Лекция 3. СОВРЕМЕННЫЕ ПРОГРАММНЫЕ СРЕДСТВА АНАЛИЗА ИНФОРМАЦИИ

До появления аналитических платформ анализ данных осуществлялся в основном в статистических пакетах. Их применение требовало от пользователя высокой квалификации. Большинство алгоритмов, реализованных в статистических пакетах, не позволяло эффективно обрабатывать большие объемы информации. Для автоматизации рутинных операций приходилось использовать встроенные языки программирования.

В конце 80-х гг. произошел стремительный рост объемов информации, накапливаемой на машинных носителях, и увеличились потребности бизнеса в применении анализа данных. Ответом стало появление новых парадигм в анализе, таких как хранилища данных, машинное обучение, Data Mining, Knowledge Discovery in Databases. Это позволило популяризировать анализ данных, поставить его на промышленную основу и решить огромное число бизнес-задач с большим экономическим эффектом.

Венцом развития анализа данных стали специализированные программные системы – аналитические платформы, которые полностью автоматизировали все этапы анализа от консолидации данных до эксплуатации моделей и интерпретации результатов.

Statistica. Statistica — программный пакет для статистического анализа, разработанный компанией StatSoft, реализующий функции анализа данных, управления данными, добычи данных, визуализации данных с привлечением статистических методов.

Statistica предоставляет мощные и удобные в использовании инструменты для статистического и графического анализа, прогнозирования, data mining, создания собственных пользовательских приложений, интеграции, совместной работы, web-доступа и др. Пакет обладает широкими графическими возможностями, позволяет выводить информацию в виде различных типов графиков (включая научные, деловые, трёхмерные и двухмерные графики в различных системах координат, специализированные статистические графики — гистограммы, матричные, категорированные графики и др.), все компоненты графиков настраиваются.

Все аналитические инструменты, имеющиеся в системе, доступны пользователю и могут быть выбраны с помощью альтернативного пользовательского интерфейса. Пользователь может всесторонне

автоматизировать свою работу, начиная с применения простых макросов для автоматизации рутинных действий вплоть до углубленных проектов, включающих в том числе интеграцию системы с другими приложениями или Интернет. Технология автоматизации позволяет даже неопытному пользователю настроить систему на свой проект.

Гибкая и мощная технология доступа к данным позволяет эффективно работать как с таблицами данных на локальном диске, так и с удаленными хранилищами данных.

Statistica состоит из набора модулей, в каждом из которых собраны тематически связные группы процедур. При переключении модулей можно либо оставлять открытым только одно окно приложения Statistica, либо все вызванные ранее модули, поскольку каждый из них может выполняться в отдельном окне (как самостоятельное приложение Windows)

Преимущества пакета Statistica:

1. Наибольший из всех изученных пакетов инструментарий визуализации полученных результатов;
2. Наличие руссифицированной версии;
3. В пакете представлена полная реализация алгоритмов статистической классификации;
4. Наличие возможности реализовать и использовать собственные алгоритмы (через написание макросов).
5. Алгоритмы интеллектуального анализа данных (Data Mining)

Недостатки пакета Statistica:

1. Применение пакета требует высокой теоретической подготовки в ТВиМС.
2. Отсутствие реализации некоторых важных тестов временных рядов (в частности – тестов на стационарность)

SPSS. SPSS Statistics (аббревиатура англ. «Statistical Package for the Social Sciences» — «статистический пакет для общественных наук») — компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в общественных науках.

Платформа IBM SPSS предлагает передовые инструменты статистического анализа, обширную библиотеку алгоритмов машинного обучения, анализа текста, расширения компонентов с открытым кодом, интеграции с большими данными и

беспрепятственного внедрения в приложения. Благодаря простоте эксплуатации, гибкости и масштабируемости IBM SPSS отлично подходит пользователям с любым уровнем подготовки для реализации проектов любого объема и сложности, направленных на поиск новых возможностей, повышение эффективности и снижение рисков.

Наряду с использованием своего собственного типа данных, пакет SPSS, может считывать данные практически из любых типов файлов и использовать их для создания отчетов в форме таблиц, графиков и диаграмм, а также вычислять описательные статистики, производить сложный статистический анализ и моделирование.

Пакет имеет модульную структуру. Модули пакета представляют собой интегрированную совокупность программных продуктов, обеспечивающих комплексное исследование – от планирования до управления данными, выполнения анализа и представления результатов.

Преимущества пакета SPSS:

1. Имеются русифицированные версии пакета.
2. Позволяет параллельно обрабатывать несколько подвыборок.
3. Простота в освоении.
4. Имеются специфические методы, нацеленные исключительно на маркетинговые и социологические исследования (например, Conjoint analysis). Удобен при обработке результатов опроса.
5. Имеется модуль для автоматизации процесса разработки анкеты и ввода результатов опросов (Data Entry).

Недостатки пакета SPSS:

1. Отсутствует возможность реализации собственных алгоритмов;
2. Существенно уступает в глубине анализа данных

EViews. EViews представляет собой современный статистический пакет, «заточенный» под анализ временных рядов. Особо широкие возможности открывает Eviews при анализе данных, представленных в виде временных рядов. Особо широкие возможности открывает Eviews при анализе данных, представленных в виде временных рядов.

Пакету свойственный обширный инструментарий, предназначенный для обработки пользовательских данных. Eviews поможет в кратчайшие сроки сделать фундаментальный анализ

введенной информации и предоставить точный прогноз на базе высчитанных взаимосвязей и статистики.

EViews дает возможность производить регрессионный анализ, выполнять графическое моделирование и составлять визуальные представления для самых разных типов данных. Программа отлично справляется с блоками информации, сформированными в виде временных рядов.

Преимущества пакета EVews:

1. Компактность: программа содержит меньшее количество модулей;
2. Наиболее полный из всех стат. пакетов набор алгоритмов анализа временных рядов (тесты на стационарность, в т.ч. расширенный тест Дики-Фулера, тест Хаусмана)
3. Возможность исследования панельных данных.
4. Возможность анализа финансовых временных рядов на основе моделей условной гетероскедастичности.

Недостатки пакета EViews:

1. Фактически отсутствует возможность реализации собственных алгоритмов
2. По сравнению с пакетом Statistica – более слабые возможности визуализации
3. Отсутствие руссифицированной версии

Excel. Microsoft Excel — программа для работы с электронными таблицами, созданная корпорацией Microsoft для Microsoft Windows, Windows NT и Mac OS, а также Android, iOS и Windows Phone. Она предоставляет возможности экономико-статистических расчетов, графические инструменты и, за исключением Excel 2008 под Mac OS X, язык макропрограммирования VBA (Visual Basic for Application). Microsoft Excel входит в состав Microsoft Office и на сегодняшний день Excel является одним из наиболее популярных приложений в мире.

API позволяет открывать таблицы Excel в ряде других приложений. Это включает в себя открытие документов Excel на веб-страницах с помощью ActiveX или таких плагинов, как Adobe Flash Player. Проект Apache POI представляет Java-библиотеки для чтения и записи электронных таблиц Excel. Также предпринимались попытки копировать таблицы Excel в веб-приложения с использованием разделённых запятыми значений (CSV).

Ценной возможностью Excel является возможность писать код на основе Visual Basic для приложений (VBA). Этот код пишется с использованием отдельного от таблиц редактора. Управление электронной таблицей осуществляется посредством объектно-ориентированной модели кода и данных. С помощью этого кода данные входных таблиц будут мгновенно обрабатываться и отображаться в таблицах и диаграммах (графиках). Таблица становится интерфейсом кода, позволяя легко работать, изменять его и управлять расчётами.

Функциональность VBA делала Excel легкой мишенью для макровирусов. И это было серьёзной проблемой до тех пор, пока антивирусные продукты не научились обнаруживать их. Фирма Microsoft, с опозданием приняв меры для уменьшения риска, добавила возможность выбора режима безопасности:

- полностью отключить макросы
- включить макросы при открытии документа
- доверять всем макросам, подписанным с использованием надёжных сертификатов.

Преимущества программного продукта Excel:

1. Наличие огромного количества встроенных стандартных функций
2. Элементы упрощенного интерфейса в программе
3. Возможность импортировать информацию из большинства учетных систем
4. Быстрое и недорогое внедрение процессов планирования

Недостатки программного продукта Excel:

1. Консолидировать данные нужно вручную
2. Отсутствуют функции защиты данных от исправлений
3. Отсутствуют гибкие механизмы разграничения доступа к данным
4. Низкая производительность при работе с большими объемами данных

RStudio. RStudio — свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графикой. RStudio представляет собой бесплатную интегрированную среду разработки (IDE) для R.

Благодаря ряду своих особенностей этот активно развивающийся программный продукт далает работу с R очень удобной.

RStudio частично написана на языке программирования C++ и использует фреймворк Qt для графического интерфейса пользователя. Работает на всех основных платформах (Windows, Mac и Linux), а также может работать как сервер, что позволяет нескольким пользователям получать доступ к RStudio IDE с помощью веб-браузера.

Редактор кода RStudio включает ряд опций для продуктивной работы, в частности подсветку кода, автоматическое завершение кода, одновременное редактирование нескольких файлов, поиск и замену определенных частей кода.

Кроме того, в RStudio имеются гибкие возможности по выполнению кода непосредственно из окна редактора. Для многих пользователей это является предпочтительным способом работы с R. Выполнение команд из окна Редактора кода вместо командной строки Консоли облегчает воспроизведение одних и тех же команд и позволяет "упаковать" такие команды в одну функцию для последующего использования.

Преимущества IDE RStudio:

1. Язык создавался специально для анализа данных: запись конструкций языка понятна многим специалистам в области.
2. Многие функции, необходимые для анализа данных, являются встроенными функциями языка. Проверка статистических гипотез зачастую занимает лишь несколько строк кода.
3. Удобный репозиторий пакетов и обилие готовых тестов практически под все методы Data Science и машинного обучения.
4. Эффективная работа с векторами и матрицами.
5. Несколько качественных пакетов для визуализации данных для различных задач (ggplot2, lattice, ggvis, googleVis, rCharts и т.д.).

Недостатки IDE RStudio:

1. Низкая производительность. Однако в системе присутствуют пакеты, позволяющие повысить скорость работы (pqR, renjin, FastR, Riposte и т. д.). При работе с

большими массивами данных рекомендуется использовать библиотеки data.table and dplyr.

2. Специфичность в сравнении со стандартными языками программирования, так как язык узкоспециализированный (например, индексация векторов начинается вместо нуля с единицы).

3. R прекрасный инструмент для статистики и соответствующих stand-alone приложений, но менее эффективен в тех областях, где традиционно применяются языки общего назначения.

Deductor. Deductor – это аналитическая платформа, основа для создания законченных прикладных решений в области анализа данных. Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от консолидации данных до построения моделей и визуализации полученных результатов.

Deductor предоставляет аналитикам инструментальные средства, необходимые для решения самых разнообразных аналитических задач: корпоративная отчетность, прогнозирование, сегментация, поиск закономерностей – эти и другие задачи, где применяются такие методики анализа, как OLAP, Knowledge Discovery in Databases и Data Mining. Deductor является идеальной платформой для создания систем поддержки принятия решений.

Объединение всех описанных выше механизмов в Deductor Studio обеспечивает принципиально новое качество анализа: быстрая разработка и адаптация решений, интеграция в существующую инфраструктуру, эволюционное развитие от простой отчетности к глубокой аналитике.

Поддерживает нерегламентированных «ad-hoc» запросов к хранилищу данных с последующей удобной визуализацией. Исправляет ошибки в данных на основе алгоритмов машинного обучения, статистики или по жестким правилам.

Преимущества программного продукта Deductor:

1. Встроенная интеграция с десятками источников данных

2. Высокая производительность

3. Все технологии анализа: Data Warehouse, OLAP, Data Mining

4. Аналитика от простых формул до самообучающихся алгоритмов
5. Апробированная платформа

Недостатки программного продукта Deductor:

1. Однопоточность
2. Стоимость продукта
3. Нестабильность при перегрузке

Weka. Weka (Waikato Environment for Knowledge Analysis) — свободное программное обеспечение для анализа данных и машинного обучения, написанное на Java в Университете Уаикато (Новая Зеландия), распространяющееся по лицензии GNU GPL.

Weka — библиотека алгоритмов машинного обучения для решения задач интеллектуального анализа данных (data mining). Система позволяет непосредственно применять алгоритмы к выборкам данных, а также вызывать алгоритмы из программ на языке Java.

Цели проекта — создать современную среду для разработки методов машинного обучения и применения их к реальным данным, сделать методы машинного обучения доступными для повсеместного применения. Предполагается, что с помощью данной среды специалист в прикладной области сможет использовать методы машинного обучения для извлечения полезных знаний непосредственно из данных, возможно, очень большого объёма.

Пользователями Weka являются исследователи в области машинного обучения и прикладных наук. Она также широко используется в учебных целях. Weka содержит средства для предварительной обработки данных, классификации, регрессии, кластеризации, отбора признаков, поиска ассоциативных правил и визуализации. Weka хорошо подходит для разработки новых подходов в машинном обучении.

Weka предоставляет прямой доступ к библиотеке реализованных в ней алгоритмов. Это позволяет легко использовать уже реализованные алгоритмы из других систем, реализованных на Java. Например, эти алгоритмы можно вызывать из MATLAB. В частности, интерфейс доступа к алгоритмам Weka из MATLAB реализован в некоторых алгоритмических пакетах машинного обучения таких, как Spider и MATLABArsenal.

Преимущества программного продукта Weka:

1. Возможность доступа к классам из любой программы, написанной на языке JAVA
2. Наличие библиотеки алгоритмов машинного обучения
3. Поддерживает поиск ассоциативных правил, формирование наборов исходных данных

Недостатки программного продукта Weka:

1. Отсутствие возможности делать сложные преобразования
2. Отсутствие гибкости визуализации

KNIME. Konstanz Information Miner является бесплатной платформой для анализа данных, отчетности и интеграции с открытым исходным кодом.

Knime Analytics Platform – open source фреймворка для анализа данных. Данный фреймворк позволяет реализовывать полный цикл анализа данных включающий чтение данных из различных источников, преобразование и фильтрацию, собственно анализ, визуализацию и экспорт.

В Knime процесс программирования логики осуществляется через создание Workflow. Workflow состоит из узлов которые выполняют ту или иную функцию (например чтение данных из БД, трансформация, визуализация).

Workflow состоит из узлов (или «нод»). Практически у каждого узла есть конфигурационный диалог в котором можно настраивать свойства.

Преимущества программного продукта Weka:

1. Позволяет конечным пользователям использовать визуальное программирование потоков выполнения анализа
2. Содержит набор инструментов для очистки, анализа и визуализации данных
3. Open Source

Недостатки программного продукта Weka:

1. Отсутствие возможности делать сложные преобразования
2. Сложная настройка параметров машинного обучения