Обзор методов и технологий Data Mining

1 Определение Data Mining

Термин Data Mining (добыча данных) тесно связан с термином KDD (Knowledge Discovery in Databases – обнаружение знаний в базах данных). Часто в публикациях эти понятия трактуются как синонимы, в то время как основатель направления KDD Г. Пиатетский-Шапиро представляет Data Mining как основной элемент технологии KDD [1].

KDD – это нетривиальный процесс извлечения из данных адекватных, новых, потенциально полезных и понятных знаний. Под данными подразумевается набор фактов, каждый из которых характеризуется множеством признаков (например, записи в базе данных). Знания формируются на основе шаблонов. Шаблон – это выражение на каком-либо формальном языке, описывающее отношения внутри соответствующего ему подмножества данных с некоторой мерой определенности. При этом под определенностью понимается ряд величин – степень полноты данных, объем исследуемого набора данных и степень поддержки шаблона доступными знаниями из исследуемой предметной области. Шаблон, который соответствует заданным пользователем критериям определенности и интереса (где интерес определяется новизной, полезностью и нетривиальностью) считается знанием [1, 2]. Поскольку получаемые знания являются результом анализа имеющегося ограниченного набора данных, они имеют характер эвристик, которые можно использовать в процессе принятия решения [3]. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений анализируемых показателей [4].

KDD – это интерактивный процесс, предполагающий обязательное участие пользователя. До его начала пользователем должна быть сформулирована задача анализа данных и выбран метод Data Mining, наиболее подходящий для решения задачи. В процессе KDD пользователь в соответствии со спецификой задачи и выбранным методом определяет подлежащий анализу подмассив данных из всего массива имеющихся. Он же определяет критерии для находимых шаблонов данных и после проведения анализа интерпретирует полученные результаты для своей предметной области.

Таким образом, процесс KDD состоит из следующих основных действий [5]:

- очистка данных от ошибочной, ложной и несущественной информации, восстановление недостающих данных;
- объединение нескольких разнородных источников данных в один;
- отбор данных, относящихся к конкретной задаче анализа;
- преобразование данных в формат, подходящий для проведения анализа с помощью процедур Data Mining;
- выполнение процедуры Data Mining – применение к данным алгоритмов анализа и обнаружения шаблонов данных;
- обработка полученных шаблонов с целью выделения из них интересных и значимых;
- представление и визуализация полученных знаний.

Таким образом, Data Mining представляет собой набор методов для получения шаблонов из уже отобранных и приведенных в определенный формат данных.

В математической статистике термин Data Mining используется уже достаточно давно, как обозначение методов анализа данных без четко сформулированных гипотез (так называемый разведочный анализ данных) [6]. Традиционно статистический анализ данных разделяется на два основных типа: подтверждающий и разведочный анализ [7]. Первый заключается в подтверждении или опровержении имеющихся гипотез, а второй в их нахождении. При этом в первом случае инициатором исследования гипотез выступает пользователь, а действия систем автоматизированного анализа ограничиваются только их проверкой. Во втором случае система самостоятельно выдвигает гипотезы в виде шаблонов информации [1, 7]. Разведочный анализ данных применяется для нахождения связей между переменными в ситуациях, когда отсутствуют полностью или частично априорные представления о природе этих связей. Методами такого анализа являются: анализ распределений переменных; кросс-табуляция; кластерный анализ; факторный анализ; анализ дискриминантных функций; многомерное шкалирование; логлинейный анализ; канонические корреляции; пошаговая линейная и нелинейная регрессия; анализ соответствий; анализ временных рядов и деревья классификации [3]. Многие из этих методов, составляющих классический разведочный анализ, позволяют в формальном виде получить зависимости в исследуемых данных, но при этом используется сложный математический аппарат, предъявляющий высокие требования к квалификации пользователя, и возникают трудности в применении и интерпретации получаемых результатов.

В связи с развитием информационных технологий записи и хранения данных, в различных областях человеческой деятельности накоплены огромные массивы разнородной информации, которые без продуктивной переработки являются трудновоспринимаемыми человеческим разумом, и могут остаться невостребованными для практического использования. Специфика

современных требований к методам переработки информации должна учитывть следующие особенности [4]:

- данные имеют большой объем;
- данные являются разнородными (количественными, качественными, текстовыми);
- результаты обработки должны быть конкретны и понятны;
- инструменты для обработки “сырых” данных должны быть понятны и просты в использовании.

Появление новых требований обусловило необходимость развития в области классического многомерного разведочного анализа технологий обнаружения характерных комбинаций и логических закономерностей в данных. Новые направления включили в себя развитие методов выделения пяти стандартных типов закономерностей: ассоциация, последовательность, классификация, кластеризация, прогнозирование. В принципе в постановке этих задач нет ничего нового. Специалисты на протяжении нескольких последних десятков лет решали подобные задачи (“поиск эмпирических закономерностей”, “эвристический поиск в сложных средах”, “индуктивный вывод” и т. п.) [4]. Сейчас в развитии данных технологий основные усилия направлены на достижение их максимальной эффективности путем применения в алгоритмах комбинаторных подходов для уменьшения пространства поиска в анализируемых данных [6], а также вовлечение в них методов искусственного интеллекта и машинного обучения (например, нейронные сети и деревья решений).

В настоящее время термин Data Mining в одних работах [1, 4, 7] сохраняет свое первоначальное значение “разведочный анализ данных”, объединяя все имеющиеся методы обнаружения зависимостей в больших массивах данных, как классические, так и новые. В других работах [3, 5, 6] в технологии Data Mining авторы выделяют главным образом направления, ориентированные в большей степени на практическое приложение полученных результатов, чем на выяснение природы явления. В области добычи данных принят такой подход к анализу данных и извлечению знаний, который иногда характеризуют словами “черный ящик”. При этом используются не только классические приемы разведочного анализа данных, но и такие методы, как нейронные сети, которые позволяют строить достоверные прогнозы, не уточняя конкретный вид тех зависимостей, на которых такой прогноз основан [3].

Таким образом, на основании проведенного анализа работ в области Data Mining можно сделать следующие основные выводы:

- Data Mining является дальнейшим развитием методов разведочного анализа данных в математической статистике, ориентированным на практическое применение;
- Data Mining объединяет в себе методы статистического анализа данных, комбинаторные методы, методы искусственного интеллекта и баз данных.

2 Классификация методов Data Mining

К Data Mining относят методы, которые позволяют обнаружить некоторые шаблоны в множестве данных. Одной из важных характеристик каждого метода является тип получаемых шаблонов. С этой точки зрения методы Data Mining разделяются на три категории: логические, вывод уравнений и кросс-табуляция [7].

2.1 Поиск логических закономерностей

Методы поиска логических закономерностей в данных учитывают информацию, заключенную не только в отдельных признаках, но и в сочетаниях значений признаков [4]. Набор всех возможных значений признаков рассматривается как множество элементарных событий. Данные рассматриваются как цепочки конъюнкций элементарных событий. Извлекаемые шаблоны имеют вид логических закономерностей: условий (ЕСЛИ/ТО), ассоциаций (КОГДА/ТО ТАКЖЕ), тенденций, отклонений, периодов, регулярных эпизодов и т.д. Логические закономерности представляются либо как наборы отдельных правил, либо в связанном в деревья решений виде [7].

За время развития теории анализа многомерных данных было предложено много различных методов поиска логических закономерностей. Наиболее популярными из них стали деревья решений и переборные алгоритмы, так называемые “алгоритмы здравого смысла” [4], не использующие сложной математической базы.

В основе методов построения деревьев решений лежит принцип циклического разбиения обучающей выборки на классы в соответствии со значениями одного из признаков. Каждый класс, выделяемый таким признаком, вновь разбивается на подклассы с использованием следующего признака. В ходе процесса образуется дерево решений. Пути движения по этому дереву с верхнего уровня на нижние определяют логические закономерности в виде цепочек конъюнкций.

Одним из первых переборных алгоритмов был алгоритм ограниченного перебора “Кора”, предложенный М.М. Бонгардом в 1967 году. Он основан на анализе частот возможных сочетаний элементарных событий в выборке. Рассматриваемые сочетания искусственно ограничиваются по числу составляющих их элементов. Алгоритм является трудоемким, так как основан на полном переборе. Поэтому он хорошо работает только при сравнительно небольших размерностях пространства признаков и невысоких значениях ограничения на число элементов в сочетаниях [4].

В настоящее время получили широкое распространение и активно развиваются и совершенствуются модификации двух алгоритмов, используемых в данной работе. Первый из них – Apriory, представленный в [8], использует ограниченный перебор сочетаний значений признаков, встречающихся в исследуемой выборке данных. Второй – Prefix-Span, представленный в [9], основан на построении дерева решений.

2.2 Вывод уравнений

Если методы вывода логических закономерностей не предполагали каких-либо ограничений на структуру анализируемых данных, то методы данной категории требуют чтобы все подлежащие анализу данные были числовыми. Эти методы разработаны для решения традиционных задач разведочного анализа данных и используют как математический аппарат прикладной статистики, так и методы искусственного интеллекта (например, нейронные сети). Обнаруживаемые с помощью этих методов закономерности выражаются формально в виде уравнений.

2.3 Кросс-табуляция

Кросс-табуляция – это процесс объединения двух (или нескольких) таблиц частот так, что каждая ячейка (клетка) в построенной таблице представляется единственной комбинацией значений или уровней табулированных переменных. Таким образом, кросс-табуляция позволяет совместить частоты появления наблюдений на разных уровнях рассматриваемых факторов. Исследуя эти частоты, можно определить связи между табулированными переменными. Обычно табулируются категориальные (номинальные) переменные или переменные с относительно небольшим числом значений [3].

Список литературы

[1] Usama Fayhad, Gregory Piatetsky-Shapiro, Padhraic Smyth, From Data Mining to Knowledge Discowery in Databases, http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, 1996.

[2] William J. Frawely, Gregory Piatetsky-Shapiro, Christopher J.Matheus, Knowledge Discovery in Databases: An Overview, http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1992.pdf, 1992.

[3] Электронный учебник по статистике, Copyright (c) StatSoft, Inc. 1984-2001, http://www.statsoft.ru/home/textbook/default.htm.

[4] Дюк В., Самойленко А., Data Mining: учебный курс. СПб: Питер, 2001.

[5] Jiawei Han, “Data Mining”, in J.Urban and Dasqupta (eds.), Encyclopedia of Distributed Computing, Kluwer Academic Publisher, 1999.

[6] Heikki Mannila, Data mining: machine learning, statistics, and databases, SSDBM 1996: 2-9, http://www.cs.helsinki.fi/ mannila/postscripts/ssdbm.ps.

[7] Parsaye K., A Characterization of Data Mining Technologies and Processes // The Journal of Data Warehousing, 1998, № 1.

[8] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, Mining Association Rules between Sets of Items in Large Databases // Proc. of the 1993 ACM SIGMOD Conf. Washington DC, USA, (May 1993) pp.207-216.

[9] Jian Pei, Juawei Han and others, PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, In Proc. 2001 Int. Conf. Data Engineering (ICDE'01), Heidelberg, Germany, April 2001, pp.215-224.