

Анализ данных: применение СПО в образовательном процессе и научно-практической деятельности

М.И. Гальченко¹, А.Г. Гущинский¹

¹Санкт-Петербургский Государственный Аграрный Университет,
maxim.galchenko@gmail.com

Аннотация — Рассматриваются вопросы применения пакета интеллектуального анализа данных KNIME в обучении студентов инженерных специальностей. Приведены примеры применения этого пакета к задачам обработки опросов и анализу надежности в системах «человек-среда-машина»

Ключевые слова — KNIME, data mining, анализ данных

I. Введение

Обработка статистических данных — это задача, с которой сталкиваются практически в любой компании. При этом, чаще всего, вопрос больших объемов информации и необходимости профессионального программного обеспечения ставится в подразделениях маркетинга. Аналитическая подготовка лучше всего, в силу профессиональной необходимости, поставлена именно у специалистов в этой области. Однако, если говорить о специфической информации, то лучшим вариантом, по мнению ряда авторов, является работа с ней специалистов-аналитиков, имеющих базовое образование в сфере, являющейся источником данных [1].

Обычно подготовка в части анализа данных в случае «непрофильности» ограничивается разве что небольшим набором сведений из раздела «Математическая статистика» и, в лучшем случае, подкрепляется небольшими навыками работы в редакторе таблиц. При этом, по оценкам специалистов Iowa State University, компетенция «Аналитические способности» - одна из важнейших в профессиональной деятельности агроинженера.

Если говорить о специализированном статистическом и data mining ПО, то наибольшее применение как в реальном секторе, так и в образовании находят такие пакеты, как SAS, SPSS Clementine, SPSS и Statistica (по данным третьего опроса компании Rexer Analytics, 2009 год), при этом удовлетворенность пользователей данными пакетами достаточно высока. Данные, на которых построены выводы, относятся к международному опросу, по опыту же работы в отечественных компаниях можно говорить лишь о том, что используются возможности данных пакетов далеко не полностью. Однако, при обучении специалистов именно SPSS и Statistica нашли признание в нашей стране.

Тот же опрос показывает, что в образовательных целях респонденты охотно используют R, WEKA и RapidMiner — свободные программные продукты. Если использовать поиск в русскоязычном сегменте сети Internet можно сделать неутешительный вывод: отсутствие большого количества учебных пособий начального уровня, описаний косвенно свидетельствует о низкой известности аналитического СПО в России.

Таким образом, мы имеем замкнутый круг: компании с радостью сокращают свои затраты с условием сохранения качества, но рынок не предлагает им специалистов, знающих как это сделать. Выходом из этого положения, как видится, стало бы включение в программы таких курсов, как «Компьютерные технологии в науке и образовании» и подобным им СПО, предназначенного для обработки статистических данных.

II. Обзор аналитического СПО

Существует три наиболее распространенных пакета, используемых для решения аналитических задач: RapidMiner (ранее YALE), язык статистического программирования R и Konstanz Information Miner (KNIME).

Язык статистического программирования R — хорошо развитый инструмент решения статистических задач. Язык позволяет эффективно решать задачи анализа статистических данных, а также решать задачи data mining. В свободном доступе размещено достаточно много руководств на русском языке, позволяющих ознакомиться с функционалом и правилами работы в среде, из которых стоит выделить «Анализ данных с R», авторы А. Б. Шипунов, А. И. Коробейников, Е. М. Балдин; «Машинное обучение. Лабораторный практикум», авторы Н.Ю. Золотых, А.Н. Половинкин. Однако, данный пакет не имеет развитого графического интерфейса, однако (рис. 1) имеющиеся интерфейсы позволяют выполнять основные задачи.

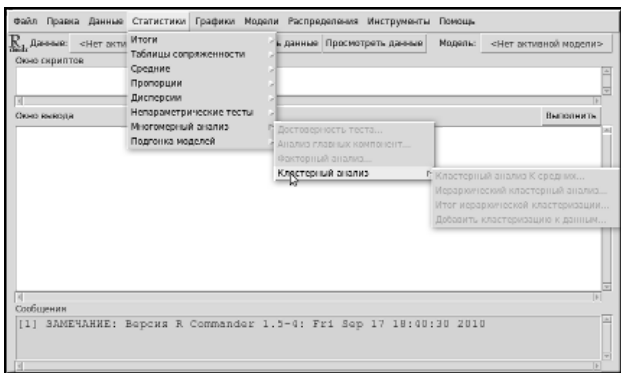


Рис. 1. R Commander — GUI для языка статистического программирования R

Пакет дает возможность строить достаточно «продвинутой» графику.

RapidMiner (актуальная версия — 5.0) — data mining пакет, разрабатываемый Rapid-I GmbH, имеет очень высокие оценки по данным отчета Rexel Analytics за 2009 год: 80% пользователей, определивших данный пакет

как основной, собирается использовать его в течение ближайших трех лет. Аналогичный показатель для R — около 50%. Пакет имеет интерфейс в корне отличающийся от того, который привычен по SPSS и прочим «классическим» программам анализа данных (рис. 2).

В этом пакете применен принцип построения потоков данных (data flows), вместо «плоской» системы обработки с помощью команд.

Центральными элементами в такой схеме обработки становятся узлы (nodes), в которых происходит обработка данных. Данные подаются в узел через порты ввода. Узлы вывода информации также могут присутствовать, однако есть и такие узлы, в которых порты вывода данных в поток не предусмотрены. Прежде всего, это узлы для построения графиков и диаграмм, а также интерактивные таблицы.

Данная схема представления работы чем-то напоминает предложенную в 70-х годах прошлого века Тони Бьюзен о технологии ментальных карт и, как показывает практика, действительно помогает в восстановлении в памяти процедуры обработки данных даже после длительного периода «простаивания» потока.

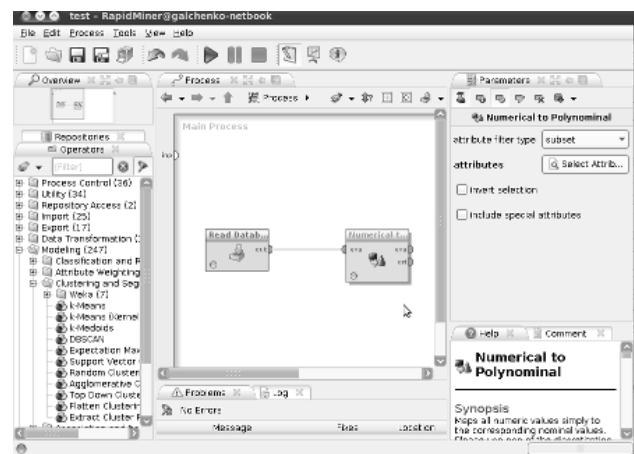


Рис. 2. Интерфейс RapidMiner 5.0

Пакет содержит действительно огромное количество узлов, реализующих те или иные алгоритмы обработки данных, в том числе и алгоритмы text mining. Кроме того, существует возможность использовать в потоках данных конструкции на языке R.

Последний пакет, на котором мы остановимся более подробно, так как именно он стал базой для работы в нашем ВУЗе — KNIME (актуальная версия — 2.2.0). Пакет изначально разрабатывался для решения задач биоинформатики в Konstanz University, причем средой разработки пакета был выбран Eclipse, язык программирования — Java. Обработка данных построена на принципе формирования потоков (рис. 3).

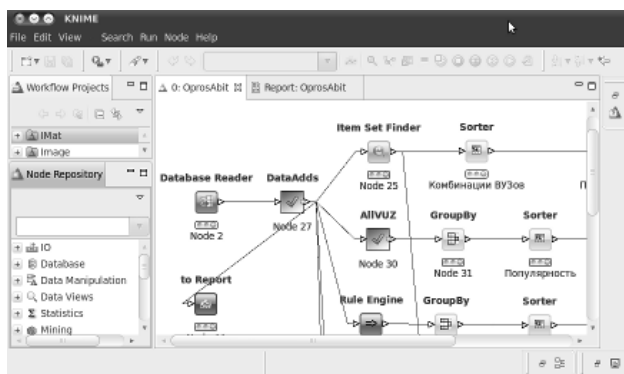


Рис. 3. Интерфейс KNIME 2.2.0

Пакет этот — самый молодой из всех рассматриваемых, его разработка началась в 2003 году, при этом его развитие показывает серьезную положительную динамику: в пакете появился базовый функционал text mining и image mining, возможность использовать в пакете на уровне пользователя, в потоках, программы написанные на языках Java и Python. В последних версиях пакета результат обработки данных может быть передан в мощный дизайнер отчетов (BIRT Reporting Tool). Отчет может быть выведен в формате Word, Excel, HTML и некоторые другие форматы.

Доступно получение данных из БД посредством JDBC и ODBC, а также и из файлов

различного формата. Более того, результат может быть сохранен в БД.

Наконец, следует отметить, что авторы пакета активно поддерживают идею развития пакета посредством создания нового функционала сообществом: действительно, модульная структура позволяет создавать новые узлы и интегрировать их не затрагивая сам пакет.

Все эти особенности, вкупе с положительным опытом использования и определили выбор именно этого пакета в качестве базового.

Пакет может успешно применяться, кроме основных задач, для:

- Препроцессинга
- Преобразования данных в БД, в том числе, традиционно выполняющиеся с помощью конструкций SQL
- Переноса данных из одной СУБД в другую

III. Использование KNIME в обучении студентов

В соответствии с указанным ранее функционалом, KNIME может быть рассмотрен в курсе обучения студентов как средство:

- Анализа данных
- Автоматизации отчетной деятельности
- Работы с БД, без привязки к сложным конструкциям языка SQL и конкретному его диалекту
- Дополнительного изучения языка Java

KNIME был успешно применен в курсе «Компьютерные технологии в науке и образовании», читавшемся на первом курсе магистратуры в 2009-2010 учебном году.

Для исследования студентам была предложена обезличенная выборка по предприяти-

ям и организациям, с указанием официальных результатов их финансовой деятельности. Работа студентов включала исследование функций пакета в части базовой обработки данных (основная статистика, построение сводных таблиц, группировка), возможностей взаимодействия с R, построения графиков и диаграмм как с помощью «родных» узлов KNIME, так и с помощью набора узлов JfreeChart, представляющих доступ к одноименной библиотеки и узла R View организующего удобный доступ к функциям построения графиков языка R, а также решения задач классификации.

Для активной работы студентов в пакете и в домашних условиях, были записаны несколько скринкастов, а также разработаны несколько лабораторных работ. Важно отметить, что пакет KNIME — кроссплатформенный и не требует установки, соответственно у студентов не возникло сложности с установкой и запуском пакета и в операционных системах семейства Windows.

В результате, студенты получили навыки в составлении достаточно больших потоков данных (рис. 4).

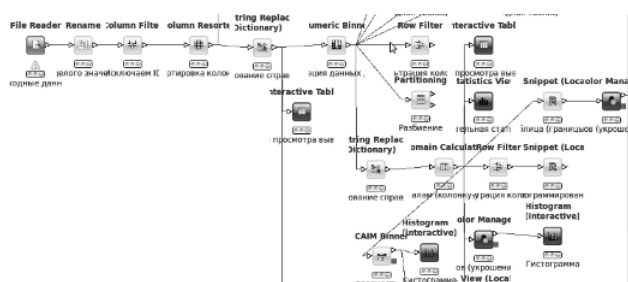


Рис. 4. Фрагмент потока данных, студенческая работа

Более эффективное изучение пакета предполагается достигнуть за счет: использования в работе наборов данных из области будущей профессиональной деятельности студентов, использования пакета в нескольких дисциплинах.

В частности, изучение основ работы с БД, языка SQL, формирования отчетов может

проходить с использованием возможностей KNIME. Как показал опыт применения на занятиях в курсе изучения основ БД утилиты MySQL Workbench, связка MySQL, MySQL Workbench, OOBase и KNIME может успешно вывести из обращения MS Access.

Здесь же, в KNIME, могут быть выполнены курсовые работы, включающие в себя создание узлов с необходимой функциональной нагрузкой.

IV. Использование KNIME в научно-практической деятельности

A. Обработка данных опросов

В 2009-2010 учебном году на Энергетическом факультете был проведен сквозной опрос студентов 1-3 курсов. Цели опроса были сформулированы следующим образом: выявление основных каналов поступления информации о СПбГАУ на момент поступления, их важности, а также оценка восприятия студентами различных аспектов жизни факультета (учебной деятельности, спортивных и культурных мероприятий).

Схема обработки данных опроса показана на рис.5. Как видно из схемы, нам удалось полностью перейти к СПО в обработке полученных данных.

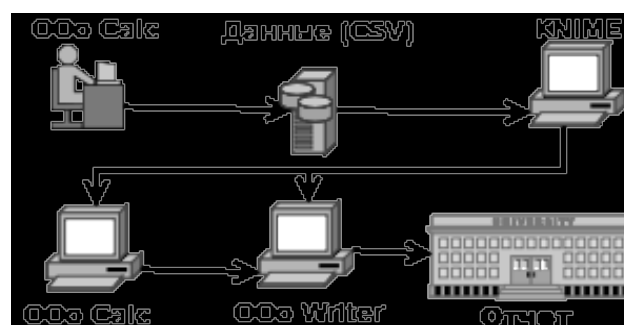


Рис. 5. Обработка информации, полученной в ходе опроса

Использование в обработке на этапе внесения данных редактора таблиц, а не БД, объясняется существенным дефицитом времени

возникшем на этапе сбора и подготовки информации и спонтанностью принятия решения о проведении опроса.

В ходе проведения приемной кампании 2010 года абитуриентам был предложен сходный по существу опросник, который они заполняли во время подачи документов. В этом цикле мы значительно упростили работу оператора и процедуру начальной обработки данных, используя возможности MySQL и OOo Base.

Основным средством обработки данных выступил как в первом, так и во втором случае, KNIME. Мы активно применяли возможности, которые предоставляет данный инструмент: начиная от построения сводных таблиц и заканчивая кластерным анализом и выявлением ассоциативных правил в наборе данных. Были использованы возможности BIRT Reporting, в частности для построения красивых диаграмм, что немаловажно для конечного пользователя.

Из новых возможностей пакета стоит отметить появление набора узлов Item Set, который позволяет вычленять в наборе данных наиболее часто встречающиеся комбинации. Так, в опросе абитуриентов респондентам предлагалось перечислить ВУЗы, в которые они собираются подавать, либо уже подали заявления в порядке приоритетности поступления. Задача выявления наиболее часто встречающихся комбинаций, таким образом, свелась к организации множества значений для каждой записи и применения указанного узла (рис. 6).

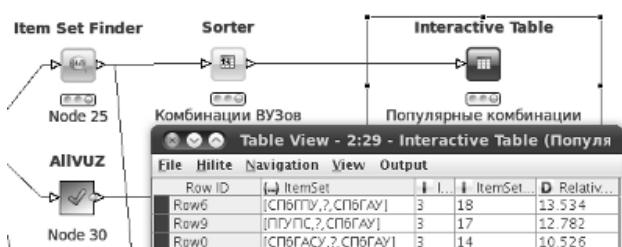


Рис. 6. Применение узла Item Set Finder к задачам опроса абитуриентов

Удобным дополнением стала возможность организации метаузлов — своеобразных аналогов процедур: набора узлов, объединенных по некоторому признаку. Применение метаузлов позволяет повысить читаемость потока и упростить добавление повторяющихся участков. На рис. 6 можно заметить метаузел AllVUZ, в котором агрегированы действия по получению общего списка ВУЗов, указанных абитуриентами, и некоторый набор действий над ним.

Б. Разработка инструментария для анализа надежности в системах «человек-среда-машина»

Изначально отметим, что мы находимся на стадии сбора статистической информации и подготовки к программной реализации задуманного комплекса.

По замыслу, рабочий комплекс должен обеспечить руководство энергетических подразделений компаний информацией о надежности обслуживаемого оборудования, с учетом самого слабого звена системы — человека.

Актуальность проблемы подтвердил проведенный в начале 2010 года экспертный опрос [2]: все данные говорят о том, что основной угрозой для технологической системы является обслуживающий персонал и операторы. В результате обработки данных были выделены основные факторы, влияющие на совершение ошибки обслуживающим персоналом, а также построен и верифицирован индикатор, позволяющий оценивать ситуацию с точки зрения опасности возникновения таких ошибок [3].

Дальнейшая разработка темы идет в двух направлениях: повышение качества построенного индикатора и построение динамической управляемой системы; реализация требуемого функционала в виде программного комплекса, включающего аналитическую систему, базу данных и интерфейс к ней (рис. 7).

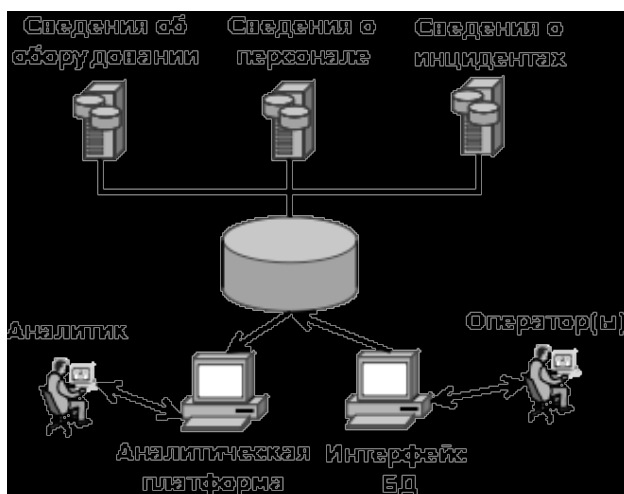


Рис. 7. Общая схема программного комплекса

Исходя из уже имеющейся информации, система должна быть построена на базе современных алгоритмов обработки данных, учитывающих значительную размытость исходной информации.

Система должна быть открыта, интегрируема в уже имеющиеся продукты. Кроме того, мы считаем, что система должна быть кросс-платформенна, хотя бы в силу той причины, что для нужд аналитического работника и оператора БД вполне достаточно рабочего места, укомплектованного исключительно на базе СПО.

На наш взгляд, изложенным выше требованиям идеально соответствует связка из СУБД MySQL, офисного пакета OpenOffice.org и платформы KNIME.

На настоящий момент идет подготовка к реализации алгоритмов, используемых в обработке данных опроса, а также метода анализа иерархий [4] и метода сужения Парето-оптимальных стратегий с учетом информации об относительной важности критериев [5].

V. Заключение

Основной проблемой перехода специализированных подразделений предприятий на

СПО было отсутствие соответствующего ПО и отсутствие подготовленных специалистов.

В сфере аналитической работы в РФ наиболее используются проприетарные пакеты SPSS и Statistica, однако, в настоящий данный класс приложений пополнился такими пакетами, как RapidMiner, R language, KNIME. Данные пакеты имеют достойный функционал, а RapidMiner и KNIME и дружелюбный пользователю графический интерфейс.

Опыт применения KNIME позволяет говорить о его высокой эффективности и рассматривать KNIME как основной пакет для обучения студентов непрофильных специальностей анализу данных (сужение круга специальностей студентов основано лишь на опыте применения пакета в образовательном процессе — пакет разрабатывался для решения задач биоинформатики, есть положительный опыт применения и в маркетинге в зарубежных компаниях).

В практических целях пакет активно использовался в обработке опросов. Опыт применения показал, что пакет содержит весь необходимый функционал. Возможности по дополнению пакета необходимыми функциями позволяют использовать его как базу для построения специализированных систем.

Литература

- [1] П.Ю.Конотопов, Ю.В.Курносков Аналитика: методология, технология и организация информационно-аналитической работы. - М.: РУСАКИ, 2004 г. - 512с.
- [2] А.Г. Гущинский, Н.И. Рузанова, М.И. Гальченко Инструментарий к опросу «неблагоприятные факторы в работе ремонтного и обслуживающего персонала электромеханических установок» // Известия СПбГАУ.- 2009.- №17.
- [3] А.Г. Гущинский, Н.И. Рузанова, М.И. Гальченко Многокритериальная оптимизация процессов в системах обслуживания энергетического оборудования // Известия СПбГАУ.- 2010.- №19.
- [4] Т.Л. Саати Принятие решений при зависимостях и обратных связях// М.-Издательство ЛКИ, 2008 - 360с.
- [5] В.Д. Ногин, Принятие решений при многих критериях.// – СПб. Издательство «ЮТАС», 2007. – 104с.

