УДК 311.312

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

М. В. Инкин, магистр 1 года обучения по программе «Управление качеством», экономический факультет ГОУВПО «Мордовский государственный университет им. Н.П. Огарева»
Е. С. Петрова, канд. экон. наук, доцент кафедры информационных систем в экономике и управлении экономического факультета ГОУВПО «Мордовский государственный университет им. Н.П. Огарева»

Одним из основных подходов в «обнаружении знаний в данных» (Data Mining) являются классификация, регрессия и поиск ассоциативных правил. Data mining "углубляется" в данные, чтобы открыть полезные закономерности и нюансы, которые "погребены" в данных из-за огромного размера этих данных и сложности проблем. Результаты интеллектуального анализа данных представляют большую ценность для руководителей и аналитиков в их повседневной деятельности.

Ключевые слова: анализ данных, классификация, регрессия, поиск ассоциативных правил, классификационные правила

"Data mining обычно используется для обнаружения (скрытых) закономерностей в ваших данных для того, чтобы помочь вам принимать более лучшие деловые решения."

Роберт Смолл

Data mining помогает найти скрытые ранее закономерности и отношения в данных для того, чтобы можно было принять более обоснованные решения. Средства формирования запросов и отчетов помогают выбрать информацию из базы или хранилища данных. Data Mining анализирует прошлое для того, чтобы предсказать будущее. Data mining хорош при обнаружении тонких нюансов и совершении индивидуальных предсказаний.

Методы Data mining помогают решить многие задачи, с которыми сталкивается аналитик. Из них основными являются: классификация, регрессия, поиск ассоциативных правил и кластеризация:

– задача классификации сводится к определению класса объекта по его

характеристикам. В этой задаче множество классов, к которым может быть отнесен объект, заранее известно;

– задача регрессии позволяет определить по известным характеристикам объекта значение некоторого его параметра. Значением параметра является множество действительных чисел:

– при поиске ассоциативных правил целью является нахождение частных зависимостей между объектами и событиями. Найденные зависимости представляются в виде правил и могут быть использованы как для лучшего понимания природы анализируемых данных, так и для предсказания появления событий;

– задача кластеризации заключается в поиске независимых групп и их характеристик во всем множестве анализируемых данных. Решение этой задачи помогает лучше понять данные. Кроме того, группировка однородных объектов позволяет сократить их число, а, следовательно, и облегчить анализ.

При анализе часто требуется определить, к какому из известных классов относятся исследуемые объекты, то есть классифицировать их. Например, когда человек обращается в банк за предоставлением ему кредита, банковский служащий должен принять решение: кредитоспособен ли потенциальный клиент или нет. Очевидно, что такое решение принимается на основании данных об исследуемом объекте (в данном случае – человеке): его месте работы, размере заработной платы, возрасте, составе семьи и т.п. В результате анализа этой информации банковской служащий должен отнести человека к одному из двух известных классов «кредитоспособен» и «некредитоспособен».

В Data mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров. Определяемый параметр часто называют зависимой переменной, а параметры, участвующие в его определении – независимыми переменными. В рассмотренном примере независимыми переменными являлись: зарплата, возраст, количество детей и т.д. Зависимыми переменными – кредитоспособность клиента (возможные значения этой переменной «да» и «нет»).

Необходимо обратить внимание, что в примере независимая переменная принимает значение из конечного множества значений: {да, нет}. Если значениями независимых и зависимых переменных являются действительные числа, то задача называется задачей регрессии.

Задачи классификации и регрессии решаются в два этапа. На первом выделяется обучающая выборка. В нее входят объекты, для которых известны значения как независимых, так и зависимых переменных. В рассмотренном примере такой обучающей выборкой может быть информация о клиентах, которым ранее выдавались кредиты на разные суммы, и информация об их погашении. На основании обучающей выборки строится модель определения значения зависимой переменной. Ее часто называют функцией классификации или регрессии. Для получения максимально точной функции к обучающей выборке предъявляются следующие основные требования:

– количество объектов, входящих в выборку, должно быть достаточно большим. Чем больше объектов, тем построенная на ее основе функция классификации или регрессии будет точнее;

– в выборку должны входить объекты, представляющие все возможные классы в случае задачи классификации или всю область значений в случае задачи регрессии;

– для каждого класса в задаче классификации или каждого интервала области значений в задаче регрессии выборка должна содержать достаточное количество объектов.

На втором этапе построенную модель применяют к анализируемым объектам.

Поиск ассоциативных правил является одним из самых популярных приложений Data mining. Суть задачи заключается в определении часто встречающихся наборов объектов в большом множестве таких наборов. Данная задача является частным случаем задачи классификации. Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах. Анализу подвергались данные о совершаемых ими покупках, которые покупатели скла-

дывают в тележку. При анализе таких данных интерес прежде всего представляет информация о том, какие товары покупаются вместе, в какой последовательности, какие категории потребителей, какие товары предпочитают, в какие периоды времени и т.п. Такая информация позволяет более эффективно планировать закупку товаров, проведения рекламной компании и прочее. Задача поиска ассоциативных правил актуальна не только в сфере торговли. Например, в сфере обслуживания интерес представляет, какими услугами клиенты предпочитают пользоваться в совокупности. В медицине анализу могут подвергаться симптомы и болезни, наблюдаемые у пациентов.

Задача кластеризации состоит в разделении исследуемого множества объектов на группы «похожих» объектов, называемых кластерами. Кластеризация может применяться практически в любой области, где необходимо исследование экспериментальных или статистических данных. Кластеризация отличается от классификации тем, что для проведения анализа не требуется иметь выделенную зависимую переменную. Эта задача решается на начальных этапах исследования, когда о данных мало, что известно. Ее решение помогает лучше понять данные, и с этой точки зрения задача кластеризации является описательной задачей. Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы информации, делать их компактными и наглядными.

Задаче кластеризации присуще ряд особенностей:

– решение сильно зависит от природы объектов данных. Так, с одной стороны, это могут быть однозначно определенные, четко количественно очерченные объекты, а с другой – объекты, имеющие вероятностное или нечеткое описание;

– решение значительно зависит также и от представления кластеров и предполагаемых отношений объектов данных кластеров. Так, необходимо учитывать такие свойства, как возможность/невозможность принадлежности объектов нескольким кластерам. Необходимо определение самого понятия принадлежности кластеру: однозначная (принадлежит/не принадлежит), вероятностная

(вероятность принадлежности), нечеткая (степень принадлежности).

Цель технологии Data mining – нахождение в данных таких моделей, которые не могут быть найдены обычными методами. Существуют два вида моделей: предсказательные и описательные.

Предсказательные модели строятся на основании набора данных с известными результатами. Они используются для предсказания результатов на основании других наборов данных. При этом требуется, чтобы модель работала максимально точно, была статистически значима и оправдана.

К ним относятся следующие модели:

1) классификация – описывают правила набора правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов. Такие правила строятся на основании информации о существующих объектах путем разбиения их на классы;

2) модель последовательностей – описывают функции, позволяющие прогнозировать изменение непрерывных числовых параметров. Они строятся на основании данных об изменении некоторого параметра за прошедший период времени.

Описательные модели уделяют внимание сути зависимостей в наборе данных, взаимному влиянию различных факторов, то есть на построение эмпирических моделей различных систем. Ключевой момент в таких моделях – легкость и прозрачность для восприятия человеком. Возможно, обнаруженные закономерности будут специфической чертой именно конкретных исследуемых данных и больше нигде не встретятся, но это все равно может быть полезно и потому должно быть известно.

К ним относятся следующие виды моделей:

1) регрессионные – описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;

2) кластеризации – описывают группы, на которые можно разделить объекты, данные о которых подвергаются анализу. Объекты внутри кластера

должны быть «похожими» друг на друга и отличаться от объектов, вошедших в другие кластеры;

3) исключений – описывают исключительные ситуации в записях, которые резко отличаются чем-либо от основного множества записей. Знания исключений может быть использовано двояким образом;

4) итоговые – выявление ограничений на данные анализируемого массива, т.е. это нахождение каких-либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей такого же формата;

5) ассоциации – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из событий X следует событие Y.

Для построения рассмотренных моделей используются различные методы и алгоритмы Data mining:

1) базовые методы. К данным методам принято относить, прежде всего, алгоритмы, основанные на переборе. Простой перебор всех исследуемых объектов требует $0(2^N)$ операций, где N – количество объектов. Для сокращения вычислительной сложности в таких алгоритмах, как правило, используют разного рода эвристики, приводящие к сокращению перебора. Основным достоинством данных алгоритмов является их простота, как с точки зрения понимания, так и реализации;

2) нечеткая логика. Основным способом исследования задач анализа данных является их отображение на формализованный язык и последующий анализ полученной модели. Основной сферой применения нечеткой логики было и во многом остается управление;

3) генетические алгоритмы относятся к числу универсальных методов оптимизации, позволяющих решать задачи различных типов и различной степени сложности. При этом генетические алгоритмы характеризуются возможностью как однокритериального, так и многокритериального поиска в большом пространстве;

4) нейронные сети – это класс моделей, основанных на биологической аналогии с мозгом человека и предназначенных после прохождения этапа так называемого обучения на имеющихся данных для решения разнообразных задач анализа данных.

Одной из наиболее распространенных задач анализа данных является определение часто встречающихся наборов объектов в большом множестве наборов. Обозначим объекты, составляющие исследуемые наборы, следующим множеством:

$$I = \{i_1, i_2, ..., i_j, ..., i_n\},$$

где i_j – объекты, входящие в анализируемые наборы; n – общее количество объектов.

Наборы объектов из множества I, хранящиеся в базе данных и подвергаемые анализу, называются транзакциями. Опишем транзакцию как подмножество множества I:

$$T_1 = \{i_j / i_j \in I\}.$$

Набор транзакций, информация о которых доступна для анализа, обозначим следующим множеством:

$$D = \{T_1, T_2, ..., T_r, ..., T_m\},$$

где m – количество доступных для анализа транзакций.

Множество транзакций, в которые входит объект i_j, обозначим следующим образом:

$$D_{i_j} = \{T_r / i_j \in T_r; j = 1..n; r = 1..m\} \subseteq D.$$

Некоторый произвольный набор объектов обозначим следующим образом:

$$\mathrm{F} = \{i_j / i_j \in I; j = 1..n\}.$$

Множество транзакций, в которые входит набор F, обозначим следующим образом:

$$D_F = \{T_r / F \subseteq T_r; r = 1..m\} \subseteq D.$$

Отношение количества транзакций, в которое входит набор F, к общему количеству транзакций называется поддержкой набора F и обозначается Supp(F):

$$Supp(F) = \frac{|D_F|}{|D|}.$$

При поиске аналитик может указать минимальное значение поддержки интересующих его наборов $Supp_{\min}$. Набор называется частным, если значение его поддержки больше минимального значения поддержки, заданного пользователем:

$$\text{Supp(F)} > Supp_{\min}.$$

Таким образом, при поиске ассоциативных правил требуется найти множество во всех частных наборов:

$$\text{L}=\{\text{F}| \text{ Supp(F)} > Supp_{\min}\}.$$

Развитие методов интеллектуального анализа, реализованных в контуре общей информационной системы, способствуют усилению обоснованности принимаемых управленческих решений. Рассмотренные методы могут быть использованы аналитиками в маркетинге, бизнесе, торговле, страховании, финансах и других областях, где требуется обработка больших объемов данных смешанного типа. Результаты анализа представляются в удобном для пользователя виде. Механизм генерации правил с оценкой их применимости обеспечивает дополнительные возможности для качественного исследования.

Библиографические ссылки

1.Барсегян А. А. Методы и модели анализа данных: OLAP и Data mining. / А. А. Барсегян, М. С. Куприяенов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург, 2004.

2.Дюк В. А. Обработка данных на ПК в примерах / В. А. Дюк. – СПб. : Питер, 1997.

3.Фролов Ю.В. Интеллектуальные системы и управленческие решения / Ю. В. Фролов. – М.: МГПУ, 2000.