DOI 10.58351/2949-2041.2023.4.4.025

Димитриади Георгий Дмитриевич,
магистрант 3 курса КубГУ, г. Краснодар

ОСОБЕННОСТИ МЕТОДИКИ ИЗВЛЕЧЕНИЯ ЗНАНИЯ ИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДИКИ KDD. ДЕТАЛИЗАЦИЯ ЭТАПОВ

Аннотация. В статье приведена детализация процесса Knowledge Discovery in Databases (KDD), а также описаны задачи, которые выполняются благодаря данной методике.

Ключевые слова. KDD, выборка данных, план-факторный анализ, управление рисками.

Методика KDD, зародившаяся в 1989 г., описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для построения модели (извлечения знания). Методика не зависит от предметной области: это набор атомарных операций, комбинируя их, можно получить нужное решение. Методика KDD включает этапы подготовки данных, выбора информативных признаков, очистки, построения моделей, постобработки и интерпретации полученных результатов. Ядром этого процесса являются методы Data Mining, позволяющие обнаруживать закономерности и знания.

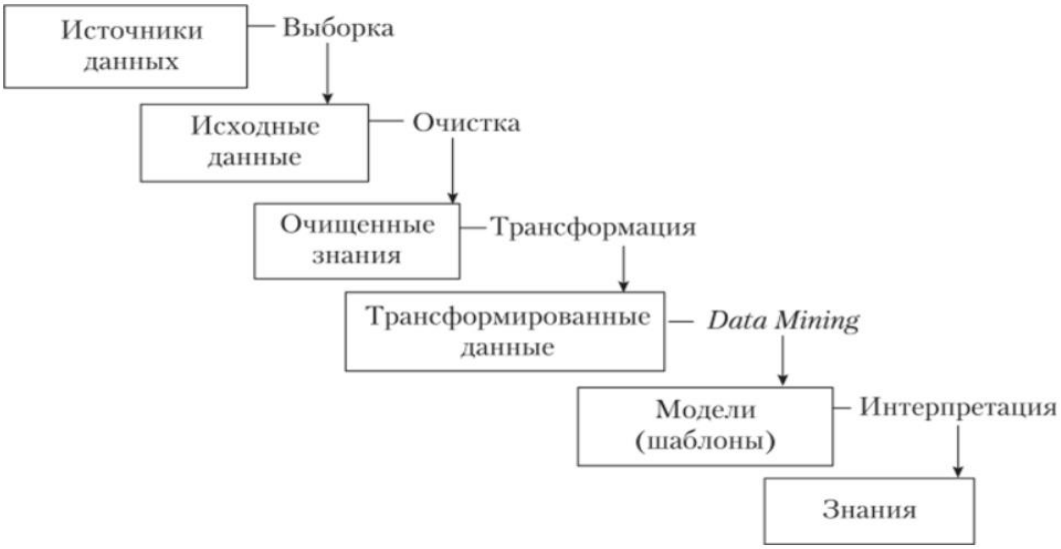

Рисунок 1 – визуализация этапов KDD
Источник: разработано автором

Таким образом, Knowledge Discovery in Databases (KDD) — процесс получения из данных знаний в виде зависимостей, правил, моделей, обычно состоящий из таких этапов, как отбор, очистка, трансформация, моделирование и интерпретация полученных результатов. Рассмотрим последовательность шагов, выполняемых на каждом этапе KDD.

Выборка данных. Первым шагом в анализе является получение исходной выборки. На основе этих данных и строятся модели. Здесь требуется активное участие экспертов для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Чаще всего в качестве источника рекомендуется использовать специализированное хранилище данных, агрегирующее всю необходимую для анализа информацию.

Очистка данных. Реальные данные для анализа редко бывают хорошего качества. Необходимость в предварительной обработке при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять

самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. К задачам очистки данных относятся: заполнение пропусков, подавление аномальных значений, сглаживание, исключение дубликатов и противоречий и пр.

Трансформация данных. Этот шаг необходим для тех методов, при использовании которых исходные данные должны быть представлены в каком-то определенном виде. Дело в том, что различные алгоритмы анализа требуют специальным образом подготовленных данных. Например, для прогнозирования необходимо преобразовать временной ряд при помощи скользящего окна или вычислить агрегированные показатели. К задачам трансформации данных относятся: скользящее окно, приведение типов, выделение временных интервалов, квантование, сортировка, группировка и пр.

Интерпретация. В случае, когда извлеченные знания не прозрачны для пользователя, должны существовать методы постобработки, позволяющие привести эти знания к интерпретируемому виду. Для оценки качества полученной модели нужно использовать как формальные методы, так и знания аналитика. Именно аналитик может сказать, насколько применима полученная модель к реальным данным. Полученные модели являются, по сути, формализованными знаниями эксперта, и, следовательно, их можно тиражировать. Найденные знания должны быть применимы и к новым данным с некоторой степенью достоверности.

Knowledge Discovery in Databases не задает набор методов обработки или пригодные для анализа алгоритмы, он определяет последовательность действий, которую необходимо выполнить для того, чтобы из исходных данных получить знания. Данный подход универсальный и не зависит от предметной области, что является его несомненным достоинством. Deductor - полнофункциональная платформа для решения задач Knowledge Discovery in Databases, позволяющая провести все вышеописанные шаги.

Несмотря на большое количество разнообразных бизнес-задач, почти все они решаются по единой методике KDD. Она описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для построения модели (извлечения знания). Данная методика не зависит от предметной области, это набор атомарных операций, комбинируя которые, можно получить нужное решение.

В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду. Для оценки качества полученной модели нужно использовать как формальные методы оценки, так и знания эксперта. Так как именно эксперт может сказать, насколько применима полученная модель к реальным данным. Полученные модели являются, по сути, формализованными знаниями эксперта, а следовательно их можно тиражировать. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Использование методов построения моделей позволяет получать новые знания, которые невозможно извлечь другим способом. Кроме того, полученные результаты являются формализованным описанием некоего процесса, а следовательно поддаются автоматической обработке. Недостатком же является то, что такие методы более требовательны к качеству данных, знаниям эксперта и формализации самого изучаемого процесса. К тому же почти всегда имеются случаи не укладывающиеся ни в какие модели. На практике подходы комбинируются, например, визуализация данных наводит эксперта на некоторые идеи, которые он пробует проверить при помощи различных способов построения моделей, а результаты построения моделей подаются на вход механизмам визуализации. Полнофункциональная система анализа не должна замыкаться на применении только одного подхода или одной методики анализа. Механизмы визуализации и построения моделей должны дополнять друг друга. Максимальную отдачу можно получить комбинируя методы и подходы к анализу данных.

С помощью KDD решаются бизнес-задачи, например:

План-факторный анализ - визуализация данных;

Анализ денежных потоков - визуализация данных;
Прогнозирование - задача регрессии;
Управление рисками - регрессия, кластеризация и классификация;
Стимулирование спроса - кластеризация, ассоциация;
Оценка эластичности спроса - регрессия;
Выявление предпочтений клиентов - последовательность, кластеризация, классификация.

Список литературы:

1. Арсеньев С.Б., Бритков В.Б., Маленкова Н.А. Использование технологии анализа данных в интеллектуальных информационных системах // Управление информационными потоками. – М.: УРСС, ИСА РАН, 2002.
2. Обнаружение знаний в хранилищах данных. [Электронный ресурс]- режим доступа https://www.osp.ru/os/1999/05-06/179852