УДК 004.82 : 004.89

В.В. Краснопрошин, В.Г. Родченко

КЛАСТЕРНЫЕ СТРУКТУРЫ И ИХ ПРИМЕНЕНИЕ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

Вводится понятие кластерных структур, которые можно использовать для выявления и представления «скрытых знаний» в интеллектуальном анализе данных. Демонстрируется возможность применения кластерных структур для решения в автоматическом режиме задач распознавания образов с обучением.

Введение

В настоящее время одной из актуальных и наиболее востребованных в практическом плане задач информатики является извлечение полезных знаний из ранее накопленных в электронном виде данных. Проводимые исследования свидетельствуют о том, что рост количества таких данных значительно превышает темпы роста как самих знаний, так и эффективности их использования.

Процесс поиска в базах данных «скрытых знаний» определяет последовательность действий, которые необходимо выполнить для того, чтобы из исходных «сырых» данных извлечь знания. Достоинством такого подхода является его универсальность, поскольку отсутствует зависимость от конкретной предметной области [1]. В целом процесс предусматривает выполнение четырех основных этапов: подготовки, предобработки, интеллектуального анализа данных и постобработки полученных результатов. При этом центральная роль отводится интеллектуальному анализу данных, который ориентирован как на разработку методов и алгоритмов извлечения знаний, так и на эффективное использование выявленных закономерностей для предсказания значений одной части информации по известным значениям ее другой части [2, 3].

В статье предлагается метод построения так называемых кластерных структур, использование которых позволяет в автоматическом режиме на основе анализа содержимого обучающей выборки выполнять процедуру поиска и формального представления «скрытых закономерностей» при решении задач распознавания образов с обучением.

1. Модели представления знаний в интеллектуальном анализе данных

Знания – новая информация, полученная решателем (человеком, компьютером) вследствие применения алгоритма к результирующей информации для решения задачи и, возможно, полезная для решения других задач [4]. Перевод информации в категорию знания может быть осуществлен только человеком или специализированной программой, анализирующей результаты решений с помощью соответствующих критериев оценки.

Центральная задача интеллектуального анализа данных заключается в выявлении «скрытых знаний» на основе анализа содержимого предварительно полученной базы данных. Под такими знаниями подразумеваются: ранее неизвестные; нетривиальные, которые нельзя обнаружить на основе визуального анализа или путем вычисления простых статистических характеристик; практически полезные для использования; знания, представляемые в наглядной пользователю форме и в терминах предметной области [5, 6].

Поскольку интеллектуальный анализ данных ориентирован на извлечение знаний, то естественным образом возникает проблема поиска эффективных моделей их представления.

На сегодняшний день известен ряд моделей представления, которые принято разделять на теоретические и эмпирические. К теоретическим моделям относятся представление знаний, основанных на исчислении высказываний, исчислении предикатов, на формальных грамматиках, а также комбинаторные и алгебраические модели. На основе указанного набора моделей пока удавалось решать небольшой класс относительно простых практических задач.

Эмпирические модели основаны на использовании принципов организации памяти человека и на моделировании механизмов решения задач [7]. К ним относятся продукционные модели, семантические сети и фреймовые модели. Сюда же можно отнести искусственные нейронные сети и генетические алгоритмы.

Продукционные модели, семантические сети и фреймовые модели акцентируют внимание на проблемах символьного представления и стратегиях формальных рассуждений и опираются на использование гипотезы о физической символьной системе [8] .

Альтернативными гипотезе о физической символьной системе являются вычислительные модели познания, в которых акцент фокусируется на проблемах обучения и адаптации [9]. Наиболее известными представителями этих моделей являются искусственные нейронные сети.

Сегодня основными требованиями к интеллектуальному анализу данных являются эффективность, простота, автоматизм. Если в этом ракурсе рассматривать эмпирические модели, то следует указать на то, что автоматизм с точки зрения извлечения «скрытых знаний» может быть в известной степени обеспечен только использованием искусственных нейронных сетей. Однако существенным недостатком таких сетей является то, что даже при правильном обучении сети нельзя проинтерпретировать в аналитических терминах сам процесс обучения.

Искусственные нейронные сети строятся и функционируют по принципу биологических нейронных сетей. На основе их использования удается строить эффективные системы распознавания объектов сложной природы, но не удается, к сожалению, находить ответы на фундаментальные вопросы, связанные с пониманием природы исследуемого процесса или явления [10].

Центральной задачей интеллектуального анализа, как уже отмечалось ранее, является извлечение знаний из данных, но поскольку из обученной нейронной сети практически не удается извлечь явный и понятный пользователю алгоритм решения задачи, то предпринимаются попытки специальным образом строить процедуры упрощения и вербализации с целью определения явного метода решения. Однако любое упрощение не проходит бесследно и в данном случае серьезно отражается на качестве модели.

В работе предпринята попытка решения проблемы с использованием машинного обучения. Machine Learning (обучение по прецедентам) основано на выявлении закономерностей внутри эмпирических данных, под которыми понимается множество фактов, полученных в результате наблюдения или эксперимента и структурированных в рамках решения задачи интеллектуального анализа данных [11]. Процедура обучения осуществляется на основе данных обучающей выборки, но в отличие от технологии искусственных нейронных сетей она ориентирована на выделение информативных признаков или их комбинаций с точки зрения разделения объектов, процессов или явлений в многомерном пространстве решений.

Формальное представление образов объектов (паттернов) и получаемых в результате обучения «скрытых закономерностей» предлагается реализовать на основе специальным образом построенных кластерных структур.

2. Метод построения кластерных структур

Пусть имеется множество объектов, каждый из которых является представителем определенного подмножества (класса) и формально описывается вектор-столбцом вида $z^{\mathrm{T}}=(z_1,z_2,\ldots,z_n)$, где $z_i \in R$ – значение i-го признака. Объединение объектов из всех классов задает так называемую *обучающую выборку*, которую формально можно записать в виде матрицы $Z_{n\,x\,m}$, где $m=m_1+m_2+...+m_k$, m_i – количество объектов i-го класса, k – количество классов.

Формальный образ каждого класса предлагается строить в виде специальных пространственных структур на основе объектов данного класса. Для этого выполним нормировку обучающей выборки $Z_{n\,x\,m}$ (где $m=m_1+m_2+...+m_k$ и m_i – количество объектов i-го класса), получим $X_{n\,x\,m}$, где $x_{ij}=(z_{ij}-z_{min})/(z_{max}-z_{min})$, и объединим все векторы i-го класса в отдельную матрицу вида

$$X^i_{n\,x\,m_i}=\begin{pmatrix} x^i_{11} & x^i_{12} & \ldots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \ldots & x^i_{2m_i} \\ \ldots & \ldots & \ldots & \ldots \\ x^i_{n1} & x^i_{n2} & \ldots & x^i_{nm_i} \end{pmatrix}, \text{ где } i=\overline{1,k},\ j=\overline{1,m_i}.$$

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.

Процесс формирования пространственной структуры начинается с начального каркаса в виде минимального остовного дерева графа, построенного на множестве вершин векторов *i*-го класса. Для этого можно, например, использовать алгоритмы Прима или Краскала [11].

Очевидно, что в этом случае исходный граф содержит m_i вершин и $C^2_{m_i}$ ребер. При этом весом каждого ребра является его длина. Будем считать, что построенная таким образом структура задает в *n*-мерном признаковом пространстве формализованный образ класса.

По аналогии с гравитационным полем в физике для каждой точки ребра построенной структуры определим «сферу влияния» – область пространства с центром в этой точке, ограниченную гиперсферой радиуса *r*, где значение радиуса является функцией от длины ребра. В результате на основе каждого ребра образуется объемный элемент и вся пространственная структура становится объемной. Предлагаемый подход позволит получать численные оценки взаимного размещения формализованных образов объектов в многомерном пространстве принятия решений.

На каждом ребре каркаса предлагается размесить по три *n*-мерных гипершара с центрами в вершинах и в средней точке ребра. В итоге строится базовый элемент пространственной структуры (рис. 1).

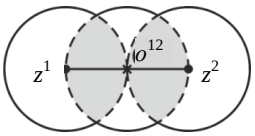


Рис. 1. Базовый элемент плоской кластерной структуры (z^1 и z^2 – вершины, o^{12} – середина ребра)

Объединяя базовые элементы с общей вершиной, получаем промежуточный вариант структуры (рис. 2).

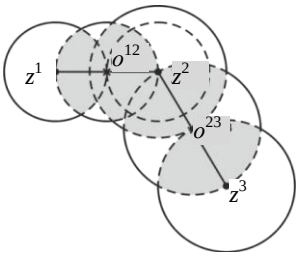


Рис. 2. Двухреберная плоская кластерная структура

Процесс построения пространственной структуры завершается, когда будут объединены все ее базовые элементы. Структуру, построенную в результате описанных выше действий, назовем кластерной.

3. Свойства кластерных структур

С точки зрения геометрической интерпретации кластерная структура представляет собой объединение взаимопересекающихся гипершаров в *n*-мерном пространстве (рис. 3).

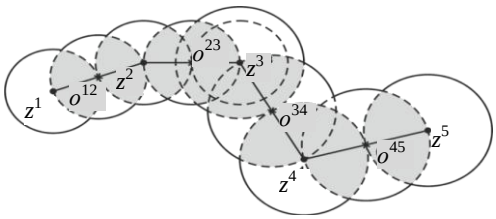


Рис. 3. Пример плоской кластерной структуры

Формально описание кластерной структуры, построенной на основе k векторов, можно представить в виде следующей таблицы.

Формальное описание кластерной структуры

Номер гипершара	Координаты центра гипершара	Радиус гипершара
1	z^1	r^1
2	o^1	r^1
3	z^2	$\max(r^1, r^2)$
….	….	….
$2k$–3	z^{k-1}	$\max(r^{k-1}, r^{k-2})$
$2k$–2	o^{k-1}	r^{k-1}
$2k$–1	z^k	r^{k-1}

Для каждой вершины $z^2,...,z^{k-1}$ в таблицу записываем максимальное значение радиуса из двух соответствующих пересекающихся гипершаров.

Очевидно, что кластерную структуру можно охарактеризовать числом вершин векторов, на основе которых построен ее каркас, а также объемом и плотностью. Для вычисления объема кластерной структуры можно воспользоваться хорошо известным методом Монте-Карло.

В n-мерном пространстве кластерная структура представляет собой геометрическое тело, которое вписано в прямоугольный гиперпараллелепипед. Длина стороны этого гиперпараллелепипеда по j-й координате будет вычисляться как разность $\max_j - \min_j$, где $\max_j$ – максимальное среди всех значений вида $(z^i_j + r^i)$, $i=\overline{1,k}$, $j=\overline{1,n}$, а $\min_j$ – минимальное среди всех значений вида $(z^i_j - r^i)$, $i=\overline{1,k}$, $j=\overline{1,n}$. Поскольку известны диапазоны изменения значений координат прямоугольного гиперпараллелепипеда, в который вписана кластерная структура, и имеется формальное описание в табличном виде кластерной структуры, то можно реализовать процедуру вычисления ее объема, применив метод Монте-Карло.

В случае когда все вершины $z^1,...,z^k$ лежат на одной прямой, для вычисления объема кластерной структуры можно воспользоваться формулой $V = \sum_{j=1}^{2k-1} V^{(j)} - U$, где $V^{(j)}$ – объем гипершара со значениями координат центра и радиусом из j-й строки таблицы; $U = \sum_{j=1}^{2k-2} U^{(j)}$, где $U^{(j)}$ – объем областей пересечения гипершаров с значениями координат центра и радиусом из j-й и (j+1)-й строк таблицы. Формулы вычисления объема гипершара в n-мерном пространстве различаются для четных и нечетных значений n. Для четных значений n объем вычисляется по формуле $V = \frac{2^{\frac{n}{2}}\pi^{\frac{n}{2}}}{n!!} r^n$, а для нечетных – по формуле $V = \frac{2^{\frac{n+1}{2}}\pi^{\frac{n-1}{2}}}{n!!} r^n$, где r – радиус гипершара, а n!! – двойной факториал [13].

Плотность кластерной структуры вычисляется как отношение количества вершин, на основе которых был построен каркас, к объему кластерной структуры. Чем выше значение плотности, тем более компактно размещено соответствующее подмножество в пространстве.

4. Использование кластерных структур для решения задач распознавания образов с обучением

Математическая постановка задачи распознавания выглядит следующим образом: пусть X – множество описаний объектов, Y – множество номеров (или наименований) классов. Существует неизвестная целевая зависимость, т. е. отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Требуется построить алгоритм, способный для произвольного объекта $x \in X$ определить его принадлежность одному из классов [14].

Для решения задачи распознавания можно воспользоваться гипотезой компактности, которая гласит, что образам классов соответствуют компактные множества в пространстве признаков [15].

Поскольку обучающая выборка содержит информацию о каждом из классов в виде множества соответствующих объектов, на их основе следует построить образы классов в виде кластерных структур. Если все полученные таким образом кластерные структуры оказываются взаимно непересекающимися или пересечение не превышает некоторого допустимого порогового значения, считается, что процедура обучения прошла успешно, т. е. в исходном признаковом пространстве образы классов размещены компактно и не пересекаются.

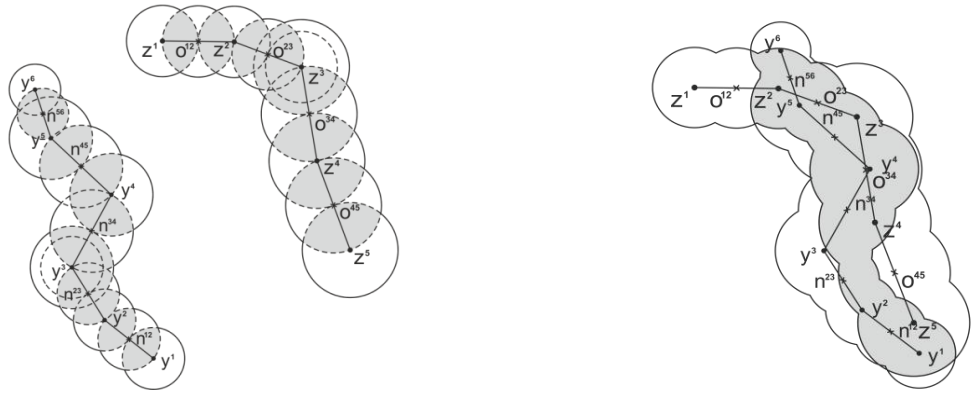
Общая схема процесса распознавания выглядит следующим образом. Произвольный объект с неизвестной классификацией представляется в виде кластерной структуры и проводится оценка ее размещения по отношению к кластерным структурам классов. Далее неклассифицированный объект либо относится к одному из существующих классов, либо на его основе формируется отдельный «джокер-класс».

Если образы классов в исходном признаковом пространстве плохо разделяются, то необходимо перейти к поиску подпространства, в котором выполнялись бы условия гипотезы компактности. В этом случае из исходного пространства набора признаков последовательно выбираются подмножества признаков и каждое подмножество анализируется на выполнение гипотезы компактности. Процедура такого анализа базируется на использовании кластерных структур, которые строятся и сравниваются в соответствующем признаковом подпространстве.

Если в результате находится признаковое подпространство, в котором выполняются условия гипотезы компактности, то переходим к классификации исследуемого объекта в этом пространстве.

В соответствии с математической постановкой задачи распознавания экземпляры *i*-го класса образуют матрицу размерности $n \times m_i$, где n – количество признаков, m_i – количество объектов *i*-го класса, а обучающая выборка представляет собой объединение соответствующих матриц, т. е. $X_{n \times m} = \bigcup_{i=1}^{k} X^i_{n \times m_i}$, где k – количество классов, $m = m_1 + m_2 + \ldots + m_k$.

Построив, например, на основе матриц $X^{i}_{n \times m_i}$ и $X^{j}_{n \times m_j}$ две кластерные структуры (рис. 4), можно получить оценку взаимного расположения образов классов в признаковом пространстве.



а) *б)*

Рис. 4. Пример взаимного размещения двух плоских кластерных структур:
а) непересекающихся; *б)* пересекающихся

Численную оценку взаимного пересечения двух кластерных структур можно представить в виде отношения $Q = \dfrac{U}{V}$, где U – объем области их пересечения, а $V = \min\{V_1, V_2\}$ (V_1 и V_2 – объемы первой и второй кластерных структур соответственно). Очевидно, что $0 \le Q \le 1$.

Для заданной обучающей выборки необходимо выполнить C_k^2 попарных сравнений кластерных структур. Если все Q_i не превышают заданного порогового значения, то это означает, что для построенных образов классов выполняется условие гипотезы компактности и можно переходить к этапу распознавания.

В случае когда хотя бы одно из значений Q_i превысит пороговое значение, переходим к поиску подпространств, в которых выполняется гипотеза компактности.

Если исходное признаковое пространство содержит n признаков, то для анализа на выполнение гипотезы компактности необходимо исследовать $L = \sum_{i=1}^{n-1} C_n^i$ всевозможных подпространств. При этом исследование каждого подпространства строится аналогично уже описанному выше процессу.

В общем случае может быть найдено $0 \le L^* \le L$ признаковых подпространств, в которых выполняется гипотеза компактности. В каждом из выявленных подпространств действует «скрытая закономерность»: комбинация соответствующих признаков обеспечивает разделение образов классов в признаковом пространстве. Такая «скрытая закономерность», с одной стороны, может быть формально представлена комбинацией кластерных структур, а с другой – проинтерпретирована в терминах предметной области.

Таким образом, описанный процесс решения задачи распознавания позволяет в автоматическом режиме выполнять процедуру извлечения и представления «скрытых закономерностей» на основе анализа содержимого обучающей выборки.

Следует отметить, что может возникнуть случай, когда L^*=0, т. е. не найдено подпространств, для которых выполняется условие гипотезы компактности, в этом случае необходимо вернуться к выбору нового набора признаков.

Заключение

В статье рассмотрены вопросы извлечения и представления скрытых закономерностей в интеллектуальном анализе данных. Анализ существующих моделей представления знаний показал актуальность поиска новых подходов, обеспечивающих автоматическое выполнение этапов обработки данных с целью извлечения и представления ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний.

Для решения в автоматическом режиме задачи извлечения и представления скрытых закономерностей и задачи распознавания образов с обучением предложен новый подход, который базируется на использовании специальным образом построенных кластерных структур.

Описан метод построения кластерных структур и продемонстрирована возможность их применения для построения практически полезных моделей представления знаний, образов классов и объектов в интеллектуальном анализе данных. Предложен вариант автоматического решения задачи распознавания образов с обучением на основе кластерных структур.

Список литературы

1. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б. Паклин, В.И. Орешков. – СПб. : Питер, 2013. – 704 с.

2. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск : ИМ СО РАН, 1999. – 270 с.

3. Журавлёв, Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлёв, В.В. Рязанов, О.В. Сенько. – М. : Фазис, 2006. –176 с.

4. Краснопрошин, В.В. Система понятий в информатике / В.В. Краснопрошин, О.А. Маркова, А.Н. Вальвачев // Информатика. – 2007. – № 3. – С. 124–130.

5. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян [и др.]. – СПб. : БХВ-Петербург, 2004. – 336 с.

6. Дюк, В. Обработка данных на ПК в примерах / В. Дюк. – СПб. : Питер, 1997. – 240 с.

7. Классификация моделей представления знаний [Электронный ресурс]. – 2015. – Режим доступа : http://www.aiportal.ru/articles/knowledge-models/classification.html. – Дата доступа : 10.03.2015.

8. Neweel, A. Computer science as empirical inquiry: symbols and search / A. Neweel, H. Simon // Communications of the ACM. – 1976. – № 19(3). – P. 113–126.

9. Люгер, Дж.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем / Дж.Ф. Люгер. – М. : Изд. дом «Вильямс», 2005. – 864 с.

10. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – М. : Изд. дом «Вильямс», 2006. – 1104 с.

11. Краснопрошин, В.В. Проблема принятия решений по прецедентности: разрешимость и выбор алгоритмов / В.В. Краснопрошин, В.А. Образцов // Выбраныя навуковыя працы Беларускага дзяржаўнага ўніверсітэта. – Минск : БДУ, 2001. – Т. 6. – С. 285–311.

12. Алгоритмы: построение и анализ / Т. Кормен [и др.] ; пер. с англ. – 2-е изд. – М. : Изд. дом «Вильямс», 2005. – 1296 с.

13. Розенфельд, Б.А. Многомерные пространства / Б.А. Розенфельд. – М. : Наука, 1966. – 647 с.

14. Задача классификации [Электронный ресурс]. – 2015. – Режим доступа : http://ru.wikipedia.org/wiki/Задача_классификации. – Дата доступа : 25.05.2015.

15. Гипотеза компактности [Электронный ресурс]. – 2015. – Режим доступа : htth://www.machinelearning.ru/wiki/index.php?title=Гипотеза_компактности. – Дата доступа : 25.05.2015.

Поступила 25.01.2016

Белорусский государственный университет,
Минск, пр. Независимости, 4
e-mail: Krasnoproshin@bsu.by,
rovar@mail.ru

V.V. Krasnoproshin, V.G. Rodchanka

CLUSTER STRUCTURES AND THEIR APPLICATIONS IN DATA MINING

A notion of cluster structures is introduced. These structures can be used to identify and present the "hidden knowledge" in data mining. It demonstrates the use of cluster structures to address automatic pattern recognition with training.