Наиболее Распространенные Задачи Data Mining

Сайяр Абдуллаев[1], Судаба Абасова[2], Сабина Фоменко[3]
[1,2,3]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1,2,3]depart5@iit.ab.az

Аннотация **– в статье рассматривается понятие Data Mining. Описываются возникновение Data Mining, основная суть задач, а в некоторых из них рассмотрены процесс и методы решения, а также применение. Сравниваются Big Data и Data Mining.**

Ключевые слова **–** ***Big Data, Data Mining, классификация, кластеризация, прогнозирование, визуализация.***

I. ВВЕДЕНИЕ

Для получения конструктивной и нужной информации прежние методы, которые применяли статики и математики, требовали большого количества времени. Это привело к формированию Data Mining.

Data Mining как термин возник в 1978 году. До середины 1990-х годов данные обрабатывались и анализировались в рамках прикладной статистики. При этом чаще всего решались задачи обработки маленьких баз данных. В этот промежуток времени понятие Data Mining начало приобретать популярность в современной интерпретации. Одним из основателей направления Data Mining является Григорий Пиатецкий-Шапиро [1].

Понятие Data Mining означает добыча (mining) данных (data). Синонимами Data Mining можно считать понятия "обнаружение знаний в базах данных" (Knowledge Discovery in Databases, KDD) и "интеллектуальный анализ данных".

Data Mining применяется везде, где имеются какие-либо данные. Но на сегодняшний день методы Data Mining в первую очередь заинтриговали коммерческие предприятия, которые развертывают проекты на основе информационных хранилищ данных.

Главной характерной чертой Data Mining является сочетание большого математического инструментария (от традиционного статистического исследования до самых новых кибернетических способов) с новейшими достижениями в области информационных технологий.

Data Mining – это мультидисциплинарная сфера, возникшая и формируемая на основе таких наук, как распознавание образов, теория баз данных, прикладная статистика, искусственный интеллект, и др. (рис. 1).

Каждый год Data Mining посвящается множество научных и практических конференций. Одной из них является Международная конференция по Knowledge Discovery Data Mining (International Conferences on Knowledge Discovery and Data Mining).

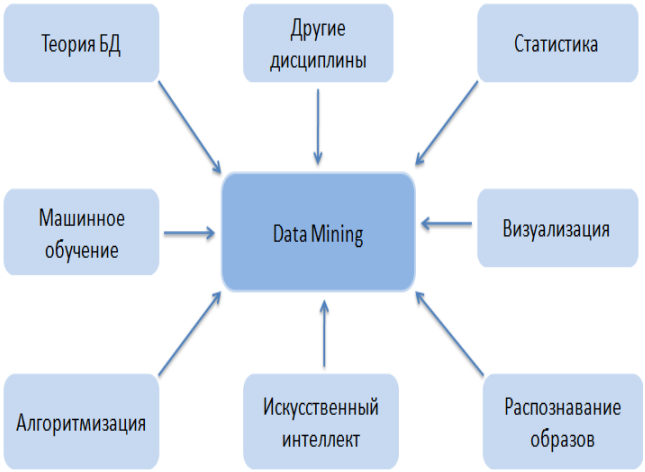

Рис. 1. Data Mining в роли мультидисциплинарной области

II. ЗАДАЧИ DATA MINING

Концепция шаблонов составляет базу технологии Data Mining. Концепция отражает в данных неожиданные, неочевидные регулярности. Эти регулярности составляют скрытые знания. Концепция представляет собой закономерности. Им свойственны выборки данных, выраженные в понятной для человека форме. Задачи Data Mining решаются результатом обнаружения незаметных закономерностей. Иногда задачи Data Mining называют закономерностями или техниками [2].

Нет единого мнения о том, какие задачи можно причислять к Data Mining. Во многих изданиях перечислены нижеприведенные задачи Data Mining:

1) классификация;
2) кластеризация;
3) прогнозирование;
4) ассоциация;
5) визуализация;
6) анализ и обнаружение отклонений;
7) оценивание;
8) анализ связей;
9) подведение итогов.

Задача классификации. Самой часто решаемой, а также самой простой задачей Data Mining является классификация.

Классификация – это упорядоченная в соответствии с определенным принципом масса объектов, имеющих схожие классификационные признаки, с помощью которых между этими объектами определяются сходства либо различия.

В задаче классификации должны соблюдаться следующие правила:

- в каждой части деления должно применяться только одно основание;
- общий объем категориальных понятий должен равняться объему родового понятия, которое делится. Деление обязательно соразмерно;
- объемы членов деления не могут пересекаться. Нужно, чтобы они взаимно исключали друг друга;
- деление обязательно должно быть последовательно.

Можно различить искусственную и естественную классификации. Искусственной называется классификация, осуществляемая по внешнему признаку. Вспомогательная классификация нужна для придания большинству процессов и явлений заданного порядка. Естественной называется классификация, которая осуществляется по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Естественная классификация есть результат и важное средство научного исследования.

Задачу классификации можно решить, используя следующие способы:

- ближайшего соседа;
- индукцию деревьев решений;
- байесовские сети;
- нейронные сети.

Задача кластеризации. В кластеризации классы исследуемого набора данных предварительно неопределенны.

Кластеризация используется с целью разбиения совокупности объектов в схожие кластеры либо классы. Представив данные выборки в виде точек в признаковом пространстве, задачу кластеризации можно свести к понятию "сгущений точек".

С помощью кластеризации можно провести исследовательский анализ с целью изучения "структуры данных".

Кластеризация применяется во многих отраслях. Первоначально она использовалась в биологии, антропологии и психологии.

Кластеры бывают непересекающимися и пересекающимися.

Иногда применение всевозможных способов кластеризации приводит к различным результатам. Это зависит от особенностей работы какого-либо алгоритма. Существует более сотни разнообразных алгоритмов кластеризации.

Примером способа решения задачи кластеризации является обучение "без учителя".

Задача прогнозирования. В самых различных сферах человеческой деятельности можно столкнуться с задачей прогнозирования. Развитие информационных технологий и методов прогнозирования напрямую связано друг с другом. Так как оно требует детального исследования исходного набора подходящих для анализа данных и методов, его можно считать одной из самых сложных задач Data Mining. Цель прогнозирования – это предсказание будущих событий.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе анализа его состояния в прошлом и настоящем. Следовательно, решение задачи прогнозирования требует некоторой обучающей выборки данных.

При решении задачи прогнозирования устанавливается функциональная зависимость между независимыми и зависимыми переменными. Прогнозирование является распространенной и востребованной задачей во многих областях человеческой деятельности. В результате прогнозирования уменьшается риск принятия неверных, необоснованных или субъективных решений.

Задача ассоциации. При решении задачи ассоциации отыскиваются закономерности среди связанных событий в наборе данных.

Ассоциации отличаются от классификации и кластеризации тем, что в них поиск закономерностей происходит между несколькими событиями, протекающими одновременно.

Задача поиска ассоциативных правил решается с помощью многих алгоритмов.

Задача визуализации. Визуализация позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом. Результатом использования визуализации является графический образ данных. Визуализация представляется в виде гистограмм, графиков, диаграмм, схем и т.д. Визуализацию используют в задачах классификации и кластеризации.

В визуализации практически полностью отсутствует необходимость в специальной подготовке пользователей.

Задачами визуализации как компонента Data Mining, являются: поддержка интерактивного и согласованного исследования; помощь в предоставлении результатов; использование зрения для создания зрительных образов и осмысление их.

Анализ связей. Это задача, с помощью которой находятся зависимости в наборе данных.

Подведение итогов. Целью этой задачи является описание конкретных групп объектов, выбранных из анализируемого набора данных [3].

III. DATA MINING И BIG DATA

Big Data и Data Mining – это два разных понятия. Каждое из понятий применяется в работе по обработке больших объемов данных. Кроме того, эти два термина используются для двух различных направлений такого рода работы.

Big Data – это термин, означающий набор больших данных. Big Data являются те данные, которые уже перерастают простые базы данных, используемые в ранние периоды, стоили они дороже и имели меньше возможностей. Примером этого могут служить наборы данных, которые слишком большие для обработки в разных системах.

Понятие Big Data в последнее время, можно сказать, стало мейнстримом. С Big Data тесно связаны такие термины, как анализ данных, наука о данных, аналитика данных, сбор данных и машинное обучение.

В отличие от Big Data, Data Mining используется для поиска нужной информации среди большого набора данных. Синонимом этого вида деятельности можно считать высказывание "искать иголку в стоге сена". То есть автоматически собираются большие наборы данных, которые могут быть однородными, похожими. Data Mining может включать в себя использование различных программ [5].

Data Mining – это процесс очистки больших данных и подготовки их последующему анализу или использованию в алгоритмах машинного обучения. Data Mining должны обладать исключительными распознавательными качествами, чудесной интуицией и техническими умениями для объединения и трансформирования огромного количества данных [6].

ЗАКЛЮЧЕНИЕ

Таким образом, технология Data Mining с каждым днем становится все более востребованной. С помощью Data Mining можно решить множество задач и обработать большое количество данных.

На сегодняшний день существует несколько точек зрения на Data Mining. Некоторые считают, что это мираж, который отвлекает внимание от классического анализа данных. Другие же считают, что Data Mining – это альтернатива традиционному подходу к анализу. А также, как и во всех случаях, существует середина, которая рассматривает возможность совместного использования современных достижений в области Data Mining и при классическом статистическом анализе данных.

Для того чтобы пользоваться Data Mining, нужна специальная квалификация. Не имея ее, очень трудно работать с Data Mining. Квалификация должна постоянно повышаться.

ЛИТЕРАТУРА

[1] И.А.Чубукова, Data Minig. Курс лекций интернет-университета INTUIT, 2008, с. 7–16.
[2] В.Дюк, А.Самойленко. Data Mining. Издательство "Питер", 2001, с. 14–20.
[3] http://bug.kpi.ua/stud/work/RGR/DATAMINING/tasksofdm.html
[4] www.kdnuggets.com
[5] https://www.techopedia.com/7/29678/technology-trends/what-is-the-difference-between-big-data-and-data-mining
[6] http://habrahabr.ru/company/io/blog/264125/