

А.Н. Мишкевич, 2 курс

Научный руководитель – О.В. Сидская

Полесский государственный университет

Банки ежедневно совершают огромное количество операций, часть из которых ориентирована не только на проведение операций внутри одной страны, но также на международную активность. Для банка важно иметь актуальную, постоянно обновляющуюся информацию по его клиентам, чтобы проводить качественный и количественный анализ деятельности и интерпретацию результатов для дальнейшего реагирования.

С каждым годом вопрос применения технологии Data Mining (интеллектуального анализа данных) в банковской сфере становится все более актуальным. Это связано с накоплением банками больших объемов информации, ужесточением конкурентной борьбы, увеличением количества случаев мошенничества и невозврата кредита.

Основные задачи банковской сферы, которые успешно решаются с использованием инструментов Data Mining:

- Анализ кредитного риска
- Привлечение и удержание клиентов
- Прогнозирование изменений клиентуры
- Обнаружение совокупностей приобретаемых клиентами банковских продуктов и услуг
- Прогнозирование остатка на счетах клиентов
- Управление портфелем ценных бумаг
- Выявление случаев мошенничества с кредитными карточками
- Оценка прибыльности инвестиционных проектов
- Оценка интенсивности конкуренции и ближайших конкурентов
- Профилирование наилучших достижений
- Повышение качества архивной финансовой информации
- Верификация данных по курсам валют

Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [1].

Можно выделить несколько уровней информации (рисунок):

- исходные данные (сырые данные, исторические данные или просто данные) – необработанные массивы данных, получаемые в результате наблюдения за некой динамической системой или объектом и отображающие его состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год);
- информация – обработанные данные, которые несут в себе некую информационную ценность для пользователя; сырые данные, представленные в более компактном виде (например, результаты поиска);
- знания — несут в себе некое ноу-хау, отображают скрытые взаимосвязи между объектами, которые не являются общедоступными (в противном случае, это будет просто информация); данные с большой энтропией (или мерой неопределенности).

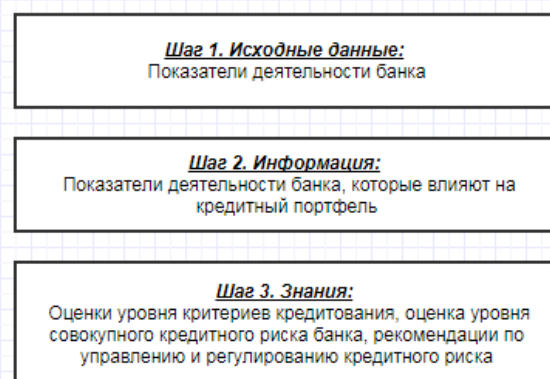


Рисунок – Уровни информации по DM в рамках данной темы

Задачи, решаемые методами Data Mining:

- Классификация – это отнесение объектов (наблюдений, событий) к одному из заранее известных классов.
- Регрессия, в том числе задачи прогнозирования. Установление зависимости непрерывных выходных от входных переменных.
- Кластеризация – это группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность этих объектов. Объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем больше похожи объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация.
- Ассоциация – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события X следует событие Y. Такие правила называются ассоциативными. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).
- Последовательные шаблоны – установление закономерностей между связанными во времени событиями, т.е. обнаружение зависимости, что если произойдет событие X, то спустя заданное время произойдет событие Y.
- Анализ отклонений – выявление наиболее нехарактерных шаблонов.

Проблемы бизнес анализа формулируются по-иному, но решение большинства из них сводится к той или иной задаче Data Mining или к их комбинации. Например, оценка рисков – это решение задачи регрессии или классификации, сегментация рынка – кластеризация, стимулирование спроса – ассоциативные правила. Фактически, задачи Data Mining являются элементами, из которых можно собрать решение подавляющего большинства реальных бизнес задач [2].

Алгоритмы обучения:

- ✓ Для задач классификации характерно «обучение с учителем», при котором построение (обучение) модели производится по выборке, содержащей входные и выходные векторы.
- ✓ Для задач кластеризации и ассоциации применяется «обучение без учителя», при котором построение модели производится по выборке, в которой нет выходного параметра. Значение выходного параметра («относится к кластеру ...», «похож на вектор ...») подбирается автоматически в процессе обучения.
- ✓ Для задач сокращения описания характерно отсутствие разделения на входные и выходные векторы. Начиная с классических работ К. Пирсона по методу главных компонент, основное внимание уделяется аппроксимации данных.

Этапы обучения. Ряд этапов решения задач методами Data Mining:

1. Постановка задачи анализа;
2. Сбор данных;
3. Подготовка данных (фильтрация, дополнение, кодирование);
4. Выбор модели (алгоритма анализа данных);
5. Подбор параметров модели и алгоритма обучения;
6. Обучение модели (автоматический поиск остальных параметров модели);
7. Анализ качества обучения, если неудовлетворительный переход на п. 5 или п. 4; Анализ выявленных закономерностей, если неудовлетворительный переход на п. 1, 4 или 5 [0].

Список использованных источников

1. Data Mining [Электронный ресурс] – Режим доступа: <https://basegroup.ru/community/articles/data-mining> – Дата доступа: 20.03.2019.
2. Data Mining: что внутри [Электронный ресурс] <https://habrahabr.ru/post/95209/> – Режим доступа: – Дата доступа: 20.03.2019.
3. Data Mining [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Data_mining – Дата доступа: 20.03.2019.