ЗАДАЧИ, МОДЕЛИ И МЕТОДЫ DATA MINING

Н. А. Никульская
Сибирский Федеральный Университет, Институт математики
г. Красноярск
Nadin-Nikulsk@yandex.ru

Аннотация. *В данной работе осуществляется обзор основных задач, моделей и методов Data Mining.*
Ключевые слова. Data Mining, методы, модели, задачи.

Для начала выясним, что такое Data Mining. Data Mining (discovery-driven data mining) дословно переводится как "добыча" или "раскопка данных". В основу современной технологии Data Mining положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные выборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск закономерностей производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Важное положение Data Mining среди других информационных технологий – нетривиальность разыскиваемых закономерностей. Это означает, что найденные закономерности должны отражать неочевидные, неожиданные регулярности в данных составляющие, так называемые скрытые знания.

Рассмотрим основные методы Data Mining и их приложения в конкретных системах.

1. Статистические методы.
   К базовым методам Data Mining традиционно причисляют все подходы, использующие элементы теории статистики. Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (компания SPSS), STATGRAPICS (компания Manugistics), STATISTICA, STADIA и другие.
2. Полный и ограниченный перебор.
   К базовым методам Data Mining принято относить также алгоритмы, основанные на переборе. Простой перебор всех исследуемых объектов требует $O(2^N)$ операций, где N – количество объектов. Следовательно, с увеличением количества данных объем вычислений растет экспоненциально, что при большом объеме делает решение любой задачи таким методом практически невозможным. Наиболее ярким современным представителем реализации этого подхода является система WizWhy компании WizSoft.
3. Нечеткая логика.
   Основным способом исследования задач анализа данных является их отображение на формализованный язык и последующий анализ полученной модели. Неопределенность по объему отсутствующей информации у системного аналитика можно разделить на три большие группы:

   1) неизвестность,
   2) неполнота (недостаточность, неадекватность),
   3) недостоверность.

   Недостоверность бывает физической (источником ее является внешняя среда) и лингвистической (возникает в результате словесного обобщения и обусловливается необходимостью описания бесконечного числа ситуаций ограниченным числом слов за ограниченное время). Основной сферой применения нечеткой логики и во многом остается управление.
4. Генетические алгоритмы.
   Генетические алгоритмы относятся к числу универсальных методов оптимизации, позволяющих решать задачи различных типов (комбинаторные, общие задачи с ограничениями и без ограничений) и различной степени сложности. Одним из наиболее востребованных приложений генетического алгоритма в области Data Mining является поиск наиболее оптимальной модели (поиск алгоритма, соответствующего специфике конкретной области). Эти алгоритмы удобны тем,

что их легко распараллеливать. Пример системы, построенной на основе технологии Data Mining с применением генетических алгоритмов: система GeneHunter фирмы Ward Systems Group.

5. Нейронные сети.
   Нейронные сети – это класс моделей, основанных на биологической аналогии с мозгом человека и предназначенных для решения разнообразных задач анализа данных после прохождения этапа, так называемого обучения на имеющихся данных. При применении этого метода, прежде всего, встает вопрос выбора конкретной архитектуры сети (числа "слоев" и количества "нейронов" в каждом из них). Размер и структура сети должны соответствовать существу исследуемого явления. Примеры нейросетевых систем: BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic).
6. Деревья решений.
   Деревья решения являются одним из наиболее популярных подходов к решению задачи классификации. Они создают иерархическую структуру классифицирующих правил типа "если-то" (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня.
   Большинство систем, построенных на основе технологии Data Mining, используют именно этот метод. Самыми известными являются See5/C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада).

Методы Data Mining помогают решить многие задачи, с которыми сталкивается аналитик. Из них основными являются: задача классификации и регрессии, задача поиска ассоциативных правил и задача кластеризации. Далее приведено краткое описание основных задач анализа данных.

Задача классификации и регрессии. Задача классификации сводится к определению класса объекта по его характеристикам. Необходимо заметить, что в этой задаче множество классов, к которым может быть отнесен объект, известно заранее. Задача регрессии подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого его параметра. В отличие от задач классификации значением параметра является не конечное множество классов, а множество действительных чисел.

Задача поиска ассоциативных правил. При поиске ассоциативных правил целью является нахождение частных зависимостей (или ассоциаций) между объектами или 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 (кластеров) 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.

Перечисленные задачи по назначению делятся на описательные и предсказательные.

Сфера применения Data Mining ничем не ограничена − она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня заинтересовали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных. Кратко приведем некоторые приложения Data Mining:

1. Электронная коммерция.
2. Розничная торговля.
3. Промышленное производство.
4. Медицина.
5. Банковское дело.
6. Страховой бизнес.
7. Телекоммуникации.
8. Молекулярная генетика и генная инженерия.
9. Прикладная химия.

Литература

[1]

[2] Дюк В.А. Data Mining — интеллектуальный анализ данных. — http://www.olap.ru/basic/dm2.asp.

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