков, обеспечивающих разделение классов. *Признак, у которого для всех C^2_k возможных комбинаций пар выборок использование непараметрического статистического критерия выявило наличие существенного различия в рассматриваемых выборках, обеспечивает разделение образов классов.*

Построение пространства решений на основе анализа данных обучающей выборки выполняется автоматически по следующему алгоритму:

1) на вход подается обучающая выборка, в которой образ каждого i-го класса формально представляется в виде матрицы $X^i_{n \times m_i}$;

2) в цикле последовательно на основе использования критерия отбора признаков, обеспечивающих разделение классов, исследуется каждый признак, и в итоговый словарь включаются только те из них, которые удовлетворяют условиям критерия;

3) если итоговый словарь оказался непустым, то обучение прошло успешно и на основе выявленных признаков строится пространство решений, в котором образы классов образуют явно выделяемые компактные сгустки. Если же словарь оказался пустым, то принимается одно из решений: либо о переформировании обучающей выборки на основе нового набора признаков, либо о применении других алгоритмов обучения.

Необходимо отметить, что, во-первых, результатом построения непустого итогового словаря является следующая скрытая закономерность: *в признаковом пространстве, построенном на основе итогового словаря, образы классов образуют явно выделяемые компактные сгустки и можно осуществлять классификацию объектов*. Во-вторых, обнаруженная закономерность может быть проинтерпретирована в рамках соответствующей предметной области на основе признаков итогового словаря, поскольку они обладают свойством разделения друг с другом образов разных классов.

Распознавание образов с обучением на основе предложенного метода

Постановка задачи распознавания следующая: пусть Z – множество описаний объектов, Y – множество номеров (или наименований) классов. Существует неизвестная целевая зависимость, т. е. отображение $y^*: Z \to Y$, значения которой известны только на объектах конечной обучающей выборки $Z^m = \{(z_1, y_1), \ldots, (z_m, y_m)\}$. Требуется построить алгоритм, способный для произвольного объекта $z \in Z$ определить его принадлежность к одному из классов [11].

Задачу распознавания будем решать в рамках использования гипотезы компактности и предположения о том, что на основе анализа данных обучающей выборки можно построить признаковое подпространство, в котором выполняется условие гипотезы компактности.

Первоначально *обучающая выборка* формально представляет собой матрицу $Z_{n\times m}$, где $m = m_1 + \ldots + m_2 + \ldots + m_k$; m_i – количество объектов i-го класса; k – количество классов; n – количество признаков. Для приведения значений данных обучающей выборки $Z_{n\times m}$ к единому масштабу осуществляем их нормировку и получаем матрицу $X_{n\times m}$, где $x_{ij} = \dfrac{(z_{ij} - z_{\min})}{(z_{\max} - z_{\min})}$, которая представляет собой объединение матриц $X^i_{n\times m_i}$, содержащих все векторы i-го класса. При этом объекты класса в n-мерном признаковом пространстве R^n представляются вектором с координатами вершины $(x_1, x_2, \ldots, x_n)$, где $x_i \in R$ – значение i-го признака.

В цикле последовательно анализируем все признаки, значения которых представлены соответствующими строками матриц $X^i_{n\times m_i}$. Придерживаемся следующего правила: *очередной j-й признак* $\left(j = \overline{1, n}\right)$ *включается в итоговый словарь, если для всех* C_k^2 *возможных комбинаций пар выборок использование непараметрического статистического критерия Лемана – Розенблатта выявило наличие существенного различия в рассматриваемых выборках* [12].

Если в результате получается непустой итоговый словарь, то процедура обучения завершается удачно и на основе признаков этого словаря создается пространство, в котором выполняется гипотеза компактности. Для формирования образов классов в этом пространстве из матриц $X^i_{n\times m_i}$ исключаем все строки, которые содержат значения признаков, не попавших в итоговый словарь, и получаем матрицы $T^i_{p\times m_i}$, где m_i – количество объектов i-го класса; p – количество признаков итогового словаря, причем $p \le n$.

Отметим, что в результате построения непустого итогового словаря была выявлена скрытая закономерность, которую можно интерпретировать следующим образом: *в признаковом пространстве, построенном на основе итогового словаря, образы классов образуют явно выделяемые компактные сгустки и можно осуществлять классификацию объектов*. Кроме того, эта скрытая закономерность может быть также интерпретирована в терминах предметной области, поскольку каждый из признаков итогового словаря несет смысловую нагрузку в рамках этой предметной области.

Классификацию исследуемого объекта можно осуществить на основе использования формальных образов классов, заданных в виде матриц $T^i_{p\times m_i}$, и применения одного из известных метрических алгоритмов классификации (например, алгоритм k ближайших соседей) [13].

Описанный выше процесс решения задачи распознавания позволяет в автоматическом режиме выполнять процедуру обнаружения скрытых и интерпретируемых закономерностей на основе анализа данных содержимого обучающей выборки. В случае когда итоговый словарь оказывается пустым, необходимо либо переформировать обучающую выборку на основе нового словаря признаков, либо применить другие алгоритмы обучения.

Таким образом, предложен один из возможных подходов к решению задачи автоматического обнаружения и интерпретации скрытых закономерностей в интеллектуальном анализе данных.

Анализ существующих методов ИАД продемонстрировал актуальность поиска новых решений, обеспечивающих автоматическое обнаружение ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации закономерностей.

Для автоматического обнаружения скрытых закономерностей и для решения задачи распознавания образов с обучением предложен оригинальный метод, предусматривающий построение на основе анализа содержимого обучающей выборки пространства решений, в котором образы классов представляют взаимно разделенные компактные сгустки.

Описан метод для автоматического обнаружения скрытых закономерностей, который базируется на использовании гипотезы компактности и анализе данных обучающей выборки. Представлена обобщенная формулировка для выявляемых закономерностей и продемонстрирована возможность их интерпретации в рамках предметной области. Предложен вариант автоматического решения задачи распознавания образов с обучением на основе разработанного метода обнаружения скрытых закономерностей.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК (*REFERENCES*)

1. *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining. СПб., 2004.

2. Интеллектуальный анализ данных // Википедия – свободная энциклопедия [Электронный ресурс]. 2001. Режим доступа: http://ru.wikipedia.org/wiki/Data_mining (дата обращения: 15.02.2016).

3. *Краснопрошин В. В., Маркова О. А., Вальвачев А. Н.* Система понятий в информатике // Информатика. 2007. № 3. С. 124 [Krasnoproshin V. V., Markova O. A., Valvachev A. N. System concepts in computer science. *Informatika.* 2007. No. 3. P. 124 (in Russ.)].

4. *Джарратано Дж.* Экспертные системы: принципы разработки и программирование. М., 2007.

5. *Краснопрошин В. В., Образцов В. А.* Проблема принятия решений по прецедентности: разрешимость и выбор алгоритмов // Выбр. навук. працы Беларус. дзярж. ун-та. 2001. Т. 6 : Матэматыка. С. 285 [Krasnoproshin V. V., Obraztsov V. A. The problem of decision-making by precedents: solvability and selection algorithms. *Vybranyja navuk. pracy Belarus. dzjarzhawnaga wniversitjeta.* 2001. Vol. 6 : Matematyka. P. 285 (in Russ.)].

6. *Хайкин С.* Нейронные сети : полн. курс. 2-е изд. М., 2006.

7. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск, 1999.

8. *Аркадьев А. Г., Браверман Э. М.* Обучение машины распознаванию образов. М., 1964.

9. Базовые эмпирические гипотезы // Институт математики им. С. Л. Соболева : сайт [Электронный ресурс]. 2009. Режим доступа: http://math.nsc.ru/AP/oteks/Russian/links/gip/index.html (дата обращения: 17.02.2016).

10. Гипотеза компактности // Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. 2008. Режим доступа: http://www.machinelearning.ru/wiki/index.php?title = Гипотеза_ компактности (дата обращения: 17.02.2016).

11. Задача классификации // Википедия – свободная энциклопедия [Электронный ресурс]. 2001. Режим доступа: http://ru.wikipedia.org/wiki/Задача_классификации (дата обращения: 17.02.2016).

12. *Орлов А. И.* О проверке однородности двух независимых выборок // Завод. лаборатория. 2003. Т. 69, № 1. С. 55 [Orlov A. I. About the Verification the Homogeneity of Two Independent Samples. *Zavod. lab.* 2003. Vol. 69, No. 1. P. 55 (in Russ.)].

13. Математические методы обучения по прецедентам (теория обучения машин) // Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. 2008. Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 22.02.2016).

Статья поступила в редколлегию 02.05.2016.
Received by editorial board 02.05.2016.