

## МЕТОДЫ DATA MINING В ЗАДАЧАХ ПРИНЯТИЯ РЕШЕНИЙ

Былина Т.А.

Марухина О.В. (научный руководитель)

Национальный исследовательский Томский политехнический университет  
bylina\_1995@mail.ru

### Введение

В век стремительного развития информационных технологий значительно увеличивается необходимость качественной обработки огромных массивов данных, собираемых в организациях. Объемы этих массивов порой достигают таких масштабов, что их обработка становится непосильной даже самым опытным экспертам. В таких случаях на помощь приходят методы интеллектуальной обработки данных Data Mining.

Data Mining – процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [1]. Цель данного исследования – обзор наиболее распространенных методов Data Mining:

### Обзор методов

Технология Data Mining представляет собой совокупность различных методов, позволяющих осуществлять самостоятельный поиск нетривиальных зависимостей и закономерностей между данными и формировать предположения, которые помогают лицу, принимающему решение, в изучении поставленной задачи [2,3,4].

- Корреляционно-регрессионный анализ. Призван производить поиск связей между двумя случайными величинами. В процессе анализа может быть выявлено наличие прямой или обратной связи или её отсутствие.

- Дерево решений. Может также именоваться деревом принятия решений, регрессионным деревом или деревом классификации. Представляет собой иерархическую структуру, построение которой осуществляется по набору определенных правил, представленных в виде конструкций «Если ..., то ...». Промежуточные узлы и ребра отражают правила, а конечные интерпретируют «корзины», в которые помещаются классифицируемые данные.

- Иерархическая кластеризация. Суть метода заключается в пошаговом объединении малых кластеров в более крупные или же, наоборот, в разделении больших кластеров на более мелкие в зависимости от условий поставленной задачи.

- Неиерархическая кластеризация. Метод имеет итеративную природу. Разбиение на кластеры происходит до тех пор, пока не будет выполнено правило остановки.

- Искусственная нейронная сеть. Представляет собой модель организации данных и процессов, интерпретирующую работу нервных

клеток в организме. Отдельные узлы сети достаточно просты, но, находясь в определенной, четко заданной взаимосвязи, способны решать достаточно сложные задачи.

- Эволюционное программирование. Основано на генетических алгоритмах, которые представляют собой эвристические алгоритмы поиска, производящие подбор и сочетание необходимых данных, применяя механизмы по аналогии с естественным отбором.

- Метод опорных векторов. Задачей данного метода является переход из области начальных векторов в новое пространство, которое имеет большую размерность, чем исходное, и поиск в этом пространстве разделяющей гиперплоскости, имеющей большой зазор.

- Байесовская сеть. Представляет собой вероятностную модель, представленную в виде графа, в котором вершины содержат случайные переменные, а ребра соответствуют вероятностным взаимосвязям между ними по Байесу.

- Методы ближайшего соседа и  $k$ -ближайшего соседа. Основаны на оценке сходства рассматриваемых объектов. Первый метод полагается на единственный ближайший сходный объект обучающей выборки, второй же менее «доверчив» и требует поиска сходств с  $k$ -ближайшими похожими объектами.

- Линейная регрессия. Отображается в виде регрессионной модели, которая описывает линейную взаимосвязь зависимой переменной от одной или нескольких независимых переменных.

- Метод поиска ассоциативных правил. Позволяет находить ассоциативные правила, на основании которых будет приниматься решение. Задача поиска правил включает в себя нахождение часто встречающихся наборов элементов и генерацию правил из ранее найденных наборов с определенной пороговой достоверностью.

- Метод ограниченного перебора. Используется для нахождения логических взаимосвязей и закономерностей в массивах данных. Алгоритмы данного метода находят частоты комбинаций простых логических событий в подгруппах данных, на основании анализа которых исследуется полезность комбинации для поставленной задачи.

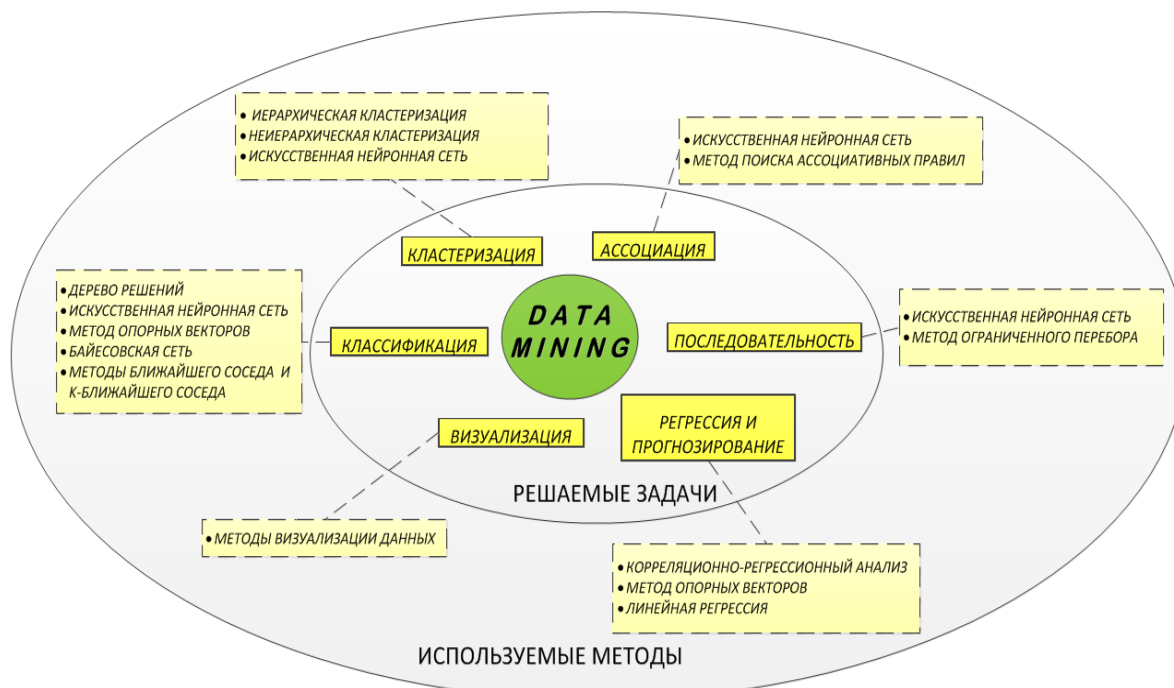


Рис. 1. Классификация методов Data Mining по типам решаемых задач

- Методы визуализации данных. Сюда относятся все методы, представляющие данные в легко воспринимаемом человеком виде.

#### Основные задачи, решаемые методами Data Mining

Все вышеописанные методы предназначены для решения определенных задач. Все задачи условно могут быть разделены на 6 больших классов (рис. 1):

1. Классификация (стратификация) – нахождение у рассматриваемых объектов специфических признаков, которые определяют их отношение к одному из заранее заданных классов.

2. Кластеризация – это несколько более трудная задача, решаемая методами Data Mining. Для этой задачи классы заранее неизвестны, их необходимо сформировать. В остальном идеи классификации сохраняются.

3. Ассоциация – выявление закономерностей среди взаимосвязанных событий. Основывается на рассмотрении одновременно произошедших событий и выявляется зависимость между произошедшими явлениями.

4. Последовательность (поиск последовательных шаблонов) – нахождение закономерностей среди взаимосвязанных по времени событий.

5. Регрессия и прогнозирование – поиск зависимости выходных данных от входных переменных и предсказание новых результатов на основе выявленных зависимостей.

6. Визуализация – графическое отображение анализируемых данных. Большие объемы сырых данных отображаются в виде наглядных таблиц, схем, диаграмм, графов и т.д.

#### Заключение

Рассмотренные методы интеллектуального анализа данных можно классифицировать в соответствии с классами задач, которые они решают.

Наибольшую сферу задач охватывает метод построения искусственных нейронных сетей. Он направлен на решение задач классификации, кластеризации, поиска ассоциаций и последовательностей. Если же рассматривать наборы методов для решения одного класса задач, то для решения задач классификации может быть использовано наибольшее число методов, среди которых методы построения деревьев, байесовских и искусственных нейронных сетей, методы опорных векторов, ближайшего соседа и  $k$ -ближайшего соседа.

#### Список использованных источников

1. Дюк В., Самойленко А. Data Mining: учебный курс. – СПб: Питер, 2001.
2. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. — М.: Изд. «Фазис», 2006. — 176 с.
3. Степанов Р.Г. Технология Data Mining: Интеллектуальный Анализ Данных. – Казанский Государственный Университет им. В.И.Ульянова-Ленина, 2008.
4. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.