УДК 519.237.8.:004.8

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЭЛЕКТРОННЫХ ХРАНИЛИЩАХ

Г. Г. Асеев

Харьковская государственная академия культуры, г.Харьков, Украина
aseev@ic.ac.kharkov.ua

Сегодня наблюдается небывалый подъем комплексной компьютеризации корпоративных организаций, причем первостепенная роль отводится построению автоматизированных систем документооборота. В статье рассматриваются современные методы статистики для построения задач классификации и многомерного анализа данных в хранилищах электронного документооборота.

СЫРЫЕ ДАННЫЕ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, КЛАСТЕРНЫЙ АНАЛИЗ, МЕРА БЛИЗОСТИ, АЛГОРИТМ ТАКСОНОМИИ, АССОЦИАТИВНЫЕ ПРАВИЛА, ТРАНЗАКЦИЯ

Введение

Кластерный анализ применяется для исследования и обнаружения «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных электронных хранилищ скрытых структур или зависимостей, которые: ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации знаний человеком, необходимых для принятия решений в различных сферах человеческой деятельности.

Ранее была рассмотрена проблема обнаружения нового знания в хранилищах данных методами Knowledge Discovery in Databases (KDD) и Data Mining [1, 2].

1. Методы извлечения и анализа данных

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других и так далее. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность заключается в том, чтобы эти знания могли принести определенную выгоду при их применении. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Методы извлечения и анализа данных делятся на **описательные**: *математическую статистику* (оценивание параметров распределения; проверка статистических гипотез; дисперсионный и регрессионный анализ; анализ временных рядов, который, в свою очередь, состоит из спектрального и корреляционного анализа и фильтрации; многомерный анализ, который, в свою очередь, состоит из кластерного, дискриминантного и факторного анализа, метода главных компонент и шкалирование) [3-7] и **предсказательные:** *эволюционное моделирование* (генетические алгоритмы; искусственные нейронные сети, которые, в свою очередь, делятся на RBF и ART сети; сети обратного и встречного распространения; сети Хемминга, Хопфилда, Кохонена и гибридно нечетко-нейронные и пр.); *машинное обучение* (деревья решений, которые, в свою очередь, делятся на энтропийную меру, ID3, C4.5, NewID и прочее) [8]. В связи с тем, что в короткой статье мы не в состоянии провести анализ, сравнение и рекомендации практического применения всех перечисленных методов, пока ограничимся основными моделями извлечения и анализа данных с помощью кластерного анализа и метода k_n ближайших соседей. Остальные модели будут рассмотрены в следующих публикациях.

2. Кластерный анализ

Задача кластерного анализа [9–11] - выделение групп в заданной (обучающей) совокупности элементов (путем итерационного слияния наиболее близких кластеров).

Кластерный анализ позволяет группировать данные вокруг нескольких центров в n–мерном пространстве. В литературе [9-11] описываются методы кластеризации полным перебором (теоретически), методами математического программирования, на основе матриц сходств, на основе оценивания функции плотности.

Первая группа методов, которые будут рассмотрены – *алгоритмы таксономии*. Кластер определяется как совокупность элементов, лежащих на расстоянии не больше r от центра (внутри гиперсферы радиуса r или гиперкуба со сторонами $2r$). При этом в качестве центра выбирается один из элементов и формируется кластер из элементов, удаленных от него не далее, чем на r. Боннер предлагает выбирать очередной центр случайно, Хиверинен предлагает в качестве очередного центра брать «типическую» точку – лежащую на минимальном расстоянии от центра оставшегося множества объектов. Далее процедура повторяется для оставшихся элементов. Элементы, не попавшие ни в один кластер после определенного числа шагов или образования кластеров с требуемыми показателями, считаются нераспознанными, могут трактоваться как шум

в системе распознавания и исключаются из обучающей совокупности элементов.

Другая важная группа эвристических методов кластеризации – методы, основывающиеся на *последовательной агломеративной процедуре*.

Все эти методы дают оптимальное решение в классе меньшем, чем класс всех возможных разбиений (кластеров), однако достоинством этих методов является простота вычислительной процедуры и алгоритмов.

Пусть анализируемая совокупность состоит из p элементов, каждый из которых характеризуется значениями n дискриминантных переменных.

На первом шаге итеративной процедуры имеется p кластеров, каждый из которых включает по одному элементу.

Определяются два наиболее близких или сходных кластера, объединяются в один кластер, количество кластеров сокращается на 1: $p \quad p-1$.

Мера близости определяется расстоянием между элементами, заносимыми в симметричную матрицу расстояний D:

$$\mathrm{D} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & & \cdot \\ d_{p1} & d_{p2} & \cdots & 0 \end{pmatrix}.$$

Наиболее близкими считаются объекты с *наименьшим расстоянием*.

Расстояние между точками, между центрами кластеров определяются разными метриками. Так, расстояние между l-ой и r-ой точками в евклидовой метрике равно: $d_{rj}^{(2)} = \sqrt{\sum_{i=1}^{n}(x_{ri} - x_{li})^2}$.

Другие метрики.

Норма:

$$d_{rj}^{(1)} = \sum_{i=1}^{n}\left(x_{ri} - x_{li}\right) - l_1.$$

Супремум-норма:

$$d_{rj}^{(\infty)} = \sup_{i=1,2,\cdots,n}\left\{\left|x_{ri} - x_{li}\right|\right\}.$$

Норма, которая охватывает и предыдущие (при $p = 2$, $p = 3$ и $p = \infty$):

$$d_{rj}^{(p)} = \left[\sum_{i=1}^{n}\left(\left|x_{ri} - x_{li}\right|^p\right)\right]^{\frac{1}{p}} - l_p.$$

Махаланобиса:

$$D^2\left(X_i, X_j\right) = \left(X_i - X_j\right)^T \mathrm{W}^{-1}\left(X_i - X_j\right),$$

где W - матрица рассеяния.

Возможны и другие способы определения *расстояния между классами*. Пусть объекты $\{X_i\}, i = 1,\ldots,k_1$ принадлежат к одному классу, а $\{Y_j\}, j = 1,\ldots,k_2$ – к другому.

Минимальное локальное расстояние:

$$D_1 = \min_{\substack{i=1,\ldots,k_1 \\ j=1,\ldots,k_2}} d\left(X_i, X_j\right).$$

Максимальное локальное расстояние:

$$D_2 = \max_{\substack{i=1,\ldots,k_1 \\ j=1,\ldots,k_2}} d\left(X_i, X_j\right).$$

Среднее расстояние:

$$D_3 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\frac{d\left(X_i, X_j\right)}{k_1 k_2}.$$

Статистическое расстояние между кластерами:

$$D_4 = \frac{k_1 k_2}{k_1 + k_2}\left(\overline{X} - \overline{Y}\right)^T\left(\overline{X} - \overline{Y}\right),$$

где

$$\overline{X} = \sum_{i=1}^{k_1}\frac{x_i}{k_1}, \quad \overline{Y} = \sum_{j=1}^{k_2}\frac{y_j}{k_2}.$$

Расстояние между центроидами:

$$D_5 = \left(\overline{X} - \overline{Y}\right)^T\left(\overline{X} - \overline{Y}\right).$$

На каждом последующем шаге агломеративной процедуры потребуется пересчет лишь для одной строки (и одного столбца) D, то есть рассчитываются расстояния от образованного кластера до каждого из оставшихся кластеров. Существует несколько методов пересчета расстояний с использованием старых значений расстояний для объединяемых кластеров, отличающихся коэффициентами в формуле:

$$d_{rs} = \alpha_p d_{ps} + \alpha_q d_{qs} + \beta d_{pq} + \gamma\left|d_{ps} - d_{qs}\right|.$$

Если кластеры p и q объединяются в кластер r и требуется рассчитать расстояние от нового кластера до кластера s, применение того или другого метода зависит от способа определения расстояния между кластерами, различные методы различаются значениями коэффициентов $\alpha_p, \alpha_q, \beta$ и γ.

Основанием для слияния кластеров может быть не только мера их близости (расстояние между ними), но и *мера сходства* – неотрицательная функция $s\left(X_i, X_j\right) = s_{ij}$, причем $0 \le s_{ij} \le 1, i \ne j, s_{ii} = 1, s_{ij} = s_{ji}$. Меры сходства всех кластеров объединяются в симметричную матрицу S:

$$\mathrm{S} = \begin{pmatrix} 1 & s_{12} & \cdots & s_{1p} \\ s_{21} & 1 & \cdots & s_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & & \cdot \\ s_{p1} & s_{p2} & \cdots & 1 \end{pmatrix}.$$

Наиболее сходными считаются объекты со значением s_{ij} наиболее близким к 1, процедура образования кластеров при этом аналогична описанной выше с использованием меры близости.

В качестве меры сходства можно использовать *коэффициент корреляции:*

$$r_{ij} = \sum_{l=1}^{n} \overset{\circ}{x}_{li} \overset{\circ}{x}_{lj} \Big/ \sqrt{\sum_{l=1}^{n} \overset{\circ}{x}_{li}^{\,2} \sum_{l=1}^{n} \overset{\circ}{x}_{lj}^{\,2}} ,$$

причем

$$\overset{\circ}{x}_{li} = x_{li} - \bar{x}_i, \quad \bar{x}_i = \sum_{l=1}^{n} x_{li}, \quad \overset{\circ}{x}_{lj} = x_{lj} - \bar{x}_j, \quad \bar{x}_j = \sum_{jl=1}^{n} x_{lj}.$$

После выполнения очередного шага агломеративной процедуры выясняется, что желательное разбиение достигнуто. Существуют различные методы определения критерия *остановки процедуры*:

получено определенное заранее количество кластеров;

все кластеры содержат более определенного числа элементов;

кластеры обладают требуемым соотношением внутренней однородности и разнородности между собой.

После анализа обучающей выборки можно решать задачу классификации новых состояний $\{X_i\}$, относя их к той или другой группе, что определяется по минимальному расстоянию до центров групп.

3. Метод k_n ближайших соседей

Здесь идея состоит в том, что вокруг распознаваемого объекта $\bar{x}$ строится ячейка объёма V. При этом неизвестный объект относится к тому образу, число обучающих представителей которого в построенной ячейке оказалось в большинстве. Если использовать статистическую терминологию, то число объектов образа s_i, попавших в данную ячейку, характеризует оценку усреднённой по объёму V плотности вероятности $p(\bar{x}/s_i)$ [12].

Для оценки усреднённых $p(\bar{x}/s_i)$ нужно решить вопрос о соотношении между объёмом V ячейки и количеством попавших в эту ячейку объектов того или иного класса (образа). Вполне разумно считать, что чем меньше V, тем более тонко будет охарактеризована $p(\bar{x}/s_i)$. Но при этом тем меньше объектов попадёт в интересующую нас ячейку, а следовательно, тем меньше достоверность оценки $p(\bar{x}/s_i)$. При чрезмерном увеличении V возрастает достоверность оценки $p(\bar{x}/s_i)$, но теряются тонкости её описания из-за усреднения по слишком большому объёму, что может привести к негативным последствиям (увеличению вероятности ошибок распознавания). При небольшом объёме обучающей выборки V целесообразно брать предельно большим, но обеспечить при этом, чтобы внутри ячейки плотности $p(\bar{x}/s_i)$ мало изменялись. Тогда их усреднение по большому объёму не очень опасно. Таким образом, вполне может случиться, что объём ячейки, уместный для одного значения $\bar{x}$, может совершенно не годиться для других случаев.

Предлагается следующий порядок действий (пока что принадлежность объекта тому или иному образу учитывать не будем).

Для того чтобы оценить $p(\bar{x})$ на основании обучающей выборки, содержащей n объектов, центрируем ячейку вокруг $\bar{x}$ и увеличиваем её объём до тех пор, пока она не вместит k_n объектов, где k_n есть некоторая функция от n. Эти k_n объектов будут ближайшими соседями $\bar{x}$. Вероятность P попадания вектора $\bar{x}$ в область R определяется выражением $P = \int_R p(\bar{x}')d\bar{x}'$.

Это сглаженный (усреднённый) вариант плотности распределения $p(\bar{x})$. Если взять выборку из n объектов (простым случайным выбором из генеральной совокупности), то k из них окажется внутри области R. Вероятность попадания k из n объектов в R описывается биномиальным законом, имеющим резко выраженный максимум около среднего значения nP. При этом k/n является неплохой оценкой для P.

Если теперь допустить, что R настолько мала, что $p(\bar{x})$ внутри неё меняется незначительно, то

$$\int_R p(\bar{x}')d\bar{x}' \approx p(\bar{x})V ,$$

где V – объём области R; $\bar{x}$ – точка внутри R.

Тогда $P \approx p(\bar{x})V$. Но $P \approx \frac{k}{n}$, следовательно, $p(\bar{x}) \approx \frac{k/n}{V}$.

Итак, оценкой $p_n(\bar{x})$ плотности $p(\bar{x})$ является величина

$$p_n(\bar{x}) = \frac{k_n/n}{V_n} . \qquad (*)$$

Без доказательства приведём утверждение, что условия

$$\lim_{n\to\infty} k_n = \infty \text{ и } \lim_{n\to\infty} \frac{k_n}{n} = 0 \qquad (**)$$

являются необходимыми и достаточными для сходимости $p_n(\bar{x})$ к $p(\bar{x})$ по вероятности во всех точках, где плотность $p(\bar{x})$ непрерывна.

Этому условию удовлетворяет, например, $k_n = \sqrt{n}$.

Теперь будем учитывать принадлежность объектов к тому или иному образу и попытаемся оценить апостериорные вероятности образов $p(s_i/\bar{x})$.

Предположим, что мы размещаем ячейку объёма V вокруг $\bar{x}$ и захватываем выборку с количеством объектов k_n, k_{ni} из которых принадлежат образу s_i. Тогда в соответствии с формулой (*) оценкой совместной вероятности $p(\bar{x},s_i)$ будет величина

$$p_n(\bar{x},s_i)=\frac{k_{ni}/n}{V_n},$$

а

$$P_n(s_i/\bar{x})=\frac{p_n(\bar{x},s_i)}{\sum_{j=1}^{M}p_n(\bar{x},s_j)}=\frac{k_{ni}}{k_n}.$$

Таким образом, апостериорная вероятность $P_n(s_i/\bar{x})$ оценивается как доля выборки в ячейке, относящаяся к s_i. Чтобы свести уровень ошибки к минимуму, нужно объект с координатами $\bar{x}$ отнести к классу (образу), количество объектов обучающей выборки которого в ячейке максимально. При $n\to\infty$ такое правило является байесовским, то есть обеспечивает теоретический минимум вероятности ошибок распознавания (разумеется, при этом должны выполняться условия(**)).

Правило ближайшего соседа. Пусть $X_n=\{\bar{x}_1,\bar{x}_2,...,\bar{x}_n\}$ – множество объектов обучающей последовательности, то есть принадлежность каждого из них тому или иному образу достоверно известна. Пусть также $\bar{x}*\in X_n$ является объектом, ближайшим к распознаваемому $\bar{x}\notin X_n$. Напомним, что при этом правило ближайшего соседа для классификации $\bar{x}$ состоит в том, что $\bar{x}$ относят к тому классу (образу), которому принадлежит $\bar{x}*$. Естественно, такое отнесение носит случайный характер. Вероятность того, что $\bar{x}$ будет отнесён к s_i, есть апостериорная вероятность $P(s_i/\bar{x}*)$. Если n очень велико, то вполне можно допустить, что $\bar{x}$ расположен достаточно близко к $\bar{x}*$, настолько близко, что $P(s_i/\bar{x}*)\approx P(s_i/\bar{x})$. А это есть не что иное, как рандомизированное решающее правило: $\bar{x}$ относят к s_i с вероятностью $P(s_i/\bar{x})$. Байесовское решающее правило основано на выборе максимальной апостериорной вероятности, то есть $\bar{x}$ относят к s_i в том случае, если

$$P(s_j/\bar{x})=\max_i P(s_i/\bar{x}).$$

Отсюда видно, что если $P(s_j/\bar{x})$ близка к единице, то правило ближайшего соседа даёт решение, в большинстве случаев совпадающее с байесовским. Напомним, что эти рассуждения имеют достаточные основания лишь при очень больших n (объёмах обучающей выборки). Такие условия на практике встречаются не так часто, но позволяют понять статистический смысл правила ближайшего соседа.

4. Анализ рыночной корзины

Анализом рыночной корзины называется задача поиска ассоциативных правил, которые описывают типичные шаблоны покупок, совершаемых в супермаркетах. Задача поиска ассоциативных правил впервые была представлена для анализа рыночной корзины, за что и получила свое название. Однако, сфера применения не ограничивается лишь одной торговлей. Ассоциативные правила также успешно применяют и в других областях: медицине, для анализа посещений веб-страниц (Web Mining), для анализа текста (Text Mining) для анализа данных по переписи населения [13], в анализе и прогнозировании сбоев телекоммуникационного оборудования и так далее.

Ассоциативные правила (*Association Rules*) [8, 14].

Впервые эта задача была предложена при поиске ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит. Такую транзакцию еще называют рыночной корзиной.

Определение 1. Пусть $I=\{i_1,i_2,i_3,...,i_n\}$ – множество (набор) товаров, называемых элементами. Пусть D – множество транзакций, где каждая транзакция T – это набор элементов из I, $T\subseteq I$. Каждая транзакция представляет собой бинарный вектор, где $t[k]$=1, если i_k элемент присутствует в транзакции, иначе $t[k]$=0. Мы говорим, что транзакция T содержит X, некоторый набор элементов из I, если $X\subseteq T$. Ассоциативным правилом называется импликация $X\Rightarrow Y$, где $X\subset I$, $Y\subset I$ и $X\cap Y=\varnothing$. Правило $X\Rightarrow Y$ имеет поддержку s (support), если s% транзакций из D, содержат $X\cup Y$, supp ($X\Rightarrow Y$) = supp ($X\cup Y$). Достоверность правила показывает какова вероятность того, что из X следует Y. Правило $X\Rightarrow Y$ справедливо с достоверностью (confidence) c, если c% транзакций из D, содержащих X, также содержат Y, conf ($X\Rightarrow Y$) = = supp ($X\cup Y$)/supp (X).

Покажем на конкретном примере: «75% транзакций, содержащих одну книгу при покупке, также содержат две одновременно. 3% от общего числа всех транзакций содержат обе книги». 75% – это достоверность (confidence) правила, 3% это поддержка (support), или «одна книга» ⇒ «две книги» с вероятностью 75%.

Другими словами, целью анализа является установление следующих зависимостей: если в транзакции

встретился некоторый набор элементов X, то на основании этого можно сделать вывод о том, что другой набор элементов Y также же должен появиться в этой транзакции. Установление таких зависимостей дает нам возможность находить очень простые и интуитивно понятные правила.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил $X \Rightarrow Y$, причем поддержка и достоверность этих правил должны быть выше некоторых наперед определенных порогов, называемых соответственно минимальной поддержкой (minsupport) и минимальной достоверностью (minconfidence).

Задача нахождения ассоциативных правил разбивается на две подзадачи:

– нахождение всех наборов элементов, которые удовлетворяют порогу minsupport. Такие наборы элементов называются часто встречающимися;

– генерация правил из наборов элементов, найденных согласно п. 1. с достоверностью, удовлетворяющей порогу minconfidence.

Один из первых алгоритмов, эффективно решающих подобный класс задач – это алгоритм APriori [15]. Кроме этого алгоритма в последнее время был разработан ряд других алгоритмов: DHP [16], Partition [19], DIC [13] и другие [20].

Значения для параметров минимальной поддержки и минимальной достоверности выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Тем не менее, большинство интересных правил находится именно при низком значении порога поддержки. Хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил.

Поиск ассоциативных правил совсем не тривиальная задача, как может показаться на первый взгляд. Одна из проблем – алгоритмическая сложность при нахождении часто встречающих наборов элементов, так как с ростом числа элементов в I ($|I|$) экспоненциально растет число потенциальных наборов элементов.

Обобщенные ассоциативные правила (*Generalized Association Rules*).

При поиске ассоциативных правил мы предполагали, что все анализируемые элементы однородны. Возвращаясь к анализу рыночной корзины, это товары, имеющие совершенно одинаковые атрибуты, за исключением названия. Однако не составит большого труда дополнить транзакцию информацией о том, в какую товарную группу входит товар и построить иерархию товаров.

Пусть нам дана база транзакций D и известно в какие группы (таксоны) входят элементы. Тогда можно извлекать из данных правила, связывающие группы с группами, отдельные элементы с группами и так далее.

Например, если «Покупатель купил товар из группы «Безалкогольные напитки», то он купит и товар из группы «Молочные продукты» или «Сок» $\Rightarrow$ «Молочные продукты»». Эти правила носят название обобщенных ассоциативных правил.

Определение 2. Обобщенным ассоциативным правилом называется импликация $X \Rightarrow Y$, где $X \subset I$, $Y \subset I$ и $X \cap Y = \varnothing$ и где ни один из элементов, входящих в набор Y, не является предком ни одного элемента, входящего в X. Поддержка и достоверность подсчитываются так же, как и в случае ассоциативных правил (см. *Определение 1*).

Введение дополнительной информации о группировке элементов в виде иерархии даст следующие преимущества:

– это помогает установить ассоциативные правила не только между отдельными элементами, но и между различными уровнями иерархии (группами);

– отдельные элементы могут иметь недостаточную поддержку, но в целом группа может удовлетворять порогу minsupport.

Для нахождения таких правил можно использовать любой из вышеназванных алгоритмов. Для этого каждую транзакцию нужно дополнить всеми предками каждого элемента, входящего в транзакцию. Однако применение «в лоб» этих алгоритмов неизбежно приведет к следующим проблемам:

– элементы на верхних уровнях иерархии стремятся к значительно большим значениям поддержки по сравнению с элементами на нижних уровнях;

– с добавлением в транзакции групп увеличилось количество атрибутов и соответственно размерность входного пространства. Это усложняет задачу, а также ведет к генерации большего количества правил.

Появление избыточных правил, противоречащих определению обобщенного ассоциативного правила, например, «Сок» $\Rightarrow$ «Прохладительные напитки». Очевидно, что практическая ценность такого «открытия» нулевая при 100% достоверности. Следовательно, нужны специальные операторы, удаляющие подобные избыточные правила.

Для нахождения обобщенных ассоциативных правил желательно использование специализированного алгоритма [17], который устраняет вышеописанные проблемы и к тому же работает в 2–5 раз быстрее, чем стандартный APriori.

Группировать элементы можно не только по вхождению в определенную товарную группу, но и по другим характеристикам, например по цене (дешево, дорого), брэнду и так далее.

Численные ассоциативные правила (*Quantitative Association Rules*).

При поиске ассоциативных правил задача была существенно упрощена. По сути все сводилось к тому, присутствует в транзакции элемент или нет. То есть, если рассматривать случай рыночной корзины, то мы рассматривали два состояния: куплен товар или нет, проигнорировав, например, информацию о том, сколько было куплено, кто купил, характеристики покупателя и так далее. И можно сказать, что рассматривали «булевские» ассоциативные правила. Если взять любую базу данных, каждая транзакция состоит из различных типов данных: числовых, категориальных и так далее. Для обработки таких записей и извлечения численных ассоциативных правил был предложен алгоритм поиска [18].

Пример численного ассоциативного правила:

[Возраст: 30–35] и [Семейное положение: женат] ⇒ [Месячный доход: 2000–2500 грн].

Помимо описанных выше ассоциативных правил существуют косвенные ассоциативные правила, ассоциативные правила с отрицанием, временные ассоциативные правила для событий связанных во времени и другие.

Выводы

Как указывалось выше, существуют два вида моделей: предсказательные и описательные. Описательная модель не сможет претендовать на абсолютное знание, но даст аналитику некоторое преимущество уже самим фактом обнаружения альтернативного статистически значимого описания. Даже богатый арсенал классической статистики используется далеко не полностью, не говоря уже о более современных методах нелинейного анализа.

Если конкретные данные хранилища данных характеризуются конечной совокупностью значимых факторов, определяющих анализируемый процесс или развитие процесса, которые могут быть объективно представлены, то в этом случае в задачах многомерного анализа можно успешно применять современные методы статистики с классификатором, построенным по принципу максимального правдоподобия на основе кластерного анализа, метода k_n ближайших соседей и анализа рыночной корзины.

Список литературы: 1. *Арсеньев, С. Б.* Использование технологии анализа данных в интеллектуальных информационных системах. Управление информационными потоками / С. Б. Арсеньев, В. Б. Бритков, Н. А. Маленкова // Сб. тр. Инта систем. анализа РАН. - М.: Эдиториал УРСС, 2002. - С. 47-68. **2.** *Асеев Г. Г.* Проблема обнаружения нового знания в хранилищах данных методами Knowledge Discovery in Databases / Г.Г. Асеев // Вестник НТУ «ХПИ». - Х.: НТУ «ХПИ», 2006. № 19. - С. 62-70. **3.** *Соколов Г.А., Гладких И.М.* Математическая статистика: учеб. / Г.А. Соколов, И.М. Гладких - М: Экзамен, 2004. - 432 с. **4.** *Елисеева И.И., Егорова И.И., Курышева С.В. и др.* Статистика: учеб. / Под ред. И.И. Елисеевой. - М: ТК Велби / Проспект, 2004. - 448 с. **5.** Статистика: учеб. / Под ред. В.С. Мхитаряна. - М: Экономист, 2005. - 671 с. **6.** *Орлов А.И.* Прикладная статистика: учеб. / Орлов А.И. - М.: Экзамен. 2006. - 452 с. **7.** *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям / Н. Хастингс, Дж. Пикок. - М.: Статистика, 1990. - 480 с. **8.** *Асеев Г.Г.* Электронный документооборот: учеб. / Г.Г. Асеев. - К.: Кондор, 2007. - 500с. **9.** Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др. - М.: Финансы и статистика, 2003. **10.** *Дюран Б.* Кластерный анализ / Б. Дюран, П. Оделл. - М.: Статистика, 2007. **11.** *Елисеева И. И.* Группировка, корреляция, распознавание образов / И. И. Елисеева, В. О. Рукавишников. - М.: Статистика, 1999. **12.** *Волошин Г. Я.* Методы распознавания образов (конспект лекций) [Электронный ресурс] / Г. Я. Волошин. - Режим доступа: http://disser.h10.ru/raspobraz.html. - Загл. с экрана. **13.** Dynamic Itemset Counting and Implication Rules for Market Basket Data / S. Brin etc. // In Proc. ACM SIGMOD Int'l Conf. Management of Data. - NY: ACM Press, 1999. **14.** *Шахиди А.* Введение в анализ ассоциативных правил [Электронный ресурс] / А. Шахиди. - Режим доступа: http://www.basegroup.ru/rules/ intro.html. - Загл. с экрана. **15.** *Agrawal R.* Fast Discovery of Association Rules / R. Agrawal, R. Srikant // In Proc. of the 20th International Conference on VLDB. - Santiago, Chile, September 2004. **16.** *Savasere E.* An Efficient Algorithm for Mining Association Rules in Large Databases / E. Savasere, E. Omiecinski, S. Navathe // In Proc. 21st Int'l Conf. Very Large Data Bases. - San Francisco: Morgan Kaufmann, 2005. **17.** *Srikant R.* Mining Generalized Association Rules / R. Srikant, R. Agrawal // In Proc. of the 21th International Conference on VLDB. - Zurich, Switzerland, 1995. **18.** *Srikant R.* Mining quantitative association rules in large relational tables / R. Srikant, R. Agrawal // In Proceedings of the ACM SIGMOD Conf. on Management of Data. - Montreal, Canada, June, 2006. **19.** *Park J. S.* An Effective HashBased Algorithm for Mining Association Rules / J. S. Park, M.-S Chen, S. Y. Philip // In Proc. ACM SIGMOD Int'l Conf. Management of Data. - NY: ACM Press, 2005. **20.** *Bremermann H. J.* Global properties of evolution processes. Natural automata and useful simulations / H. J. Bremermann, J. Roghson, S. Salaff. - London: Macmillan, 2006. - P. 3-42.

Поступила в редколлегию 29.01.2009

УДК 519.237.8.:004.8

Методи інтелектуального аналізу даних в електронних сховищах / Г. Г. Асєєв // Біоніка інтелекту: наук.-техн. журнал.- 2009.- №1(70).- С. 28-33.

Стаття присвячена проблемі дослідження і виявлення «машиною» у сирих даних схованих структур або залежностей, якщо конкретні дані сховища даних характеризуються кінцевою сукупністю значущих факторів, що обумовлюють аналізований процес чи розвиток процесу, методами кластерного аналізу, асоціативних правил та іншими.

Бібліогр.: 20 найм.

UDC 519.237.8.:004.8

Methods of intellectual analysis of data in electronic warehouses / G. G. Aseyev // Bionics of Intelligence: Sci. Mag. - 2009.- №1(70).- P. 28-33.

An unprecedented increase of complex computerization of corporate organizations is being observed at present, and at that the development of automated systems of document flow is of great importance. The paper deals with up-to-date methods of statistics for development of classification tasks and multivariate data analysis in electronic document warehouses.

Ref.: 20 item.