УДК 004.891.2

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ОБРАЗОВАТЕЛЬНЫХ ДАННЫХ В СРЕДЕ ПРИЛОЖЕНИЯ RStudio

канд. техн. наук, доц. А.Ф. ОСЬКИН
(Полоцкий государственный университет);
Д.А. ОСЬКИН
(Белорусский государственный экономический университет, Минск)

Описывается разработка программного комплекса для интеллектуального анализа образовательных данных. Обосновывается выбор языка программирования и среды разработки. Описывается структура создаваемого информационно-аналитического комплекса. Приводятся результаты работы одного из модулей комплекса.

***Ключевые слова**: программный комплекс, интеллектуальный анализ образовательных данных, информационно-аналитический комплекс, язык программирования, хранилище данных, OLAP.*

Введение. В течение длительного времени в информационных системах высших учебных заведений накапливались значительные объемы информации, содержащей сведения об успеваемости студентов, их личных данных, увлечениях, участии в жизни университета и т.д. Кроме этого, университетские базы данных хранят информацию о преподавателях, учебных программах, учебных планах, компьютерных тестах, экзаменационных вопросах и задачах, а также другие разнообразные методические и научно-педагогические материалы. До последнего времени эти горы информации оставались невостребованными и по-настоящему неиспользованными. Это положение стало меняться с появлением нового научного направления, получившего название «Интеллектуальный анализ образовательных данных» (Educational Data Mining, EDM) [1].

Как пишут Н.Н. Горлушкина, И.Ю. Коцюба, М.В. Хлопотов: «наиболее характерными из задач интеллектуального анализа образовательных данных являются следующие:

– мониторинг сформированности профессиональных компетенций;
– проектирование учебных планов, индивидуальных учебных планов, программ учебных дисциплин;
– анализ и прогнозирование повышения конкурентоспособности студентов на рынке труда;
– прогнозирование и проектирование тех качеств выпускника, которые предполагается получить «на выходе» образовательного процесса;
– диагностика уровня качества образования для своевременной компенсации нежелательных отклонений;
– оценка реального качества образования на его соответствие стандартам.

Основные цели EDM:

– улучшить образовательный процесс путем поддержки принятия рациональных решений;
– направить студентов по целесообразной образовательной траектории;
– дать рекомендации студентам и преподавателям по корректировке образовательного процесса;
– вникнуть в саму суть учебного процесса – выявить неявные взаимосвязи данных и, как следствие, понять, каким образом человек усваивает информацию, приобретает навыки и умения» [2].

Таким образом, в настоящее время разработка программного комплекса для интеллектуального анализа образовательных данных представляется нам весьма актуальной и важной задачей.

Выбор среды разработки. Известны по крайней мере три языка программирования, которые могут быть эффективно использованы для создания задуманного нами программного комплекса. Это C/C++, Python и R. Мы оценили популярность каждого из этих языков и их пригодность для решения поставленной задачи, анализируя статистику запросов для соответствующих поисковых образов в поисковой системе Яндекс. Частью этой системы является ресурс https://wordstat.yandex.by, позволяющий получить общее число запросов по заданному поисковому образу за текущий месяц. Выяснилось, что с запросом «анализ данных в R» в течение последнего месяца к системе обращались 546 раз (рисунок 1), с аналогичными запросами по языку Python – 372 раза, C/C++ – 59 раз.

Таким образом, на основании этого мини-исследования можно сделать вывод, что на текущий момент наиболее популярным языком для построения программного комплекса интеллектуального анализа образовательных данных является язык R. В принципе, этот вывод не является неожиданным, так как

R – специализированный язык, ориентированный прежде всего на статистический анализ данных. Кроме того, несомненными достоинствами этого языка являются его следующие качества:

– бесплатность. R – это свободное программное обеспечение, его код открыт;

– расширяемость. Язык имеет модульную структуру и состоит из набора пакетов. При этом вновь создаваемые пакеты легко интегрируются в единую программную среду;

– кроссплатформенность и переносимость. Интерпретаторы языка R разработаны для всех популярных операционных систем и перенос кода из одной системы в другую не представляет никаких трудностей;

– великолепная графика, позволяющая выполнять визуальный анализ данных и создавать прекрасно иллюстрированные отчеты по выполненным исследованиям;

– огромное количество встроенных функций для проведения анализа данных;

– встроенные статистические тесты и алгоритмы.

Учитывая сказанное выше, мы остановили свой выбор на языке программирования R, используя в качестве среды разработки приложение RStudio.

Что искали со словом «анализ данных в r» — 546 показов в месяц

Статистика по словам	Показов в месяц
анализ данных +в r	273
язык r анализ данных	100
r анализ +и визуализация данных	92
анализ данных +с помощью r	57
r studio анализ данных	18
анализ данных +в r stepic ответы	13
r +или python +для анализа данных	9
разведочный анализ данных r	8
search rsl ru анализ данных r	6
анализ панельных данных +в r	6

Рисунок 1. – Число запросов в Яндекс для поискового образа «анализ данных в R»

Структура программного комплекса. Анализ образовательных данных, проводимый с использованием технологий EDM, состоит из четырех этапов [3].

1. *Построение хранилища данных*.

Основой аналитической системы является хранилище данных. Оно должно актуализироваться с определенной периодичностью, пополняясь данными из баз, результатами контрольных мероприятий, итогами сессий и т.д. Данные проверяются, очищаются, проходят предварительную обработку, приводятся к единому формату и загружаются в хранилище данных.

2. *Построение многомерного OLAP-куба*.

Многомерный OLAP-куб строится на основе хранилища данных. С его помощью становится возможным осуществлять в режиме реального времени анализ данных и формировать отчеты в различных аспектах с произвольной глубиной детализации.

3. *Формирование системы ключевых показателей*.

На основе многомерных OLAP-кубов может быть сформирована система ключевых показателей, позволяющая проводить мониторинг и оценку бизнес-процессов и информировать администраторов об имеющих место фактах отклонения.

4. *EDM – интеллектуальный анализ образовательных данных*.

На основе данных, загруженных в хранилище, могут быть построены модели интеллектуального анализа, позволяющие реализовать процедуры прогнозирования наиболее важных показателей учебной деятельности, а также выявить ее скрытые и неочевидные закономерности.

В соответствии с описанной выше структурой программный комплекс должен включать в себя следующие модули (рисунок 2).

1. *Модуль хранения данных*. Данный модуль представляет собой хранилище данных, дополненное инструментами для интеграции с различными внешними источниками, такими как данные, хранящиеся в БД «Деканат», таблицы Excel, данные произвольных форматов из сторонних информационных систем. Здесь же будет выполняться предварительная обработка данных, их очистка и приведение к единому формату.

2. *Модуль анализа данных*. Функционал этого модуля должен позволять формировать многоаспектные отчеты на базе OLAP-технологий. Предполагается, что будет также реализована возможность графического представления результатов анализа в виде столбиковых, круговых, кольцевых и других диаграмм.
3. *Модуль прогнозирования*. В этом модуле будут реализованы известные методы Data Mining, такие как классификация, кластеризация, ассоциация, прогнозирование, анализ отклонений.
4. *Модуль администрирования*, содержащий функции управления приложением.

Рисунок 2. – Структура информационно-аналитической системы

Реализация комплекса в IDE RStudio. Как уже указывалось выше, работы по созданию программного комплекса велись в интегрированной среде разработки (IDE) RStudio. Это бесплатное программное обеспечение, позволяющее выполнить весь цикл работ по проектированию, созданию, отладке, тестированию и эксплуатации программных продуктов, написанных на языке R.

После запуска RStudio, перед разработчиком открывается четыре окна (рисунок 3).

Окно «Редактор» представляет собой удобный редактор кода, имеющий ряд опций, существенно повышающих продуктивность работы. Это, в частности, автоматическое завершение кода, подсветка кода, возможность одновременного редактирования нескольких файлов, поиск и замена выделенных частей кода. Интересной особенностью редактора является возможность выделить часть кода, проанализировать его и автоматически преобразовать в функцию для последующего многократного использования.

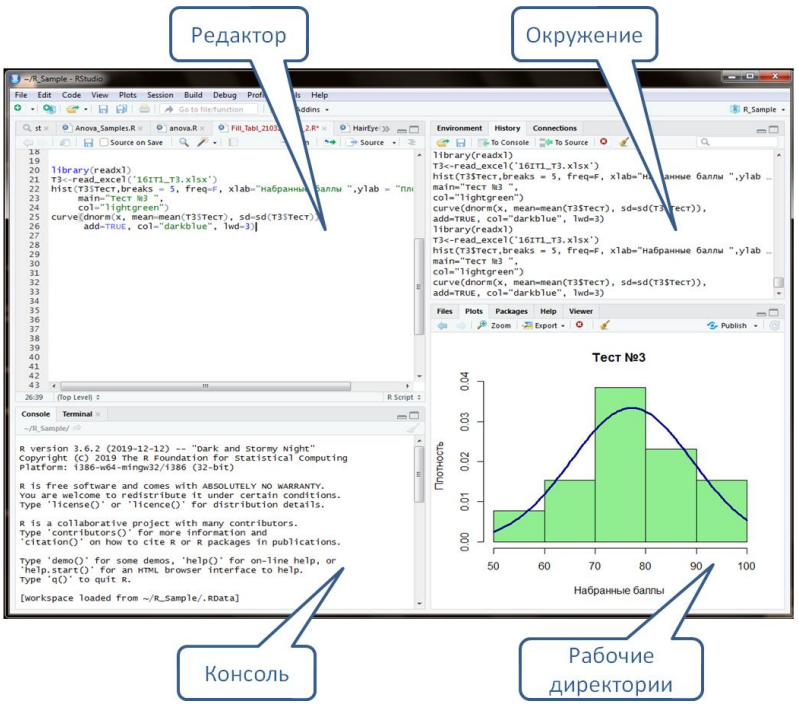

Рисунок 3. – Интегрированная среда разработки RStudio

Имеется также возможность исполнения кода непосредственно из редактора, без ручного копирования кода в консоль.

Кроме создания и редактирования кода в редакторе RStudio можно создавать и редактировать файлы документов, используя язык R Markdown, с последующим экспортом в форматы TeX или pdf. Эти файлы могут также содержать фрагменты исполняемого кода. Все это позволяет писать тезисы докладов, научные статьи, отчеты, другие документы, содержащие R-программы, прямо в редакторе, не тратя усилий на их последующее оформление.

Консоль RStudio представляет собой собственно среду для исполнения скриптов, написанных на языке R. Так же, как и в редакторе в консоли имеется опция автоматического завершения кода, что повышает продуктивность разработки и отладки программ.

Так как при работе с R постоянно возникает необходимость повторить выполненную ранее команду, в консоли Rstudio, как и в стандартной R-консоли, поддерживается функция навигации по ранее выполненным командам. Для этого используются клавиши со стрелками.

Окно «Окружение» содержит сведения о текущем состоянии рабочего окружения и позволяет просматривать векторы, матрицы, списки, датафреймы, доступные для использования в текущей сессии.

Кроме того, можно просматривать историю выполненных ранее программ, осуществлять поиск по истории и выполнять найденные команды.

Окно «Рабочие директории» позволяет просматривать списки файлов, созданных в процессе реализации проекта и манипулировать ими.

Таким образом, RStudio представляет собой гибкий и удобный инструмент, позволяющий вести разработку программного обеспечения с максимальной эффективностью.

По состоянию на начало марта 2020 года нами реализованы следующие пакеты, входящие в состав перечисленных выше модулей:

– пакет Load_IREN, предназначенный для извлечения, преобразования и загрузки данных из базы данных системы online-тестирования IREN [4] в хранилище разрабатываемого программного комплекса;
– пакет OLAP_Cube_Builder, строящий OLAP-куб на основе данных из хранилища;
– пакет Make_Report, ответственный за формирование отчетов;
– пакет Make_Visualization, предназначенный для визуализации результатов анализа.

На рисунках 4 и 5 показаны результаты работы пакета Make_Visualization. Это результаты прохождения тестов № 1 и № 3 по дисциплине «Методы и алгоритмы принятия решений» студентами одной из групп дневной формы обучения. Видно, что первый тест пройден гораздо хуже третьего. Это можно объяснить той методикой, которая использовалась нами при проведении тестирования знаний студентов. После изучения первой лекции студенты сдавали тест по материалам только этой лекции. После второй тест включал в себя вопросы как из первой, так и из второй лекций, после третьей – вопросы из первой, второй и третьей лекций и т.д. Таким образом, постоянное повторение вопросов способствовало лучшему усвоению материала и, как следствие, получению лучших результатов при тестировании. Все сказанное хорошо иллюстрируется гистограммами, приведенными на рисунках 4 и 5.

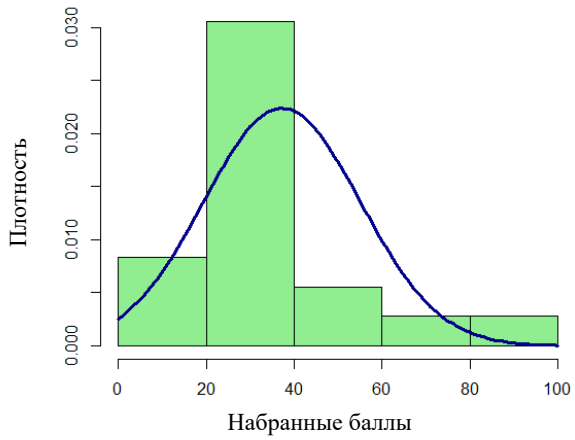

Рисунок 4. – Результаты теста № 1

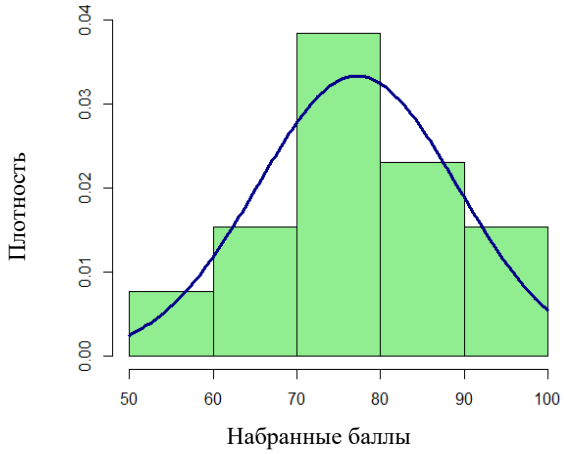

Рисунок 5. – Результаты теста № 3

Выводы:

1. Интеллектуальный анализ образовательных данных – перспективное и динамично развивающееся научное направление, результаты внедрения которого могут существенно изменить подходы к управлению учебным процессом.

2. Разработка программного обеспечения для интеллектуального анализа образовательных данных является важной и актуальной задачей.

3. Для быстрого построения систем интеллектуального анализа данных может быть весьма эффективно использован язык программирования R.

4. Для разработки, отладки и эксплуатации программ на R удобно использовать интегрированную среду разработки RStudio.

ЛИТЕРАТУРА

1. Romero, C. Data mining in education / C. Romero, S. Ventura // Wiley interdisciplinary reviews. Data mining and knowledge discovery. – 2013. – 3(1). – P. 12–27.
2. Горлушкина, Н.Н. Задачи и методы интеллектуального анализа образовательных данных для поддержки принятия решений / Н.Н. Горлушкина, И.Ю. Коцюба, М.В. Хлопотов // Образовательные технологии и общество. – 2015. – Т. 18, № 1. – С. 474.
3. KAI Development. IT решения для бизнеса и государства [Электронный ресурс]. – Режим доступа: http://kaidev.ru. – Дата доступа: 14.03.2020.
4. Айрен. Программа тестирования знаний [Электронный ресурс]. – Режим доступа: https://irenproject.ru. – Дата доступа: 14.03.2020.

Поступила 17.03.2020

SOFTWARE PACKAGE FOR INTELLECTUAL ANALYSIS OF EDUCATIONAL DATA IN THE ENVIRONMENT OF THE RSTUDIO APPLICATION

A. OSKIN, D. OSKIN

Describes the development of a software package for the educational data mining. The choice of a programming language and development environment is substantiated. The structure of the created information-analytical complex is described. The results of one of the modules of the complex are presented.

Keywords: *software complex, educational data mining, information and analytical complex, programming language, data warehouse, OLAP.*