ПОДХОДЫ К ОПРЕДЕЛЕНИЮ СРЕДСТВ ДЛЯ ПОСТРОЕНИЯ МЕТОДИКИ ОБУЧЕНИЯ РАБОТЕ С БОЛЬШИМИ ДАННЫМИ

А. Г. Степанов[1], **Г. А. Плотников**[1], **В. С. Васильева**[1]

[1] ***Санкт-Петербургский государственный университет аэрокосмического приборостроения***
190000, Россия, г. Санкт-Петербург, ул. Большая Морская, д. 67

Аннотация

В статье актуализируется необходимость обучения студентов технологиям работы с большими данными. Big data — это перспективная и фундаментальная отрасль, требующая большого количества квалифицированных специалистов в различных областях. Целью работы является описание концепции определения набора аппаратных, программных, алгоритмических и методических средств (с учетом контингента обучаемых и возможностей учебного заведения) для построения методики преподавания дисциплины, связанной с изучением методов обработки больших данных. Выделяются два основных сектора заинтересованных лиц, которым необходимы специалисты в сфере Big Data. Производится сравнительный анализ программных решений, поддерживающих обработку больших данных. Описывается методика построения курса по обучению студентов технологиям обработки и анализа больших данных. Предлагается план организации лекционного курса и лабораторного практикума с рассмотрением подзадач для выполнения во время обучения. Обговариваются состав и методика самостоятельной работы студентов по дисциплине, связанной с изучением Big Data, с использованием системы управления обучением типа Moodle. Представляется пример реализации обработки данных средствами пакета RapidMiner Studio с использованием алгоритма обучения многослойной нейронной сети методом обратного распространения ошибки.

Ключевые слова: RapidMiner Studio, Moodle, Data Mining, нейронная сеть, классификация, кластеризация, регрессия, поиск ассоциативных правил, обработка естественного языка, NLP.

DOI: 10.32517/0234-0453-2021-36-4-54-62

Для цитирования:

Степанов А. Г., Плотников Г. А., Васильева В. С. Подходы к определению средств для построения методики обучения работе с большими данными // Информатика и образование. 2021. № 4. С. 54–62.

Статья поступила в редакцию: 28 января 2021 года.

Статья принята к печати: 6 апреля 2021 года.

Сведения об авторах

Степанов Александр Георгиевич, доктор пед. наук, доцент, профессор кафедры информационных технологий предпринимательства, Институт технологий предпринимательства, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия; georgich_spb@mail.ru; ORCID: 0000-0002-3922-9684

Плотников Григорий Александрович, ассистент кафедры информационных технологий предпринимательства, Институт технологий предпринимательства, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия; dim111077@mail.ru; ORCID: 0000-0002-8924-1382

Васильева Виктория Сергеевна, магистрант кафедры информационных технологий предпринимательства, Институт технологий предпринимательства, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, Россия; victoria-vasilyeva@mail.ru; ORCID: 0000-0002-1429-0358

1. Введение

Удивительные результаты применения современных информационных технологий, использующих методы обработки больших объемов данных (Big Data), производят на непосвященных ошеломляющее впечатление. Ездящие без водителя автомобили и разговаривающие человеческим языком роботы еще вчера были чем-то из области фантастики. Сегодня это предмет для публикаций газет уже даже не на первых полосах, а завтра они станут привычными и общеупотребительными. Мы должны быть готовы к тому, что создание и применение технологий Big Data станут, в свою очередь, массовой сферой приложения человеческого труда. И, как следствие, обязаны уже сегодня учить своих студентов вопросам проектирования, конструирования и эксплуатации этих технологий.

Достигнутые результаты были бы невозможны без взрывоподобного развития современной информатики. Опираясь на понятие информации как на некий феномен, принимаемый без определения [1], информатика в своей основе базируется на трех своих связанных с компьютерами составляющих:

- аппаратных средствах (*hardware*);
- программных средствах (*software*);
- методах, моделях и алгоритмах (*brainware*).

Развитие *аппаратных средств* на сегодняшний день во многом сдерживается физическими ограничениями, однако современный уровень этих средств обеспечивает решение большинства актуальных практических задач.

Программные средства непрерывно развиваются в сторону повышения производительности труда программистов. На сегодняшний день эти средства представлены широкой номенклатурой трансляторов, компиляторов и интерпретаторов, позволяющих реализовать методы процедурно-ориентированного, структурного, объектно-ориентированного, логико-ориентированного программирования и визуального конструирования.

Продолжаются развитие *методов* решения практических задач и уточнение *моделей* (в первую оче-

редь математических) реальных процессов, а также их *алгоритмизация* в виде создания специализированных библиотек и программных средств.

Все перечисленное говорит о необходимости адаптации возможностей педагогики к решению современных задач информатики.

Одними из ведущих направлений исследований в области информатики являются вопросы работы с данными больших объемов. Ограничения в емкости устройств хранения информации сейчас уже не играют определяющей роли. Развитие направления во многом сдерживается временны́ми характеристиками вычислителей и требует внедрения в практику многопроцессорных распределенных систем. Как следствие, возникает необходимость развития специальных методов обработки информации, учитывающих эти обстоятельства.

Целью настоящей статьи является *описание концепции определения набора аппаратных, программных, алгоритмических и методических средств (с учетом контингента обучаемых и возможностей учебного заведения) для построения методики преподавания дисциплины, связанной с изучением методов обработки больших данных.*

2. Концепция определения набора аппаратных средств и выбора программных и алгоритмических средств для построения дисциплины, связанной с обработкой больших данных

Круг лиц, заинтересованных в изучении методов обработки больших данных, весьма широк и неоднороден. В первую очередь к нему можно отнести представителей всей информационной отрасли современной экономики. Р. М. Юсупов разделяет ее на первичный и вторичный информационные сектора [2].

В первичном информационном секторе сосредоточено производство информации, информационных услуг и информационных средств. Представители этого сектора экономики являются профессионалами в своем деле и используют для решения своих задач всю номенклатуру инструментария информатики.

Вторичный сектор информационной отрасли входит в состав других структур экономики, связанных с материальным производством, производством энергии, обеспечением жизнедеятельности человека, транспортом, военным делом и т. п. Информационные технологии здесь играют вспомогательную по отношению к основной производственной деятельности роль [3, 4]. Тем не менее именно тут и ожидается наибольший прогресс в части использования средств и методов обработки больших данных, поскольку именно здесь они могут оказать существенное влияние на эффективность работы экономики в целом. Кроме этого к заинтересованным в освоении обсуждаемых информационных технологий можно отнести студентов и аспирантов самых разных областей знаний. Информационная подготовка уже стала неотъемлемой частью системы высшего образования, но именно подрастающее поколение специалистов должно и будет внедрять сегодняшние достижения информатики в практическую деятельность.

Оставим профессионалам из первого сектора информационной отрасли их профессиональные заботы и сосредоточимся на методике подготовки тех, для кого она не является профильной. Очевидно, что такая подготовка выдвигает более жесткие ресурсные и организационные ограничения. В этом случае требуется гораздо более строгий подход к выбору инструментальных средств.

Распространенным средством программирования обсуждаемого класса задач является язык Python. Он представляет собой универсальный язык программирования, реализованный в большинстве случаев как интерпретатор. На практике его использование требует владения навыками программирования на процедурно-ориентированных языках высокого уровня. Важным достоинством языка является возможность подключения к нему специализированных библиотек, однако доступные учебники, например [5], сосредоточиваются только на достаточно сложном синтаксисе Python. В то же время основной интерес представляет не столько сама технология программирования, сколько возможность освоения обучающимися новых для них категорий, относящихся непосредственно к интеллектуальному анализу: классификация, регрессия, кластеризация и т. п. Хотелось бы при обучении освободиться от традиционных методов программирования со всеми их недостатками и сложностями и использовать *специализированные программные средства визуального проектирования и программирования*. В таблице представлены некоторые из них.

Платформа **IBM SPSS Modeler**, как следует из ее названия, — это продукт основного игрока на рынке программных средств — компании IBM [6]. Утверждается, что она обладает низким порогом входа для начинающих. Доступность для новичков обеспечивается режимами «автопилота». Встроенные средства автоматизации (Auto Numeric, Auto Classifier) перебирают несколько возможных моделей с разными параметрами, определяя среди них лучшую в условиях конкретной задачи. Как следствие, не слишком опытный аналитик может построить на таком инструменте адекватную модель.

KNIME — еще одна бесплатная система для интеллектуального анализа данных, которая даже в базовой версии обладает хорошим функционалом [7]. Программное средство предлагает интуитивно понятную рабочую среду без необходимости программировать. Пользователю предоставляется набор типовых операторов, позволяющих решать стандартные задачи (в KNIME они называются узлами).

Qlik Analytics Platform предоставляет разработчикам все необходимые инструменты для управления данными [8]. Компания Qlik является лидером в области визуальной аналитики, а ее платформа для анализа данных поддерживает создание и разработку как пользовательских, так и заказных аналитических приложений, в том числе мэшапов

Таблица

Программные средства, обеспечивающие поддержку технологий обработки больших данных

Название	Графическое программирование	Процедурное программирование	Обработка текста	Регрессия	Классификация	Кластеризация	Нейронные сети	Деревья решений	Создание циклов	Облачное хранение	Help-система		
											Встроенные примеры	Подсказки	Уроки
IBM SPSS Modeler	Да	Нет	Платно	Да	Да	Да	Да	Да	Да	Нет	Нет	Да	Да
KNIME	Да	Да	Да	Да	Да	Да	Да	Да	Да	Да	Нет	Да	Да
Qlik Analytics Platform	Да	Нет	Да	Да	Да	Да	Да	Да	Да	Да	Нет	Да	Да
STATISTICA Data Miner	Да	Нет	Отдельно	Да	Да	Да	Да	Да	Нет	Нет	Нет	Да	Да
RapidMiner Studio	Да	Да	Да	Да	Да	Да	Да	Да	Да	Да	Да	Да	Да

(мэшап — веб-приложение, объединяющее данные из нескольких источников в один интегрированный инструмент).

STATISTICA Data Miner — это платформа российской разработки. Система предоставляет наиболее полный набор методов для Data Mining [9]. В частности, в STATISTICA Data Miner реализованы инструменты предварительной обработки, фильтрации и чистки данных, что позволяет эффективно отбирать признаки из сотен тысяч возможных предикторов.

Наконец, известен и свободно распространяется в минимальной, но достаточной для начального обучения версии программный пакет **RapidMiner Studio** [10]. Мы пришли к заключению, что все перечисленные программы имеют схожий функционал, в некотором смысле копируют друг друга, и остановили свой выбор именно на RapidMiner Studio. Он имеет англоязычный сайт — источник основной документации — и развитую внутреннюю систему помощи. Как следствие, при работе с этим пакетом пользователю можно достаточно просто найти всю необходимую информацию*.

Наличие в лицензированных вузах электронной образовательной среды обязательно предусматривает возможность выхода в интернет. Скорее всего, высшее учебное заведение имеет дисплейные классы, объединенные внутренней локальной сетью. Состав аппаратуры дисплейного класса может быть самым разным, однако с высокой вероятностью можно предположить, что его компьютеры удовлетворяют следующим **минимальным требованиям пакета RapidMiner Studio** [11]:

- двухъядерный процессор 2 ГГц;
- 4 ГБ оперативной памяти;
- больше 1 ГБ свободного места на диске;
- разрешение дисплея 1280×1024;
- операционная система Windows 7, Windows 8, Windows 8.1, Windows 10 (производитель настоятельно рекомендует 64-разрядную версию), Linux (только 64-разрядная версия), MacOS X 10.10–10.14.

Традиционное планирование учебного процесса предусматривает семестровую организацию обучения и нормирование его объема в академических часах. Нами была взята за основу следующая **структура курса** [12]:

- лекционные занятия — 16 академических часов;
- лабораторный практикум — 16 часов;
- самостоятельная работа студента под руководством преподавателя и промежуточная аттестация в виде экзамена в форме компьютерного тестирования — 144 часа.

3. Содержание обучения и особенности преподавания дисциплины

Информация может существовать, анализироваться и обрабатываться в следующих формах: данные, знания, смысл и чувства (их эмоциональность или тональность). Составляя представленное на рисунке 1 содержание дисциплины, мы попытались затронуть все перечисленные составляющие.

Теоретическая часть дисциплины поддерживается лекционным курсом, который может быть реализован как в традиционном, так и в дистанционном вариантах [13].

Подготовка данных подразумевает поиск, сбор, кодирование, хранение, сортировку и фильтрацию данных, а также выделение пропущенных или ошибочных значений. Она может быть проведена, например, с помощью стандартных функций Excel или с использованием возможностей баз данных. Новыми для студента оказываются освоение операции импорта данных, сведения об используемых в пакете форматах данных и методы подготовки и отображения данных самого пакета. Лабораторные работы Лр1 — Лр3 формируют необходимые навыки в соответствии с особенностями решаемых

* При составлении таблицы мы испытывали определенные трудности с заполнением некоторых позиций.

Рис. 1. Содержание дисциплины

задач. Назначение лабораторной работы Лр4 — обеспечить обучаемого материалом для последующих исследований.

Под знаниями обычно понимают законы и закономерности [14]. Очевидно, что выделение знаний из имеющихся наборов данных представляет собой типовую задачу Data Mining, которая решается методами классификации, регрессии, кластеризации и поиска ассоциативных правил. На обучение использованию этих методов (постановке задачи и интерпретации результатов) направлено выполнение лабораторных работ Лр5 — Лр10.

Понятие смысла требует уточнения. Будем считать, что смысл — это «идеальное содержание, идея, сущность, предназначение, конечная цель (ценность) чего-либо (смысл жизни, смысл истории и т. д.); целостное содержание какого-либо высказывания, несводимое к значениям составляющих его частей и элементов, но само определяющее эти значения (например, смысл художественного произведения и т. п.); в логике, в ряде случаев в языкознании — то же, что значение» [15]. Выделение смысла в процессе компьютерной обработки информации является перспективным направлением, но на настоящий момент находится только в стадии разработки, начальной постановки и получения первых практических результатов. Подобного рода задачи решаются методами Natural Language Processing (обработка естественного языка) и в литературе обычно обозначаются аббревиатурой NLP [16]. Особый интерес сегодня представляют методы обработки естественного языка, в частности выделения смысла и его тональности [17, 18]. Формированию общего подхода к решению подобных задач посвящена лабораторная работа Лр11.

Лабораторный практикум, при необходимости, также может быть реализован в дистанционном формате. Такой вариант его проведения потребует наличия у обучаемых собственной аппаратной базы и установки на нее программного продукта RapidMiner Studio.

В рамках одной сравнительно небольшой дисциплины изучить все возможности пакета нельзя. Поэтому студентам предоставляется возможность ознакомиться с избранными методами в зависимости от их профессиональных интересов. Организационно в план выполнения лабораторных работ включаются:

- обязательное выполнение лабораторных работ Лр1 — Лр3;
- при отсутствии у обучаемого собственных статистических данных — лабораторная работа Лр4;
- лабораторная работа ЛрN по выбору обучающегося из числа работ Лр5 — Лр12 в зависимости от тематики его научных исследований.

4. Организация самостоятельной работы обучаемых

Самостоятельная работа студентов при изучении рассматриваемой дисциплины включает деятельность нескольких видов — репродуктивную, познавательно-поисковую, творческую (рис. 2), — которая осуществляется по заданию преподавателя и при его методическом руководстве [19].

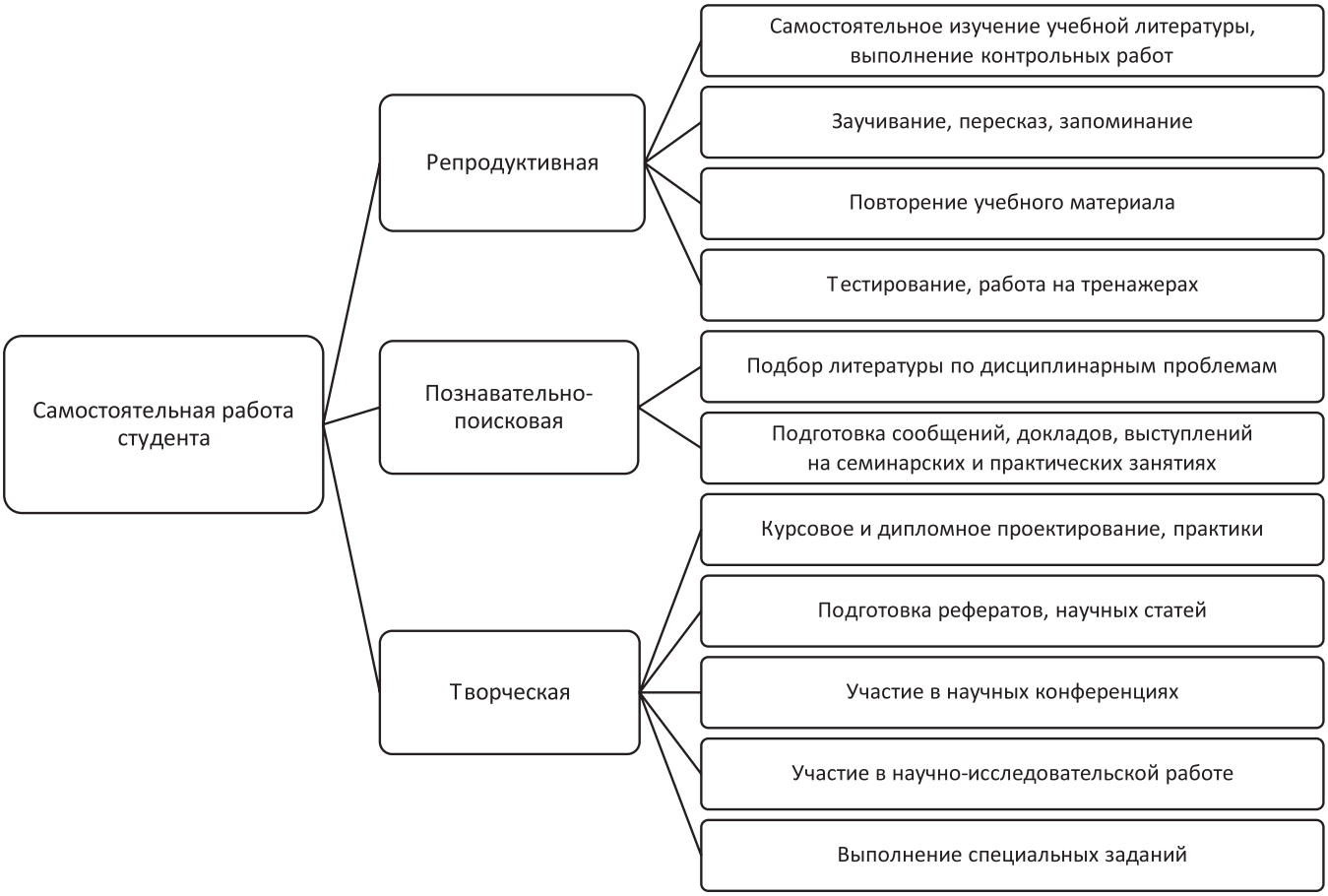

Рис. 2. Возможные виды самостоятельной работы обучаемого

Учебный процесс по дисциплине целесообразно поддерживать системой управления обучением типа Moodle [20]. Для этого в системе должна быть создана страница дисциплины, на которой размещаются рабочая программа дисциплины, презентация к лекциям, методические указания для самостоятельной работы, тесты и позиции, в которые обучающийся должен присылать преподавателю результаты выполнения самостоятельной работы. При дистанционном формате работы обучающихся на странице дисциплины должны быть установлены средства организации телеконференции. Промежуточная аттестация может выполняться либо в традиционном формате, либо в форме компьютерного теста.

5. Пример реализации обработки данных средствами пакета RapidMiner Studio

Рассмотрим вариант реализации алгоритма классификации с использованием многослойной нейронной сети персептронного типа методом обратного распространения ошибки.

Сеть должна распознать известные ей четыре символа: У, Б, 2, 3 (рис. 3). Как следствие, она имеет четыре выхода.

Начальная подготовка данных проводилась средствами Excel. Закрашенный пиксель отображается символом 1, а незакрашенный — символом 0 (рис. 4).

Далее исходные данные импортированы в репозиторий пакета RapidMiner Studio и им присвоено название «Обучение нейронной сети», а атрибуту «ОБРАЗЕЦ» обучающей выборки задана роль label.

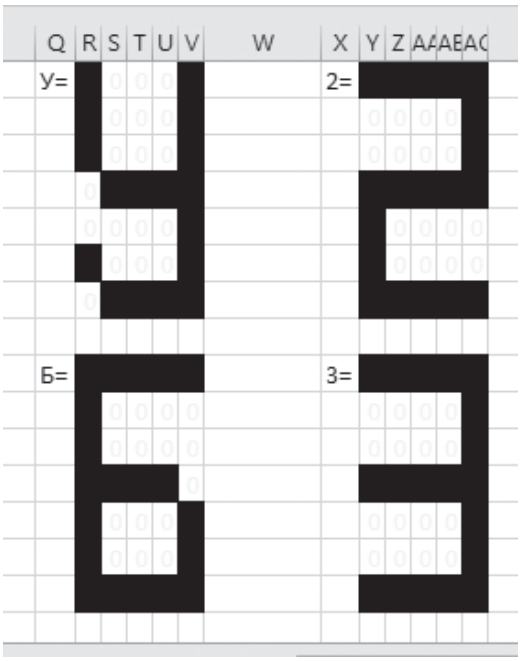

Рис. 3. Графическое представление символов

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	ОБАЗЕЦ	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	X33	X34	X35	
2	2	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	
3	3	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	
4	У	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	0	1	1	1	1	0	0	0	0	1	1	0	0	0	1	0	1	1	1	1	
5	Б	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	1	0	0	0	1	1	0	0	0	1	1	1	1	1	0	
6																																					

Рис. 4. Представление исходных данных (обучающая выборка)

Import Data - Format your columns.

Format your columns.

Replace errors with missing values

	ОБАЗЕЦ *polynominal label*	X1 *integer*	X2 *integer*	X3 *integer*	X4 *integer*	X5 *integer*	X6 *integer*	X7 *integer*
1	2	1	1	1	1	1	0	0
2	3	1	1	1	1	1	0	0
3	У	1	0	0	0	1	1	0
4	Б	1	1	1	1	1	1	0
5	1	0	0	0	0	1	0	0
6	8	1	1	1	1	1	1	0
7	Ч	1	0	0	0	1	1	0
8	В	1	1	1	1	0	1	0

PREVIEW

Рис. 5. Данные распознаваемой выборки (тестирование нейронной сети)

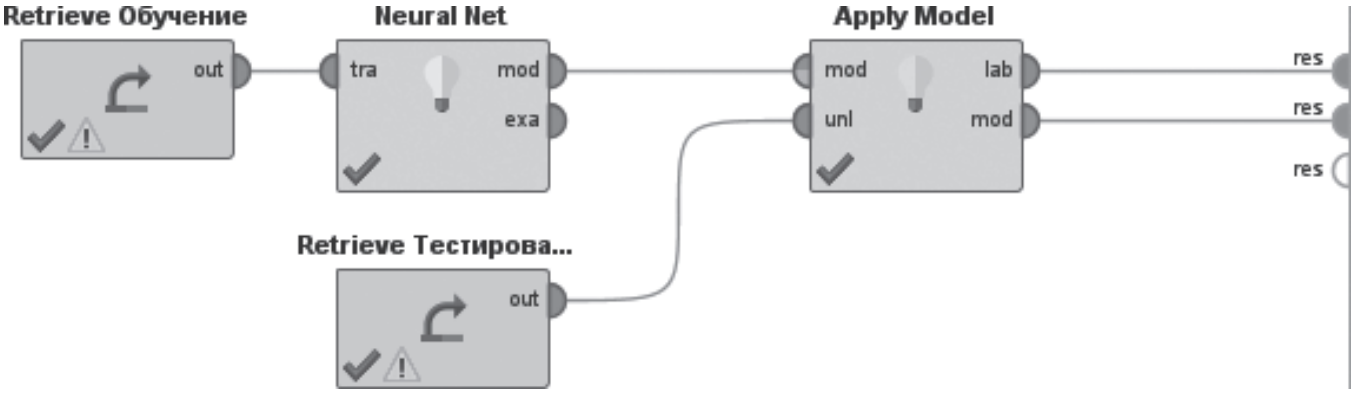

Рис. 6. Процесс классификации с помощью нейронной сети

Набор данных распознаваемой (классифицируемой) выборки называется «Тестирование нейронной сети» и представлен на рисунке 5. Он включает в себя четыре символа, полностью совпадающих с символами обучающей выборки, а также четыре символа, отличных от заданных.

Средствами RapidMiner Studio с помощью оператора Neural Net построен и выполнен процесс классификации (рис. 6). На рисунке 7 представлены настройки этого оператора.

Оператор Apply Model используется для объединения обученной сети и новых данных. Результаты работы процесса показаны на рисунке 8. Колонка predication представляет собой результат классификации, а колонки confidence показывают оценку совпадения символа с соответствующим классом.

Сеть «правильно» распознала символы У, Б, 2, 3, заданные в обучающей выборке (рис. 9), а прове-

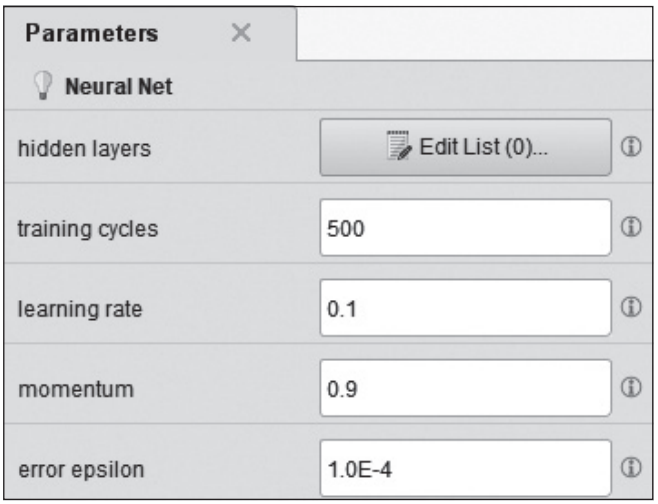

Рис. 7. Настройки оператора Neural Net

Row No.	ОБАЗЕЦ	prediction(ОБАЗЕЦ)	confidence(2)	confidence(3)	confidence(У)	confidence(Б)
1	2	2	0.966	0.013	0.008	0.012
2	3	3	0.013	0.961	0.015	0.012
3	У	У	0.009	0.011	0.971	0.009
4	Б	Б	0.012	0.011	0.009	0.967
5	1	У	0.013	0.476	0.511	0.001
6	8	Б	0.227	0.090	0.131	0.552
7	Ч	У	0.002	0.115	0.873	0.009
8	В	Б	0.024	0.016	0.018	0.942

Рис. 8. Результаты классификации

Row No.	ОБАЗЕЦ	predicti...	confidence(2) ↑	confidence(3)	confidence(У)	confidence(Б)
3	У	У	0.009	0.011	0.971	0.009
4	Б	Б	0.012	0.011	0.009	0.967
2	3	3	0.013	0.961	0.015	0.012
1	2	2	0.966	0.013	0.008	0.012

Рис. 9. Примеры «правильной» классификации

Row No.	ОБАЗЕЦ	predicti...	confidence(2)	confidence(3)	confidence(У)	confidence(Б)
1	1	У	0.013	0.476	0.511	0.001
2	8	Б	0.227	0.090	0.131	0.552
3	Ч	У	0.002	0.115	0.873	0.009
4	В	Б	0.024	0.016	0.018	0.942

Рис. 10. Примеры «неправильной» классификации

рочные символы 1, 8, Ч и Б были классифицированы «неправильно», но отнесены к наиболее близкому классу (рис. 10).

Таким образом, построенный процесс может быть использован для классификации символов.

6. Выводы

Подводя итоги изложенному, отметим, что:

- актуальность преподавания вопросов обработки больших данных определяется потребностями развития экономики и наличием соответствующих программных средств;
- в высших учебных заведениях существует вся необходимая инфраструктура для создания учебной дисциплины, связанной с обработкой больших данных;
- определены структура и содержание обучения по дисциплине с учетом возможности реализации ее изучения в рамках магистерской подготовки или системы повышения квалификации как в традиционном, так и в дистанционном форматах;
- разработана технология организации самостоятельной работы обучающихся;
- показана возможность практического решения задач Data Mining средствами пакета RapidMiner Studio.

Список использованных источников

1. *Юсупов Р. М., Юсупов Ю. В.* Состояние и перспективы развития информатики // Труды СПИИРАН. 2007. № 5. С. 10–45. http://www.mathnet.ru/links/79ca198018afb2fc5c21394ee2c7b0ce/trspy297.pdf

2. *Юсупов Р. М.* Информационные технологии и экономика информационного общества // Инновации. 2013. № 11. С. 40–46. https://www.elibrary.ru/item.asp?id=22306752

3. *Voronin D., Shevchenko V., Chengar O., Mashchenko E.* Conceptual Big Data processing model for the tasks of Smart Cities environmental monitoring // DTGS 2019: Digital Transformation and Global Society. Communications in Computer and Information Science. Cham: Springer, 2019. Vol. 1038. P. 212–222. DOI: 10.1007/978-3-030-37858-5_17

4. *Sheng H., Wang X.* Intelligent retrieval system for ship fault information based on big data analysis // Journal of Coastal Research. 2019. Vol. 93. Is. 1. P. 1019–1025. DOI: 10.2112/SI93-147.1

5. *Златопольский Д. М.* Основы программирования на языке Python. М.: ДМК Пресс, 2018. 396 с.

6. IBM SPSS Modeler. https://www.ibm.com/ru-ru/products/spss-modeler

7. End to End Data Science // KNIME. https://www.knime.com/

8. Qlik Analytics Platform. https://www.qlik.com/us/products/qlik-analytics-platform

9. Уникальные возможности STATISTICA Data Miner. http://statsoft.ru/products/STATISTICA_Data_Miner/unique-opportunities-statistica-data-miner.php

10. RapidMiner. https://rapidminer.com/

11. RapidMiner Documentation. https://docs.rapidminer.com/

12. *Stepanov A. G., Plotnikov G. A., Kosmachev V. M.* Assurance of safety of computer-based testing system for midterm knowledge assessment // The European Proceedings of Social and Behavioural Sciences. 2020. Vol. 90. P. 1473–1485. DOI: 10.15405/epsbs.2020.10.03.170

13. *Космачев В. М., Плотников Г. А., Степанов А. Г.* Преподаватель в дистанционном образовательном процессе // Актуальные проблемы экономики и управления. 2020. № 4. С. 135–138. https://fs.guap.ru/emtp/journals/2020-4.pdf

14. *Никифоров А. Л.* Анализ понятия «знание»: подходы и проблемы // Эпистемология и философия науки. 2009. Т. 21. № 3. С. 61–73. http://www.intelros.ru/pdf/eps/03/05.pdf

15. *Прохоров А. М.* Большой энциклопедический словарь. М.: Норинт, 2004. 1456 с.

16. *Хобсон Л., Коул Х., Ханнес Х.* Обработка естественного языка в действии. СПб.: Питер, 2020. 576 с.

17. *Рубцова Ю. В.* Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1. С. 72–78. DOI: 10.15827/0236-235X.109.072-078

18. *Narynov S., Mukhtarkhanuly D., Omarov B.* Dataset of depressive posts in Russian language collected from social media // Data in Brief. 2020. Vol. 29. DOI: 10.1016/j.dib.2020.105195

19. *Будагов А. С., Дмитриева А. В., Космачев В. М., Мартыненко С. А., Москалева О. И., Степанов А. Г.* Самостоятельная работа студентов в магистратуре // Актуальные проблемы экономики и управления. 2016. № 2. С. 68–75. https://www.elibrary.ru/item.asp?id=26202416

20. *Lahoti U., Joshi R., Vyas N., Deshpande K., Jain S.* Drowsiness detection system for online courses // International Journal of Advanced Trends in Computer Science and Engineering. 2020. Vol. 9. No. 2. P. 1930–1934. DOI: 10.30534/ijatcse/2020/158922020

APPROACHES TO THE CHOICE OF TOOLS FOR CONSTRUCTING A METHODOLOGY FOR LEARNING TO WORK WITH BIG DATA

A. G. Stepanov[1], **G. A. Plotnikov**[1], **V. S. Vasilyeva**[1]

[1] ***Saint Petersburg State University of Aerospace Instrumentation***
190000, Russia, Saint Petersburg, ul. Bolshaya Morskaya, 67

Abstract

The article actualizes the need for teaching students to work with Big Data technologies. Big Data is a promising and fundamental industry that requires a large number of qualified specialists in various fields. The aim of the work is to describe the concept of determining a set of hardware, software, algorithmic and methodological tools (taking into account the contingent of students and the capabilities of the educational institution) for building a methodology for teaching a discipline related to the study of Big Data processing methods. There are two main sectors of stakeholders who need specialists in the field of Big Data. A detailed comparative analysis of software solutions that support Big Data processing is carried out. The article describes the methodology for constructing a course for teaching students technologies for processing and analyzing Big Data. A plan for organizing a lecture course and laboratory practice with consideration of subtasks is proposed for students to perform during training. The composition and methodology of independent work of students in the discipline related to the study of Big Data, using a learning management system such as Moodle, are discussed. An example of implementing data processing by means of the RapidMiner Studio package using a multi-layer neural network training algorithm using the error back propagation method is presented.

Keywords: RapidMiner Studio, Moodle, Data Mining, neural network, classification, clustering, regression, search for association rules, natural language processing, NLP.

DOI: 10.32517/0234-0453-2021-36-4-54-62

For citation:

Stepanov A. G., Plotnikov G. A., Vasilyeva V. S. Podkhody k opredeleniyu sredstv dlya postroeniya metodiki obucheniya rabote s bol'shimi dannymi [Approaches to the choice of tools for constructing a methodology for learning to work with Big Data]. *Informatika i obrazovanie — Informatics and Education*, 2021, no. 4, p. 54–62. (In Russian.)

Received: January 28, 2021.

Accepted: April 6, 2021.

About the authors

Alexander G. Stepanov, Doctor of Sciences (Education), Docent, Professor at the Department of Entrepreneurial Information Technologies, Institute of Entrepreneurship Technologies, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia; georgich_spb@mail.ru; ORCID: 0000-0002-3922-9684

Grigori A. Plotnikov, Assistant at the Department of Entrepreneurial Information Technologies, Institute of Entrepreneurship Technologies, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia; dim111077@mail.ru; ORCID: 0000-0002-8924-1382

Viktoria S. Vasilyeva, a master student at the Department of Entrepreneurial Information Technologies, Institute of Entrepreneurship Technologies, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia; victoria-vasilyeva@mail.ru; ORCID: 0000-0002-1429-0358

References

1. *Yusupov R. M., Yusupov Yu. V.* Sostoyanie i perspektivy razvitiya informatiki [State and prospects for the development of informatics]. *Trudy SPIIRAN — SPIIRAS Proceedings*, 2007, no. 5, p. 10–45. (In Russian.) Available at: http://www.mathnet.ru/links/79ca198018afb2fc5c21394ee2c7b0ce/trspy297.pdf

2. *Yusupov R. M.* Informatsionnye tekhnologii i ehkonomika informatsionnogo obshhestva [Information technologies and information society economy]. *Innovatsii — Innovations*, 2013, no. 11, p. 40–46. (In Russian.)

3. *Voronin D., Shevchenko V., Chengar O., Mashchenko E.* Conceptual Big Data processing model for the tasks of Smart Cities environmental monitoring. *DTGS 2019: Digital Transformation and Global Society. Communications in Computer and Information Science*. Cham, Springer, 2019, vol. 1038, p. 212–222. DOI: 10.1007/978-3-030-37858-5_17

4. *Sheng H., Wang X.* Intelligent retrieval system for ship fault information based on big data analysis. *Journal of Coastal Research*, 2019, vol. 93, is. 1, p. 1019–1025. DOI: 10.2112/SI93-147.1

5. *Zlatopolsky D. M.* Osnovy programmirovaniya na yazyke Python [Basics of programming in Python]. Moscow, DMK Press, 2018, 396 p. (In Russian.)

6. IBM SPSS Modeler. Available at: https://www.ibm.com/products/spss-modeler

7. End to End Data Science. *KNIME*. Available at: https://www.knime.com/

8. Qlik Analytics Platform. Available at: https://www.qlik.com/us/products/qlik-analytics-platform

9. Unikal'nye vozmozhnosti STATISTICA Data Miner [Unique features of STATISTICA Data Miner]. (In Russian.) Available at: http://statsoft.ru/products/STATISTICA_Data_Miner/unique-opportunities-statistica-data-miner.php

10. RapidMiner. Available at: https://rapidminer.com/

11. RapidMiner Documentation. Available at: https://docs.rapidminer.com/

12. *Stepanov A. G., Plotnikov G. A., Kosmachev V. M.* Assurance of safety of computer-based testing system for midterm knowledge assessment. *The European Proceedings of Social and Behavioural Sciences*, 2020, vol. 90, p. 1473–1485. DOI: 10.15405/epsbs.2020.10.03.170

13. *Kosmachev V. M., Plotnikov G. A., Stepanov A. G.* Prepodavatel' v distantsionnom obrazovatel'nom protsesse [Teacher in remote educational process]. *Aktual'nye problemy ehkonomiki i upravleniya — Actual Problems of Economics and Management*, 2020, no. 4, p. 135–138. (In Russian.) Available at: https://fs.guap.ru/emtp/journals/2020-4.pdf

14. *Nikiforov A. L.* Analiz ponyatiya "znanie": podkhody i problemy [Analysis of the concept of "knowledge": approaches and problems]. *Ehpistemologiya i filosofiya nauki — Epistemology and Philosophy of Science*, 2009, vol. 21, no. 3, p. 61–73. (In Russian.) Available at: http://www.intelros.ru/pdf/eps/03/05.pdf

15. *Prokhorov A. M.* Bol'shoj ehntsiklopedicheskij slovar' [Big encyclopedic dictionary]. Moscow, Norint, 2004. 1456 p. (In Russian.)

16. *Hobson L., Cole H., Hannes H.* Obrabotka estestvennogo yazyka v dejstvii [Natural Language Processing in Action. Understanding, analyzing, and generating text with Python]. Saint Petersburg, Piter, 2020. 576 p. (In Russian.)

17. *Rubtsova Yu. V.* Postroenie korpusa tekstov dlya nastrojki tonovogo klassifikatora [Constructing a corpus for sentiment classification training]. *Programmnye produkty i sistemy — Software & Systems*, 2015, no. 1, p. 72–78. (In Russian.) DOI: 10.15827/0236-235X.109.072-078

18. *Narynov S., Mukhtarkhanuly D., Omarov B.* Dataset of depressive posts in Russian language collected from social media. *Data in Brief*, 2020, vol. 29. DOI: 10.1016/j.dib.2020.105195

19. *Budagov A. S., Dmitrieva A. V., Kosmachev V. M., Martynenko S. A., Moskaleva O. I., Stepanov A. G.* Samostoyatel'naya rabota studentov v magistrature [Independent work of students in the master's degree]. *Aktual'nye problemy ehkonomiki i upravleniya — Actual Problems of Economics and Management*, 2016, no. 2, p. 68–75. (In Russian.)

20. *Lahoti U., Joshi R., Vyas N., Deshpande K., Jain S.* Drowsiness detection system for online courses. *International Journal of Advanced Trends in Computer Science and Engineering*, 2020, vol. 9, no. 2, p. 1930–1934. DOI: 10.30534/ijatcse/2020/158922020

НОВОСТИ

Дети с ограниченными возможностями будут учиться в московских школах по программам «Мобильного электронного образования»

Компания «Мобильное электронное образование» (совместный проект «1С» и группы «Ланит») выиграла грант правительства Москвы на интеграцию контента для учеников первых—третьих классов с ограниченными возможностями здоровья в «Московскую электронную школу» («МЭШ»).

Благодаря интеграции в МЭШ программ «Мобильного электронного образования» дети с особенностями развития (с нарушениями речи, задержкой психического развития и легкой степенью интеллектуальных нарушений) смогут получить адаптированный для них образовательный контент.

Платформа «Мобильное электронное образование» предлагает персональные траектории обучения не только детям с ограниченными возможностями здоровья, но и всем желающим обучаться по индивидуальным программам. Это возможно благодаря уникальному инструменту платформы — матрице назначения заданий, с помощью которой учитель может в несколько кликов подобрать отдельные задания для любого ребенка. Каждая программа в матрице в зависимости от уровня сложности отмечена индикатором определенного цвета. О том, что программа «особенная», известно только педагогу, в личном кабинете ученика это никак не обозначено.

(По материалам CNews)