

УДК 004.89

DATAMINING И ОБЛАСТИ ЕГО ПРИМЕНЕНИЯ

Л. М. Савченко

Научный руководитель – С. С. Бежитский

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: qcva@yandex.ru

В настоящее время DataMining достаточно распространена в коммерческой деятельности, также она представляет большую ценность для руководителей и аналитиков, так как сфера ее влияния абсолютно ничем не ограничена, она может применятся везде, где есть какие-либо данные. В данной статье мы углубимся в суть DataMining и рассмотрим ее основные грани и понятия.

Ключевые слова: DataMining, анализ данных.

DATAMINING AND ITS APPLICATIONS

L. M. Savchenko

Scientific supervisor – S. S. Bezitskii

Reshetnev Siberian State Aerospace University
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
E-mail: qcva@yandex.ru

Currently DataMining is quite common in business, also it is of great value to managers and analysts, because the sphere of influence is not limited, it can presentsa wherever there is any data. In this article we will delve into the essence of DataMining and consider its main facets and concepts.

Keywords: DataMining, analysis of data.

До начала девяностых годов в сфере обработки знаний царствовала прикладная статистика. Она оценивала, проверяла гипотезы, но давала грубые и усреднённые результаты. Благодаря техническому прогрессу люди начали хранить огромные объёмы информации, которая была разнородной, и которую, естественно, надо было обрабатывать. Стало ясно, что необходима новая технология обработки данных, которая сможет обеспечить глубинный анализ и более точные результаты.

В 1989 году Григорий Пятецкий-Шапиро проводит семинар, с которого и начинается история DataMining. DataMining – это интеллектуальный анализ данных. Это технология, которая помогает выявлять скрытые связи в базах данных очень больших размеров.

Применение рассматриваемой технологии не имеет границ, единственное условие – наличие обилия данных. Так как DataMining развился на базе прикладной статистики, методов искусственного интеллекта, теории баз данных и многого другого, то отсюда проистекает изобилие методов и алгоритмов [1].

DataMining отказывается от концепции усреднения по выборке, чтобы не производить какие-либо действия над «фиктивными» величинами, вместо этого в основу закладывается концепция шаблонов, которые отражают фрагменты взаимоотношений данных. Шаблоны представляют собой некие закономерности, «подвыборки» данных. Благодаря такой концепции можно формулировать задачу как «Имеются ли характерные портреты клиентов, которые могут отказаться от услуг нашей компании?», вместо «Каковы средние счета нынешних клиентов в сравнении со счетами ушедших?». Невтривальность таких найденных шаблонов и будет отражать неожиданные, скрытые знания.

Целью DataMining является нахождение таких моделей, которые не могут быть найдены обычными методами. И существует два вида моделей: предсказательные и описательные.

Предсказательные модели: позиционируются на наборе данных с известными результатами. И используются для предсказания результатов на основании других наборов данных. Это модели

классификации (описывают правила, по которым можно отнести описание объекта к одному из классов) и модели последовательностей (они описывают функции, по которым можно прогнозировать изменение непрерывных числовых параметров).

Описательные модели: они уделяют особое внимание сути зависимостей в наборе данных, взаимному влиянию различных факторов, построению эмпирических моделей. Являются легкими для восприятия человеком.

Типы закономерностей, которые может найти технология DataMining, согласно В. А. Дюку:

- **Ассоциация** – применяется, когда несколько событий связаны между собой. Например, исследования показали, что 59 % купивших чипсы берут также и газированную воду, а если есть скидка на такой комплект, то газированную воду приобретают в 79 % случаев. Если менеджеры располагают подобными данными, то им достаточно легко оценить действенность предполагаемой скидки.

- **Классификация** – выявление черт, которые будут характеризовать группу, к которой принадлежит объект, на основе обучения на уже классифицированных объектах.

- **Кластеризация** – отличается от классификации тем, что группы заранее не известны и средства DataMining самостоятельно выявляют различные однородные группы данных.

- **Последовательность** – применяется при существовании цепочки событий, связанных во времени. Например, при приобретении квартиры в течение месяца приобретается кухонная плита в 49 % случаев, а в течение трех недель – холодильник в 73 %.

- **Прогнозирование** – создание или нахождение шаблонов, которые будут истинно показывать тенденция поведения необходимых показателей по временным рядам. При помощи них можно предсказать поведение системы в будущем [2].

Вернемся к тому, что сферы применения технологии интеллектуального анализа данных ограничивается исключительно наличием обилия информации. Поэтому руководители и аналитики, использующие данную технологию, могут получать ощутимые преимущества в конкурентной борьбе.

Банковское дело – анализ кредитных рисков, привлечение и удержание клиентов, управление ресурсами. В частности, выявление стереотипов мошенничества с кредитными картами; помощь в построении модели «ценности» клиентов и соответствующих типов обслуживания.

Телекоммуникации – выявление категорий клиентов со схожими стереотипами использования услуг, как следствие – разработка ценовой политики; проведение анализа отказов; предсказание пиковых нагрузок.

Прикладная химия – выяснение особенностей химического строения соединений, определяющих их основные свойства [3].

В силу того, что DataMining развился на стыке многих областей, то можно привести классы и методы, использующие эту технологию: нейронные сети, деревья решений, генетические алгоритмы, алгоритмы ограниченного перебора, ассоциативные правила, кластерный анализ и многое другое [4].

Более того, методы интеллектуального анализа данных позволяют эффективно решать задачи структурного инженерного проектирования инновационных технических систем в металлургической промышленности [5].

Существует мнение, что интеллектуальный анализ данных это лишь веяние моды, которое довольно скоро отойдет на задний план. Но таковым DataMining нельзя назвать, потому что его появление это следующий этап процесса, развивающегося с XX века. Сильный рост вычислительной мощности компьютеров вместе с появлением дешевых электронных методов сбора информации последовательно вывели на этот этап. И игнорировать его нельзя: применяемые методы – многочисленны, а преимущества – огромны [6].

Выражаю благодарность Бежитскому С. С. за участие в поиске и подготовке материалов на тему данной работы.

Библиографические ссылки

1. Елманова Н. И. Введение в DataMining [Электронный ресурс] // КомпьютерПресс. 2003. № 8. URL: <http://compress.ru> (дата обращения: 07.04.2015).
2. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP : учеб. пособие. 2-е изд. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. СПб.: 2007. 59 с.
3. Учебно-методический комплекс по дисциплине «Информационные технологии» института экономики и управления [Электронный ресурс]. URL: <http://www.sergeeva-i.narod.ru/inform/page11.htm> (дата обращения: 07.04.2015).

4. Чубукова И. А. Курс DataMining [Электронный ресурс]. URL: <http://www.intuit.ru/department/database/datamining/> (дата обращения: 07.04.2015).

5. О решении задачи оптимального выбора параметров питания линейной индукционной машины генетическим алгоритмом с локальным поиском / С. С. Бежитский, Е. А. Головенко, В. А. Горемыкин, М. В. Первухин // Вестник СибГАУ. 2010. Вып. 4. С. 23–27.

6. Шурига Л. DataMining: новое оружие для малого бизнеса [Электронный ресурс]. URL: <http://datareview.info/article/data-mining-novoe-oruzhie-dlya-malogo-biznesa> (дата обращения: 07.04.2015).

© Савченко Л. М., 2015