Лекция № 1

Введение. Основные определения и понятия курса

Структура курса

- Лекции – 34 ч.
- Лабораторные работы – 17 ч. (4)
- Самостоятельная работа – 64 ч.
- Контрольная работа
- Зачет
- Всего – 119 ч.

Литература

- Методичка «Решение задач ИАД» в среде Statistica
- А.А. Барсегян «Методы и модели анализа данных: OLAP и Data Mining», Санкт-Петербург, изд-во БХВ-Петрбург, 2004 г. (коллектив авторов Санкт- Петербургский гос. тех. Университет – ЛЭТИ и компания ZSoftLtd – разработка информационно-аналитических систем). Книга – обзор технологий обработки данных, первая на русском языке.
- Факторный, дискриминантный и кластерный анализ/Пер. с англ. А.М. Хотинского. Под ред. И.С. Енюкова. -М.: Финансы и статистика, 1989.
- Электронный учебник StatSoft по анализу данных.

- Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. – М.: Финансы и статистика, 1974. – 240 с.
- Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности.- М.: Финансы и статистика, 1989.
- Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ/Пер. с нем. В.М. Ивановой. –М.: Финансы и статистика, 1985.
- Боровиков В.П. Statistica. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.
- Боровиков В.П., Боровиков И.П. Statistica - Статистический анализ и обработка данных в среде Windows. – М.: «Филин», 1997. – 608 с.

- Бериков В.Б. Анализ статистических данных с использованием деревьев решений: Учебное пособие. – Новосибирск. Изд-во НГТУ, 2002. – 60 с.
- Авдеенко Т.В. Компьютерные методы анализа временных рядов и прогнозирования. – Новосибирск: НГТУ, 2008. – 271 с.
- Дайитбегов Д.М. Компьютерные технологии анализа данных в эконометрике. – Изд-во Инфра-М, 2008. – 578 с.

1. Определение ИАД

- **Интеллектуальный анализ данных (ИАД, data mining) представляет собой новое направление в области информационных систем (ИС), ориентированное на решение задач поддержки принятия решений на основе количественных и качественных исследований сверхбольших массивов разнородных ретроспективных данных.**

1. Определение ИАД

- **ИАД (Data Mining) - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации). При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.**

1. Определение ИАД

- Data Mining" – это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого закономерностей с целью достижения преимуществ в бизнесе (SAS Institute).

1. Определение ИАД

- **ИАД “Data Mining” – это процесс, цель которого – обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образов и методов математической статистики (Gartner Group).**

1. Определение ИАД

- **“Data Mining” – технология поиска характеризующих объект скрытых зависимостей и взаимосвязей, проявляющихся через данные о нем.**

1. Определение ИАД (StatSoft)

- **ИАД (*Data Mining*) - процесс аналитического исследования больших массивов информации (обычно экономического характера) с целью выявления определенных закономерностей и систематических взаимосвязей между переменными, которые затем можно применить к новым совокупностям данных.**

1. Определение Data Mining

Data Mining – исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком. *(Григорий Пятецкий-Шапиро, 1996 г. – основатель направления)*

1. Определение Data Mining
Основные свойства знаний:

- знания должны быть новые, ранее неизвестные. Затраченные усилия на открытие знаний, которые уже были известны пользователю – не окупаются.

- знания должны быть нетривиальны. Результаты анализа должны отражать неочевидные, неожиданные закономерности в данных, составляющие так называемые скрытые знания. Наример, если знания получены простым просмотром – привлечение мощных средств Data Mining не оправдывается.

- знания должны быть практически полезны. Знания должны быть применимы на новых данных с достаточно высокой степенью достоверности и приносить выгоду при их применении.

- знания должны быть доступны для понимания человеку. Закономерности д.б. логически объяснимы, иначе они могут быть случайны и представлены в понятном для человека виде.

В этом контексте знания представляют собой краткое обобщенное описание основного содержания информации, представленной в данных (скрытые закономерности, корреляции, тенденции, обобщенные характеристики данных типа “если-то” и т.д.).

1. Определение *KNOWLEDGE DISCOVERY IN DATABASES (POLYANALYST)*

- «ОБНАРУЖЕНИЕ ЗНАНИЙ В БАЗАХ ДАННЫХ») – АНАЛИТИЧЕСКИЙ ПРОЦЕСС ИССЛЕДОВАНИЯ ЧЕЛОВЕКОМ БОЛЬШОГО ОБЪЕМА ИНФОРМАЦИИ С ПРИВЛЕЧЕНИЕМ СРЕДСТВ АВТОМАТИЗИРОВАННОГО ИССЛЕДОВАНИЯ ДАННЫХ С ЦЕЛЬЮ ОБНАРУЖЕНИЯ СКРЫТЫХ В ДАННЫХ СТРУКТУР ИЛИ ЗАВИСИМОСТЕЙ.
- ПРЕДПОЛАГАЕТСЯ ПОЛНОЕ ИЛИ ЧАСТИЧНОЕ ОТСУТСТВИЕ АПРИОРНЫХ ПРЕДСТАВЛЕНИЙ О ХАРАКТЕРЕ СКРЫТЫХ СТРУКТУР И ЗАВИСИМОСТЕЙ.

1. Этапы *KDD*

- ПОСТАНОВКА ЗАДАЧИ (В ТЕРМИНАХ ЦЕЛЕВЫХ ПЕРЕМЕННЫХ);
- ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА (ПРЕОБРАЗОВАНИЕ ДАННЫХ К ДОСТУПНОМУ ДЛЯ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ФОРМАТУ)
- ОБНАРУЖЕНИЕ СРЕДСТВАМИ АВТОМАТИЧЕСКОГО ИССЛЕДОВАНИЯ ДАННЫХ (DATA MINING) СКРЫТЫХ СТРУКТУР ИЛИ ЗАВИСИМОСТЕЙ;
- АПРОБАЦИЯ ОБНАРУЖЕННЫХ МОДЕЛЕЙ НА НОВЫХ, НЕ ИСПОЛЬЗОВАВШИХСЯ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ ДАННЫХ И ИНТЕРПРЕТАЦИЯ ЧЕЛОВЕКОМ ОБНАРУЖЕННЫХ МОДЕЛЕЙ.

1. Определение ИАД, ВМ (Губарев В.В.)

- Одно из направлений ИАД: поиск, выбор, синтез методов и средств обработки и анализа данных с учетом поставленных целей исследования.
- Технология, которая реализует этот вариант ИАД – вариативное моделирование (ВМ).
- ВМ – есть метод исследования, основанный на замене исследуемого объекта-оригинала набором разнообразных моделей его и на работе с ними.

1. Определение ВМ (Губарев В.В.)

- Отличительная особенность ВМ от обычного (классического) заключается в том, что здесь обязательным является построение и применение в процессе моделирования не менее двух разных моделей исследуемого (моделируемого) объекта.
- *Это могут быть модели разных классов (познавательные и прагматические; материальные и идеальные; микро, макро и мегамодели; реальные, виртуальные и абстрактные; априорные и апостериорные; регулярные и иррегулярные; стохастические и хаотические и т.п.), одного класса, но разных типов, склонностей; использующие разные уровни описания объекта, средства и технологии их построения, интерпретации и применения и т.п.*
- Виды моделей зависят от метода их создания. Наиболее распространенные: правила, деревья решений, кластеры, математические функции.

2. Классификация задач ИАД

1. **Выявление ассоциативных взаимосвязей в данных**

- **Ассоциация используется для определения закономерностей в событиях или процессах.**
- **Ассоциации связывают различные факты одного события.**
- **Найденные закономерности представляются в виде правил и используются как для лучшего понимания природы явления так и для предсказания появления события.**

1. Выявление ассоциативных взаимосвязей в данных

Результатом ассоциативного анализа являются правила вида:

Если факт *A* является частью события, то с вероятностью *X%* факт *B* будет частью того же события.

2. Классификация задач ИАД

2. **Выявление последовательностей**

- **Последовательные шаблоны аналогичны ассоциациям с той лишь разницей, что связывают события, разнесенные во времени.**
- **Такая задача является разновидностью задачи поиска ассоциативных правил и называется сиквенциальным анализом.**

2. Классификация задач ИАД

3. Кластеризация объектов – разделение исследуемого множества объектов на группы «похожих» объектов, называемых кластерами.

- В процессе кластеризации методами ИАД определяются схожие характеристики объектов и на их основе объединяются объекты в классы (кластеры).

2. Классификация задач ИАД

4. Классификация объектов – отнесение объектов к одному из известных классов на основе их характеристик.

2. Классификация задач ИАД

5. Нахождение исключений, исключительных ситуаций, записей, которые резко отличаются чем-либо от основного множества записей (группы больных).

2. Классификация задач ИАД

6. Задачи регрессии – задача определения значения одного из параметров анализируемого объекта (характеристики) на основе значений других характеристик (все характеристики – количественные).

- Задачи взаимосвязаны, из одной вытекает другая.

3. Области применения ИАД

- **Сфера применения Data Mining ничем не ограничена - Data Mining нужен везде, где имеются какие-либо данные.**

3. Области применения ИАД

- **Интернет-технологии. Применяется для построения рекомендательных систем Интернет-магазинов и для решения проблемы персонализации (автоматическое распознавание принадлежности клиента к определенной целевой группе) посетителей web-сайтов.**
- **Понятие web-Mining - применение технологий DM для анализа информации, содержащейся на web-узлах. Например, обнаружение закономерностей в поведении пользователей конкретного web-узла: в какой последовательности и какие страницы запрашиваются пользователями и какими группами пользователей.**

3. Области применения ИАД

- **Банковское дело. Сегментация клиентов, выявление мошенничества с кредитными картами, прогнозирование изменения клиентуры, анализ финансовых рисков.**

- **Торговля. Анализ потребительской корзины, исследование временных шаблонов, создание прогнозирующих моделей, оптимизация складских запасов.**

3. Области применения ИАД

Страховой бизнес. Сегментация клиентов, выявление фактов мошенничества, анализ страховых рисков, разработка новых продуктов, расчет страховых премий.

Телекоммуникации. Анализ лояльности клиентов, сегментирование клиентской базы и услуг, анализ внешних факторов на отказы оборудования, выявление случаев несанкционированного доступа к сети.

3. Области применения ИАД

Производственные предприятия. Оптимизация закупок, диагностика брака на ранних стадиях, диагностика оборудования, маркетинг.
Нефтегазовая отрасль. Диагностика оборудования и нефте/газопроводов, прогнозирование цен, разведка месторождений, анализ влияния внешних и внутренних факторов на объемы продаж.

4. Математический аппарат ИАД

ИАД – это многодисциплинарный подход, который включает в себя методы математической статистики и теории вероятности, методы визуализации данных, нейросетевые методы, методы деревьев решений, нечеткую логику, экспертный анализ, эволюционное программирование, генетические алгоритмы и т.д.

4. Классификация методов ИАД

- Методы статистической обработки данных
- Кибернетические методы оптимизации
- Традиционные методы решения оптимизационных задач
- Экспертные методы
- Интегрированные технологии, вариативное моделирование

Методы статистической обработки данных

- Предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения и ее параметров).
- Выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ).
- Многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластер-анализ, компонентный анализ, факторный анализ).
- Динамические модели и прогноз на основе временных рядов.

Методы статистической обработки данных

- Достоинства
- Построенные модели “прозрачны” и допускают интерпретацию.
- Возможно оценить статистическую значимость полученных результатов.
- Разработано много алгоритмов и накоплен большой опыт их применения в научных и инженерных приложениях.

Методы статистической обработки данных

- Недостатки
- Требуют сохранение неизменных условий эксперимента (требования статистического ансамбля).
- Требуют априорных допущений об исследуемых данных (закон распределения исследуемых данных, отсутствие пропусков в данных, отсутствие аномальных выбросов и т.д.).

Методы статистической обработки данных

- Программное обеспечение
- Statistica (Statsoft), SAS (компания SAS Institute), SPSS (SPSS), Statgraphics (Statistical Graphics).

Кибернетические методы оптимизации

- Нейронные сети (Neural Nets)
- Генетические алгоритмы (Genetic algorithms)
- Эволюционное программирование (Evolutionary programming)

Нейронные сети

- Достоинства
- Не требуют априорных допущений о природе исследуемых данных.
- Удобны при работе с нелинейными зависимостями, зашумленными и неполными данными.

Нейронные сети

- Недостатки
- “Черный ящик”: модель не может объяснить выявленные знания (не поддается интерпретации).
- Программное обеспечение
- BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic), 4Thought.

Генетические алгоритмы

- Достоинства
- Красота подхода, близость метода к природному механизму (имитация процесса естественного отбора в природе).
- Высокая скорость решения задач большой размерности.

Генетические алгоритмы

- Недостатки
- Невозможно оценить статистическую значимость результата.
- Сложность использования метода (сложность постановки задачи, сложность определения критерия отбора хромосом и т.д.).
- Программное обеспечение
- GeneHunter (Ward Systems Group)

Эволюционное программирование

- Достоинства
- Высокая степень автоматизации (автоматическое обнаружение в массивах данных кластеров, случайных выбросов, скрытых закономерностей, фильтрация шумов; визуализация обнаруженных зависимостей, оценка статистической значимости результатов и т.д.).

Эволюционное программирование

- Недостатки
- Сложность (невозможность) содержательной интерпретации полученных результатов
- Программное обеспечение
- PolyAnalyst (Мегапьютер Интеллидженс).

Традиционные методы решения оптимизационных задач

- Методы исследования операций, включающие в себя различные виды математического программирования (линейное, нелинейное, дискретное, целочисленное)
- динамическое программирование,
- методы теории систем массового обслуживания
- Программное обеспечение
- MathCAD и MatLab.

Экспертные методы

- Деревья решений
- Ассоциативный анализ
- Предметно-ориентированные системы анализа ситуаций
- Методы визуализации

Деревья решений

- Достоинства
- Наглядность (возможность графического представления результатов, иерархическая структура дерева).
- Простота интерпретации полученных результатов.

Деревья решений

- Недостатки
- Проблема оценки статистической значимости результатов.
- Программное обеспечение
- C5.0 (RuleQuest, Австралия); Clementine (Integral Solutions, Великобритания); SIPINA (University of Lyon, Франция); IDIS (Information Discovery, США), Scenario.

Ассоциативный анализ

- Достоинства
- Простота (для осуществления прогноза или выбора решения в прошлом находятся аналоги наличной ситуации, и выбирается тот же ответ, который был для них правильным).

Ассоциативный анализ

- Недостатки
- В процессе решения не создаются модели и правила, обобщающие предыдущий опыт. Программное обеспечение
- KATE tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США).

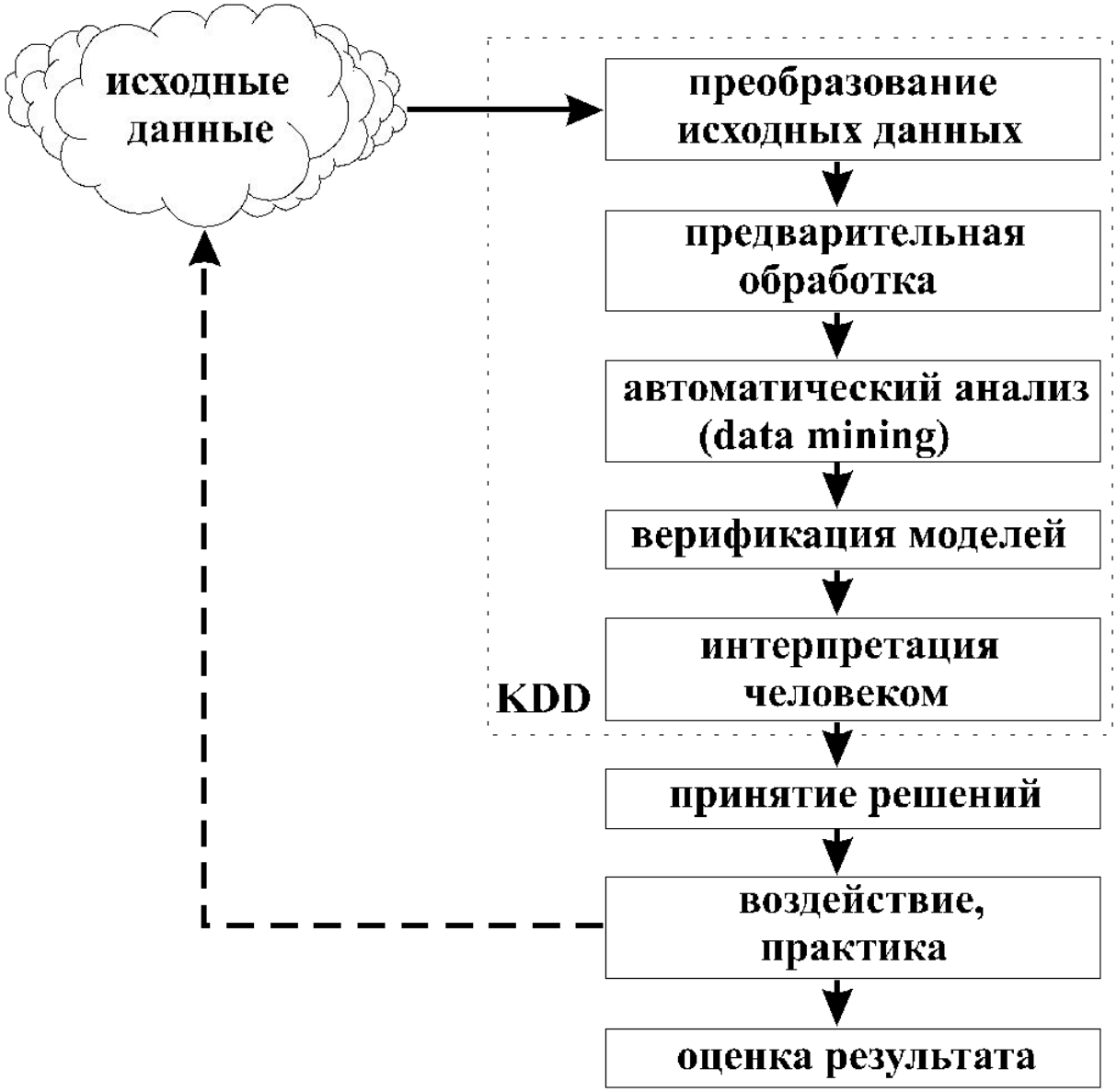
Методы визуализации

- Достоинства
- Наглядность, простота.
- Недостатки
- Высокая доля субъективизма в интерпретации результатов.
- Отсутствие аналитических моделей.
- Программное обеспечение
- MineSet (Silicon Graphics).

Интегрированные технологии, вариативное моделирование

- Достоинства
- Эффективность (можно выбирать подходы адекватные задачам, или сравнивать результаты применения разных подходов).
- Недостатки
- Сложные средства поддержки (программное и аппаратное обеспечение), высокая стоимость.
- Программное обеспечение: Scenario, MineSet, Statistica.

Технология KDD



Особенности технологий ИАД

- Технологии ИАД в большей степени ориентированы на практическое приложение полученных результатов, чем на выяснение природы явления.
- При ИАД нас не очень интересует конкретный вид зависимости между переменными. Основное внимание уделяется поиску решений, на основе которых модно получить достоверный прогноз.
- В ИАД широко используют модели типа «черный» ящик.

Требования к результатам ИАД

- Результат должен быть понятен пользователю-нематематику.
- Результат должен быть пригодным для дальнейшей обработки компьютерными программами, т.е. требование «прозрачности» для человека и машины.
- Например, правила «если-то» таким условиям удовлетворяют.

Связь технологий Data Warehousing и OLAP с методами ИАД