Лекция 1 **Методы анализа данных**

В результате развития информационных технологий количество данных, накопленных в электронном виде, растет быстрыми темпами.

Эти данные существуют в различных форматах: тексты, изображения, аудио, видео, гипертекстовые документы, реляционные базы данных и т.д.

Однако подавляющая часть доступной информации не несет для конкретного человека какой-либо пользы, так как он не в состоянии переработать такое количество сведений.

Возникает проблема извлечения полезной для пользователя информации из большого объема данных.

Лекция 1 **Интеллектуальный анализ данных**

Понятие *интеллектуального анализа данных* соответствует широко распространенному термину *Data Mining*, который часто переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний в базах данных.

Data Mining можно охарактеризовать как технологию, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей:

- не очевидных, так как найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем;
- объективных, так как обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным;
- практически полезных, так как выводы имеют конкретное значение, которому можно найти практическое применение.

Состав рынка интеллектуальных технологий

Содержит набор программных продуктов следующих классов:

- средства построения хранилищ данных (*Data Warehousing, ХД*);
- системы оперативной аналитической обработки (*OLAP*);
- информационно-аналитические системы (*Enterprise Information Systems, EIS*);
- средства интеллектуального анализа данных (*Data Mining*);
- инструменты для выполнения запросов и построения отчетов (*query and reporting tools*).

Данные обеспечивают получение информации, поддерживающую решения

Информационная пирамида

Набор данных и их атрибутов

Данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Атрибут – свойство, характеризующее объект: цвет глаз человека,температура воды и т.д.

Выборка - часть всей совокупности, позволяющая получить интересующую нас информацию на ее основе. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности.

Измерение – процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу. Измеряется не сам объект, а его характеристики.

Шкала – правило, в соответствии с которым объектам присваиваются числа.

Наиболее часто встречаются данные, состоящие из *записей*: табличные данные, матричные данные, документальные данные, транзакционные или операционные

Метаданные – это данные о данных (каталоги, справочники и др.)

Задачи *Data Mining*

Задачи подразделяются по типам производимой информации:

- классификация,
- прогнозирование.

В результате решения *задачи классификации* обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных – *классы*; по этим признакам новый объект можно отнести к тому или иному классу (кластеризация).

В ходе решения задачи поиска *ассоциативных правил* отыскиваются закономерности между связанными событиями в наборе данных, которые происходят одновременно.

Задачи *Data Mining* в зависимости от используемых моделей могут быть *дескриптивными (описательными) и прогнозирующими*. Концепция *описательных задач* подразумевает характеристику и сравнение наборов данных.

Прогнозирующие задачи основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных

Основы анализа данных

Описательная статистика, включающая технологии сбора и суммирования количественных данных, используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения.

Цель описательной статистики – обобщить первичные результаты, полученные в результате наблюдений и экспериментов.

В состав описательной статистики входят такие характеристики: среднее; стандартная ошибка; медиана; мода; стандартное отклонение; дисперсия выборки; эксцесс; асимметричность; интервал; минимум; максимум; сумма; счет.

Корреляционный анализ

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде.

Коэффициент корреляции r используется для определения наличия взаимосвязи между двумя свойствами.

Связь между признаками оценивается по шкале Чеддока.

Любая зависимость между переменными обладает двумя важными свойствами: *величиной и надежностью*.

Величина коэффициента корреляции, r	**0,1-0,3**	**0,3-0,5**	**0,5-0,7**	**0,7-0,9**	**0,9-1**
Характеристика силы связи	**Слабая**	**Умеренная**	**Заметная**	**Высокая**	**Весьма высокая**

Надежность зависимости характеризует вероятность, что эта зависимость будет снова найдена на других данных. С ростом величины зависимости переменных ее надежность обычно возрастает

Последовательность этапов регрессионного анализа

При помощи *регрессионного анализа* можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Последовательность этапов регрессионного анализа:

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений;
2. Определение зависимых и независимых (объясняющих) переменных;
3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель;
4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная);
5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии);
6. Оценка точности регрессионного анализа;
7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов;
8. Предсказание неизвестных значений зависимой переменной.

Задачи, решаемые регрессионным анализом

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации.

Основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной.

Определение функции регрессии сводится к выяснению действия на зависимую переменную главных факторов или причин при неизменных прочих равных условиях.

Функция регрессии определяется в виде математического уравнения того или иного типа.

Прогнозирование

Прогнозирование – установление функциональной зависимости между зависимыми и независимыми переменными. Целью прогнозирования является предсказание будущих событий.

Основой для прогнозирования служит историческая информация, хранящаяся в базе данных в виде временных рядов.

В процессе определения структуры и закономерностей временного ряда предполагается обнаружение: шумов и выбросов, тренда, сезонной компоненты, циклической компоненты.

Горизонт прогнозирования должен быть не меньше, чем время, которое необходимо для реализации решения, принятого на основе этого прогноза.

Прогноз может быть краткосрочным, среднесрочным и долгосрочным. *Для построения краткосрочных и среднесрочных прогнозов вполне подходят статистические методы, нейронные сети, деревья решений и линейная регрессия.*

Задача кластеризации

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры, или классы).

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Цель кластеризации – поиск существующих структур.

Кластерный анализ позволяет анализировать показатели различных типов данных (интервальных данных, частот, бинарных данных). *При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах*.

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.

Наряду со *стандартизацией переменных*, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной.

Методы кластерного анализа: иерархический и неиерархический.

Иерархический алгоритм кластерного анализа

Факторный анализ

Факторный анализ – это метод, применяемый для изучения взаимосвязей между значениями переменных. Факторный анализ преследует две цели: сокращение числа переменных и классификацию переменных – определение структуры взаимосвязей между переменными.

При помощи факторного анализа большое число переменных сводится к меньшему числу независимых влияющих величин, которые называются факторами.

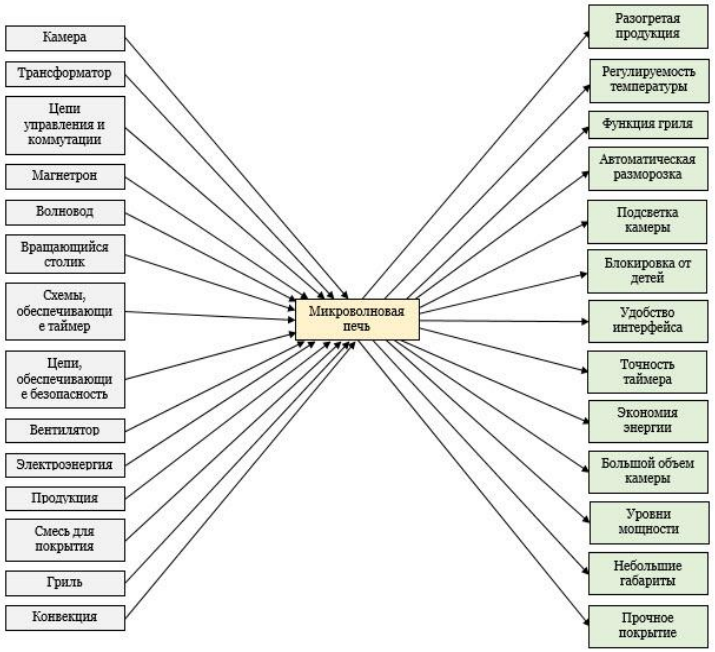
Фактор в "сжатом" виде содержит информацию о нескольких переменных. В один фактор объединяются переменные, которые сильно коррелируют между собой.

В результате факторного анализа отыскиваются такие комплексные факторы, которые как можно более полно объясняют связи между рассматриваемыми переменными.

Задача визуализации

В результате использования визуализации создается графический образ данных. К способам визуального или графического представления данных относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т.д.

Существует такой распространенный и наиболее простой способ представления модели, как *"черный ящик".*



Этапы интеллектуального анализа

Процесс интеллектуального анализа и обработки данных состоит из следующих шести этапов: отбор данных, очистка, обогащение, кодирование, извлечение знаний и сообщение.

Отбор данных - выбор то их подмножество, которое будет подвергнуто анализу.

Очистка - удаление дублирующих записей, исправление типографских ошибок, добавление отсутствующей информации и т.д.

Извлечение знаний - обнаружение ядра процесса интеллектуального анализа и обработки знаний.

Различают четыре различных типа знаний, которые могут быть извлечены из данных:

1. Поверхностное знание. Это информация находится из баз данных, используя (SQL) запросы;
2. Многомерное знание. Это информация может быть проанализирована при использовании интерактивных аналитических инструментальных средств обработки OLAP.
3. Скрытое знание. Это информация может быть найдена с помощью алгоритмов распознавания образов или машинного обучения.
4. Глубокое знание. Это информация, которая хранится в базе данных, но может быть обнаружена только в том случае, если имеется ключ, который сообщит нам, где смотреть.

Сообщение о результатах процесса обнаружения знаний. Можно использовать любой редактор сообщений или графическое инструментальное средство.

Инструментальные средства анализа данных

Инструменты *Data Mining* во многих случаях рассматриваются как составная часть *BI*-платформ, в состав которых также входят средства построения хранилищ и витрин данных, средства обработки неожиданных запросов (*ad-hoc query*), средства отчетности (*reporting*), а также инструменты *OLAP*.

К категории наборов инструментов относятся универсальные средства, которые включают методы классификации, кластеризации и предварительной подготовки данных. К этой группе относятся такие известные коммерческие инструменты, как:

- *IBM Intelligent Miner for Data* - поддерживает полный *Data Mining-процесс* ;
- *SPSS* - один из наиболее популярных инструментов, поддерживается множество методов *Data Mining*;
- *Statistica Data Miner* - инструмент обеспечивает всесторонний, интегрированный статистический анализ данных, имеет мощные графические возможности, управление базами данных, а также приложение разработки систем;
- и др.

Литература

Интеллектуальный анализ данных: учеб. пособие для студентов специальности 080801.65 «Прикладная информатика (в экономике)» / Саратовский государственный социально-экономический университет. – Саратов, 2012. – 92 с.
ISBN 978-5-4345-0113-2