4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова / МИЭМ. М., 2011. 272 с.

5. Татарникова Т. М. Защищенные корпоративные сети: задачи по защите информации / РГГМУ. СПб., 2012. 113 с.

6. Aggarwal C. C., Zhai C. Mining Text Data. Springer, 2012. 527 p.

7. Almeida T. A., Yamakami A. Advances in spam filtering techniques // Computational Intelligence for Privacy and Security Studies in Computational Intelligence. 2012. Vol. 394. P.199–214.

8. Berry M. W., Browne M. E-mail surveillance using nonnegative matrix factorization // Computational & Mathematical Organization Theory. 2005. Vol. 11, № 3. P. 249–264.

9. Татарникова Т. М. Анализ данных / СПбЭУ. СПб., 2018. 82 с.

10. Горковенко Д. К. Применение методов text mining для классификации информации, распространяемой в социальных сетях // Молодой ученый. 2016. № 4. С. 66–72.

B. Ya. Sovetov, T. M. Tatarnikova, A. I. Yashin
Saint Petersburg Electrotechnical University «LETI»

USE OF TECHNOLOGY TEXTMINING FOR IDENTIFYING HIDDEN THREATS IN COMMUNICATIONS DISTRIBUTED BY SOCIAL NETWORKS

A solution to the problem of text analysis using TextMining technology to detect threats hidden in messages exchanged between users in social networks has been proposed. The possibilities of TextMining technology in the tasks of knowledge detection in unstructured information arrays are discussed. The sequence of text analysis is presented in the form of a methodology. The content of the stages of the methodology is disclosed and the main techniques used at each stage are considered. The methods of calculating the weighting functions are selected, on the basis of which a list of keywords and phrases is formed. Ways of building semantic networks based on a set of keywords are considered. To automate text analysis, a software package that implements TextMining technology has been developed. The functions of the software package include the identification of keywords, relationships, and the emotional portrait of the user, which allows you to move from data to their meaning and draw conclusions about the information security of the text.

TextMining, semantic network, keywords, word occurrence frequency, text meaning, theme, text classification, dictionary text tonality, software package

УДК 004.942

Е. Е. Котова, А. С. Писарев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Задача классификации учащихся с использованием методов интеллектуального анализа данных

Технологически поддерживаемые учебные среды генерируют большое количество данных, которые могут быть собраны и проанализированы с помощью релевантных алгоритмов. Функции аналитики обучения необходимы для планирования и внесения изменений в организацию процессов обучения, обеспечения адаптивных рекомендаций и персонализированного анализа учебной деятельности. Существует несколько различных методов классификации, используемых в обнаружении знаний и добыче данных (Knowledge Discovery and Data Mining). Каждый метод или методика имеет свои преимущества. В статье применяются методы анализа данных к задаче классификации обучающихся. Один из вопросов – определение признаков дифференциации обучающихся. Предложены признаки дифференциации, характеризующие индивидуальные параметры познавательной сферы и составляющие модель когнитивно-стилевого потенциала обучающихся. Методы интегрированы в веб-среду интеллектуальной поддержки процессов обучения. Анализируются результаты, полученные при помощи нескольких классификаторов. Классификация по методу, предложенному в статье, с применением признаков когнитивно-стилевого потенциала дает более точные результаты по сравнению с полученными в других исследованиях и опубликованными в доступных источниках.

Анализ данных, классификация обучающихся, когнитивно-стилевой потенциал, веб-среда интеллектуальной поддержки

Разработка методов использования данных, получаемых из образовательного контекста, входит в активно развивающуюся область междисциплинарных исследований Educational Data Mining (EDM).

Область EDM включает методы, инструменты и исследования, предназначенные для автоматического извлечения уникальных знаний из больших репозиториев сгенерированных данных, связанных с учебной деятельностью и учебным процессом. Базы данных, накапливаемые в процессе обучения, хранят широкий диапазон переменных. Содержание баз данных растет от семестра к семестру и традиционно содержит информацию о студентах, баллах, оценках, рейтингах и др. Электронные и дистанционные формы обучения позволяют расширить диапазон накапливаемых данных более детальной информацией о результатах учебной деятельности обучающихся, формировании компетенций и динамики траекторий обучения.

В данном исследовании применяются методы интеллектуального анализа данных с целью получения ответов на вопросы: какие параметры познавательно-мыслительной деятельности могут быть преимущественными в учебной деятельности студентов? Можно ли найти такие *классы* или группы студентов, которые имеют сходные индивидуальные характеристики? Если да, то возможно ли определить, к какому классу будет относиться отдельный студент? И с помощью этой информации определить как помочь студенту, принадлежащему к определенному классу, использовать дидактические ресурсы лучше, основываясь на его личностном когнитивном потенциале и предпочтениях?

В статье описывается метод классификации учащихся на условные группы на основе диагностики параметров модели когнитивно-стилевого потенциала (КСП) обучающихся. Для тестовых данных были определены экспериментальные группы студентов разных курсов СПбГЭТУ «ЛЭТИ» и проведен ряд экспериментов. Представлены результаты применения методов анализа данных, применяемые для задач классификации и прогнозирования успеваемости студентов.

Текст статьи упорядочен следующим образом. Раздел 1 представляет постановку задачи. Разделы 2 и 3 представляют соответственно первый и второй этапы исследования. В разделе 4 рассматривается процедура идентификации параметров модели. В разделе 5 приводятся обсуждения полученных экспериментальных результатов и выводы по работе.

1. Постановка задачи. Для многих студентов адаптация к студенческой среде вуза проходит поразному. Выяснение причин академической неудачи среди студентов первого курса является важным вопросом для понимания и объяснения сложностей в процессе обучения студентов. Исследование направлено на то, чтобы как можно раньше, на ранних стадиях учебного процесса, идентифицировать обучающихся, испытывающих сложности в обучении. Задачей является классификация учащихся на условные группы по уровню риска академический неудачи: группу студентов с низким уровнем риска, которые имеют высокую вероятность успеха в дальнейшем обучении; группу среднего уровня риска – студентов, которые могут успешно обучаться по общим программам, принятым в университете, и группу студентов с высоким уровнем риска – студентов, которые имеют высокую вероятность неудачи («отсева», или отчисления). Для формализации задачи вводятся условные обозначения типовых профилей обучающихся (групп): High-profile (H), Average-profile (A), Special-profile (S). H-группа – группа студентов с низким уровнем риска, или группа так называемого продвинутого уровня обучаемости, характеризующаяся высоким уровнем продуктивности учебной деятельности. A-группа – группа среднего уровня риска, или группа минимально допустимого/достаточного/среднего уровня продуктивности учебной деятельности. S-группа – группа студентов высокого уровня риска, или недопустимого/недостаточного уровня продуктивности учебной деятельности (ниже среднего). Экспериментальную выборку составили учащиеся университета, обучающиеся на первом – четвертом курсах.

Первый этап исследования заключался в определении тестового набора диагностической системы экспресс-диагностики КСП обучающихся. Второй этап заключался в определении признаков внутренней дифференциации обучающихся на условные группы (классы H/A/S).

Реализация метода классификации учащихся предполагает разработку сценариев проведения исследований, обработку данных, программную реализацию и визуализацию результатов в вебсреде с применением агентного подхода.

2. Первый этап исследования. Определение тестового набора системы экспресс-диагностики когнитивно-стилевого потенциала. Тестовый набор системы диагностики КСП включает компьютерные версии методик, диагностирующих когнитивные стили, реализованных в подсистеме диагностики программного комплекса ОнтоМАСТЕР (ОнтоМАСТЕР-Диагностика). Программный комплекс реализует 9 модифицированных диагностических методик и позволяет диагностировать когнитивные параметры в виде

комплексной модели, в состав которой входят параметры КСП.

Диагностика КСП основана на понимании стиля, представленного в публикациях, в частности, книге под редакцией А. Либина «Стиль человека: психологический анализ» (1998) [1], где стиль человека определяется как устойчивая субъектно-специфическая система способов, или приемов осуществления человеком разных типов активности, в том числе интеллектуальной. Стиль определяется как интегральная характеристика формально-динамической сферы индивидуальности, проявляющаяся в предпочтении субъектом определенной формы взаимодействия с физической (предметной) и социальной (коммуникативно-символической) средой. «Стиль человека – это устойчивый целостный паттерн индивидуальных проявлений, выражающийся в предпочтении индивидуумом формы (способа) взаимодействия с физической и социальной средой».

Основные особенности стилевых различий обосновываются подходом, предложенным в рамках дифференциальной психологии (психологии индивидуальных различий) в системе «личность – познавательные процессы» [1], и подходом к представлению персонального познавательного стиля как интеграции познавательных стилей разных уровней [2].

Методики для диагностирования параметров КСП реализованы в подсистеме диагностики. Представим кратко их назначение.

Методика 1 представляет собой модифицированный компьютерный вариант методики «Скрытые фигуры» Л. Терстоуна (Thurstone L.L., 1944) [3]. Методика является разновидностью перцептивных тестов. При помощи методики диагностируются индивидуальные различия в познавательно-ориентировочной деятельности. Диагностируется параметр когнитивного стиля «поленезависимость-полезависимость» (ПЗ-ПНЗ), т. е. структурирующая способность в восприятии информации. Параметр ПЗ-ПНЗ характеризует индивидуальные различия в способах познавательной деятельности обучающихся и продуктивности работы с информацией: способах восприятия и структурирования информации, выделения наиболее значимых черт и элементов в какомлибо объекте. Параметр ПЗ-ПНЗ является одним из важных показателей обучаемости в условиях высокой информационной нагрузки учебного процесса, что подтверждено рядом исследований зарубежных и отечественных ученых [2].

Методика 2 представляет модифицированный компьютерный вариант теста MFFT–12 (Matching Familiar Figures Test) – «Выбор парной фигуры», разработанного Дж. Каганом (Kagan J., 1966) [4]. Диагностируемые параметры характеризуют когнитивный стиль «импульсивность – рефлексивность» (И-Р). При помощи данного когнитивного стиля определяются индивидуальные различия относительно времени отбора информации и оценивания гипотез для принятия решений, точности перцептивного сканирования как аспекта интеллектуальной продуктивности. Данное свойство проявляется в условиях неопределенности как стилевая характеристика индивидуальных особенностей решения задач при выборе из нескольких альтернатив. Различия в этом индивидуальном стиле проявляются при понимании сложных смысловых связей. Параметр И-Р выполняет функцию регуляции умственных действий при обучении и характеризует соотношение темпа и качества мыслительных процессов.

Методика 3 представляет модифицированный вариант теста Струпа (J. Ridley Stroop, 1935) «Словесно-цифровая интерференция» [5], с помощью которого измеряется уровень интерференции – параметр, характеризующий когнитивный контроль. Всю проблематику когнитивных стилей разделяют на две большие группы: группу когнитивных стилей, в которую входят такие стили, как ПЗ-ПНЗ, И-Р, и группу когнитивных контролей [6]. Контроли определяются как индивидуальные стратегии решения определенного класса интеллектуальных задач (перцептивных, мнестических). Различные виды контролей, по мнению исследователей, создают индивидуальный «паттерн» установок, или стиль. Контроль играет роль «медиатора» во взаимоотношениях субъекта с внешней средой, «запускается» пониманием условий задачи и зависит от отношения к требованиям инструкций [7]. При помощи методики 3 диагностируется параметр «гибкий–узкий (ригидный) когнитивный контроль (КК)» и ряд дополнительных параметров, которые возможно получить при помощи компьютерной обработки данных [8]. Выявление когнитивного контроля наиболее важно при наличии ситуаций когнитивного конфликта, характерных для учебного про-

цесса, насыщенного информацией различной модальности и когнитивной сложности. В ситуациях выполнения различных заданий студентами может проявиться эффект интерференции, являющийся следствием взаимодействия двух или более потоков информации в процессе их обработки, например вербальной и перцептивной информации и др., что является существенным фактором увеличения когнитивной нагрузки.

В описываемом исследовании при формировании модели КСП принята иерархическая структура, основанная на идее о иерархической структуре стиля, которая представляет собой совокупность определенных типов связей между разными параметрами индивидуальности [1]. Верхний уровень иерархии модели КСП представлен группой параметров когнитивного контроля, которые позволяют предварительно дифференцировать обучающихся на группы в зависимости от получаемых в результате диагностирования параметров. Следующий уровень модели представлен параметрами когнитивных стилей ПЗ-ПНЗ и И-Р для более глубокого уточнения параметров индивидуальных моделей учащихся группы риска, которые позволяют адаптировать учебный контент.

3. Второй этап исследования. Классификация обучающихся на основе данных экспресс-диагностирования. Параметры когнитивного контроля позволяют определить учащихся, попадающих в условную группу высокого уровня риска, или недостаточного уровня продуктивности учебной деятельности, которая требует особого внимания в организации индивидуального процесса обучения.

Компьютерная модификация методики Струпа позволяет рассматривать процесс выполнения заданий на основе динамического подхода. Динамический подход к проблеме когнитивного роста рассматривается в исследованиях процесса обучения (например, [9]–[15] и др.). При динамическом подходе «рост» уровня знаний и навыков происходит поэтапно во времени как результат усвоения последовательности элементов учебного процесса.

Результатом диагностирования когнитивного контроля по классической методике Струпа является суммарное время выполнения последовательно усложняющихся серий (наборов) из 100 заданий. Фиксация суммарного времени T_c предполагает, что время выполнения одного задания постоянно и равно усредненному значению. На рис. 1 представлена гистограмма распределения суммарного времени T_c (выборку составили 225 студентов I курса Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В. И. Ульянова (Ленина) (СПбГЭТУ «ЛЭТИ»)).

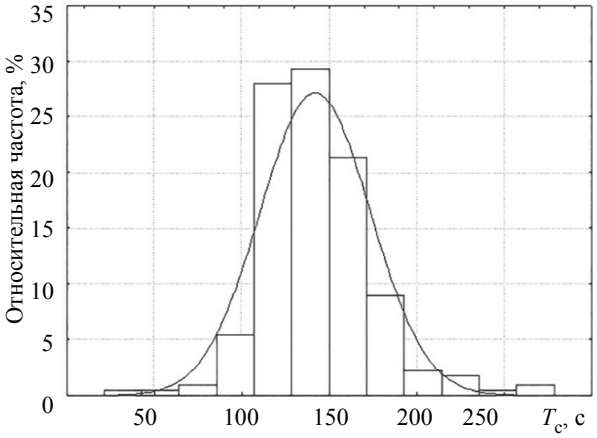
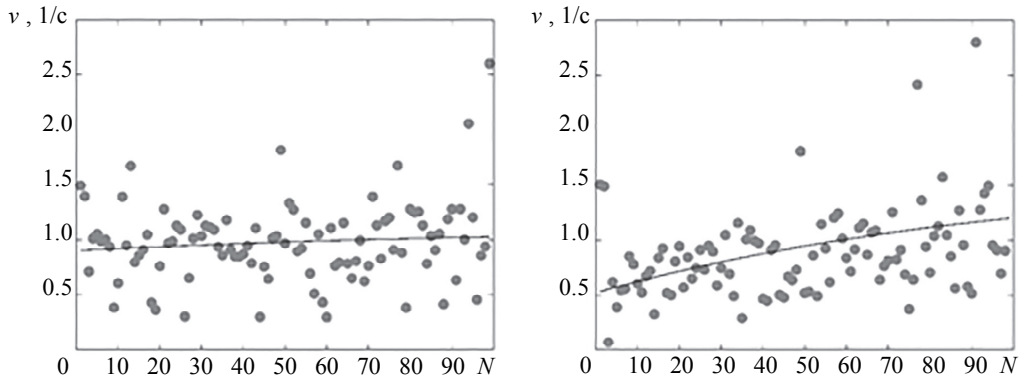


Рис. 1

Данные позволяют оценить границы суммарного времени выполнения заданий (min: 23.72 с, max: 275 с), среднее время выполнения (mean: 142 с), стандартное отклонение (std: 31.45). Анализ данных показывает, что испытуемые, которые затратили на выполнение заданий менее 90 с, совершают недопустимое число ошибок или не соблюдают инструкцию по выполнению заданий. Невозможность учета числа ошибочных решений в классическом варианте методики (в бланковом исполнении), когда учитывается только время выполнения заданий, приводит к некорректным выводам о способностях диагностируемых. Кроме того, единственного показателя (времени выполнения заданий T_c) не достаточно для оценки индивидуальных параметров когнитивной сферы и выявления динамики роста знаний и навыков. Ввиду высказанных соображений в компьютерную версию методики введены для обработки дополнительные показатели, которые можно получить из экспериментальных данных, характеризующие деятельность по освоению новой информации и автоматизации навыков восприятия и обработки информации:

$T_c3 - T_c2$ – показатель ригидности/гибкости когнитивного контроля, разница во времени выполнения третьей и второй серии заданий соответственно. Чем больше значение данного показателя, тем сильнее выражен эффект интерференции и, соответственно, более выражена ригидность (узость, жесткость) познавательного контроля;



а *б*

Рис. 2

T_c2 / T_c1 – показатель «вербальности», отношение времени выполнения второй и первой серии заданий соответственно. Высокие значения этого показателя свидетельствуют о преобладании словесного способа переработки информации, низкие – о преобладании сенсорно-перцептивного способа переработки информации;

$v_{уст}$ – установившееся значение скорости выполнения заданий; $T_{пер}$ – постоянная времени переходного процесса; $v_{нач}$ – начальное значение скорости выполнения заданий; $N_{ерг}$ – число допущенных ошибок.

Ранее авторами статьи была разработана компьютерная реализация модифицированной версии методики Струп-М [15] в среде Matlab (MathSoft), которая позволила фиксировать и получать данные диагностирования в виде процессов: кумулятивного времени выполнения заданий $T[n]$, времени выполнения одного задания $d[n]$ и скорости выполнения $v[n]=1/d[n]$. Время выполнения заданий фиксировалось с помощью встроенного модуля таймера с точностью до 0.01 с. Предварительная версия Matlab-программы необходима для тестирования и отладки методики при дальнейшей реализации программного инструмента в веб-среде и рассматривается как версия исследовательского прототипа. Для количественной оценки процессов диагностирования в [15] было предложено использовать 2 прямых показателя – длительность переходного процесса, т. е. число заданий $N_{пер}$, по истечении которого устанавливается время выполнения задания, и установившуюся скорость $v_{уст}$ выполнения заданий.

На рис. 2 изображены примеры данных диагностирования конкретных студентов технического вуза в виде графиков $v[n]$, полученных по методике Струп-М, с различающимися типами когнитивного потенциала (студент Н-группы (рис. 2, *а*) и студент S-группы (рис. 2, *б*) в принятых обозначениях групп обучающихся. Из полученных графиков видно, что кривые существенно различаются. После определенного числа выполненных заданий N скорость v устанавливается, что позволяет выделить переходную и установившуюся составляющие процесса.

Длительность переходного процесса и установившуюся скорость выполнения заданий можно рассматривать как приближенные оценки динамических и статических характеристик объектов диагностирования. Например, на рис. 2, *а* кривой соответствуют показатели $N_{пер}$ = 15, $v_{уст}$ = 1.1 с$^{-1}$; на рис. 2, *б* – $N_{пер}$ >100, $v_{уст}$ > > 1.21 с$^{-1}$. Результаты кривых интерпретируются следующим образом: студент 1 (рис. 2, *а*) достаточно быстро (на уровне 15-го задания) проявил ориентировку в восприятии новой информации и стал выполнять задания стабильно с постоянной скоростью. Студент 2 (рис. 2, *б*) практически испытывал сложности в выполнении, ориентировался в задании до его окончания. Студенты различаются и по показателю начальной скорости выполнения заданий.

В табл. 1 приведены значения параметров модифицированного теста Струп-М для двух студентов (студенты 1, 2), данные получены в веб-среде ОнтоМАСТЕР.

Таблица 1

Параметр	Студент Н-группы	Студент S-группы
$T_c3–T_c2$	23.98	22.98
T_c2/T_c1	0.95	1.09
T_c3	119.34	143.12
$N_{пер}$	15	>100
$v_{уст}$	1.1	>1.2
$v_{нач}$	0.91	0.53
$N_{ерг}$	1	3

При значительном количестве потенциальных учащихся становится актуальной автоматическая классификация, для чего необходимо формализовать оценку показателей процесса диагностирования по результатам экспресс-диагностирования.

В [15] была принята гипотеза о структуре модели: приращение скорости Δv_n пропорционально достигнутой скорости v_n и сложности задания u_n в виде линейного разностного уравнения первого порядка

$$\Delta v_n \equiv v_{n+1} - v_n = a v_n + b u_n, \qquad (1)$$

где a и b – подлежащие оцениванию в результате обработки данных коэффициенты.

Диагностирование по методике Струп-М предоставляет данные активного эксперимента для идентификации параметров модели (1). Входная последовательность из 100 заданий формально представляет единичную последовательность. Выходная последовательность представляет реакцию объекта на тестовую последовательность в виде скорости выполнения задания $v[n]$, $n = 1, \ldots, 100$, зависящей от номера задания n.

Оценивание параметров модели в [15] осуществлялось с помощью инструмента Parameter Estimation Tool программы Simulink [16]. Для идентификации параметров модели в автоматическом режиме программу Струп-М целесообразно дополнить модулем параметрической идентификации. Было отмечено, что результаты, получаемые методом наименьших квадратов для модели (1), весьма чувствительны к помехам. В связи с этим, для учета случайных помех был разработан вариант модели (1) и метод параметрической диагностики на основе стохастического дифференциального уравнения.

4. Процедура идентификации параметров модели. Модель процесса обучения (решения задач) представлена в виде обыкновенного дифференциального уравнения 1-го порядка:

$$T\frac{dy}{dt} + y = ku, \qquad (2)$$

где k – коэффициент усиления, интерпретируемый как коэффициент эффективности восприятия студентом управляющего дидактического воздействия; y – установившееся значение выходной величины $y(t_\infty) = ku$; T – постоянная времени апериодического звена, за которое выходная величина достигает приблизительно 0.63 от установившегося значения; $u = 1(t)$ – единичное ступенчатое управляющее дидактическое воздействие (дидактический ресурс в условных дидактических единицах).

Стохастический вариант модели (2) дополнительно учитывает воздействие случайных факторов ξ на параметр k:

$$\frac{dy}{dt} = -\frac{1}{T}y + \frac{(k + \sigma\xi)}{T}. \qquad (3)$$

Получим выражение для стохастического дифференциального уравнения в следующей форме:

$$dy_t = a\left(b - y_t\right)dt + c dW_t, \qquad (4)$$

где $a\left(b - y_t\right)$ – дрейф (drift); c –диффузия (diffusion); $dW_t = \xi(t)dt$ – винеровский процесс; σ – среднеквадратическое отклонение ($a = 1/T$, $b = k$, $c = \sigma/T$).

В обобщенной векторной форме система стохастических дифференциальных уравнений (СДУ) имеет следующий вид:

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t,$$

где X_t – случайный процесс с непрерывным временем.

Специальными случаями обобщенной модели СДУ являются: процесс Орнштейна–Уленбека (Ornstein-Uhlenbeck process); среднеинвертированный процесс Орнштейна–Уленбека (mean-reverting Ornstein-Uhlenbeck process); процесс Васичека (Vasicek process); броуновское движение (Brownian Motion – BM); геометрическое броуновское движение (Geometric Brownian Motion – GBM); модель Кокса–Ингерсолла–Росса (Cox-Ingersoll-Ross – CIR); модель Халла–Уайта–Васичека (Hull-White/Vasicek –HWV) и др. [17]–[20].

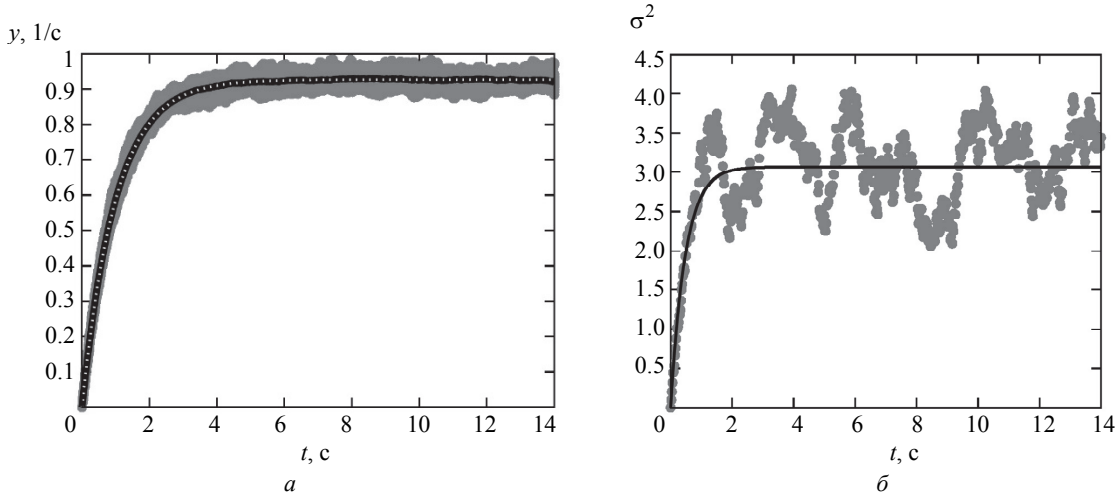
Модель (4) представляет собой модификацию случайных процессов Орнштейна–Уленбека: $dy_t = a y_t dt + c dW_t$ или Васичека: $dy_t = a\left(b - y_t\right)dt + c dW_t$ [17], [18].

Уравнение (4) имеет аналитическое решение, содержащее интеграл Ито:

$$y(t) = y(0)\, e^{-at} + b\left(1 - e^{-at}\right) + c e^{-at} \int_0^t e^{as} dW_s.$$

Параметры модели (4) могут быть найдены из совместного решения системы уравнений для средних значений выборки и дисперсий:

$$E(y_t) = b\left(1 - e^{-at}\right) + y_0 e^{-at}. \qquad (5)$$



а *б*

Рис. 3

$$\mathrm{Var}(y_t) = \frac{c^2}{2a}\left(1 - e^{-2at}\right). \qquad (6)$$

Параметрическая идентификация модели (4) в виде стохастического дифференциального уравнения осуществлялась с применением программы ОнтоМАСТЕР–Управление [19]. Для этого диагностические данные, полученные по модифицированной методике Струп-М, подвергались усреднению с получением зависимостей от времени для средних значений выборки и дисперсий (5) и (6) соответственно.

Метод был проверен на искусственных данных, сформированных в соответствии с моделью (4) с помощью метода Мильштейна (Milstein) [20].

На основе данных тестовой выборки была произведена идентификация параметров модели (4) и реконструкция зависимостей (5) и (6).

На рис. 3 приведены примеры траекторий в виде облака точек, соответствующих стохастической модели (4) с параметрами: $T = 1$, $k = 0.925$, $\sigma = 0.025$.

Реконструированная траектория (пунктирная линия) соответствует значениям параметров стохастической модели, идентифицированным по усредненной траектории и зависимости дисперсии от времени (рис. 3). Рис. 3, *а* – реконструкция зависимости (5) (пунктирная линия) с параметрами, идентифицированными по усредненной траектории (сплошная линия), соответствующей множеству экспериментально полученных траекторий (облако точек на графике). Рис. 3, *б* – реконструкция зависимости (6) (сплошная линия) по экспериментальным данным (облако точек на графике) с использованием идентифицированных значений параметров.

В табл. 2 представлены идентифицированные значения параметров уравнений (4), (5) и (6), которые демонстрируют возможность применения разработанного метода и программного обеспечения при обработке данных по методике Струп-М.

Таблица 2

Параметр	Исходное значение	Идентифицированное значение
T	1	0.9925
k	0.925	0.9266
σ	0.025	0.02499

На рис. 4 изображены примеры идентификации параметров модели по усредненным данным скоростей решения задач $E(y_t)$, обозначенных на графике <*y*>, и дисперсий $\mathrm{Var}(y_t)$, обозначенных σ^2, полученным из экспериментальных данных (см. рис. 2). Примеры усредненных значений <*y*> экспериментальных данных и дисперсий, полученных по методике Струп-М студентов 1 и 2 с различающимися типами когнитивного потенциала: «Н» (рис. 4, *а*) и «S» (рис. 4, *б*).

Наборы характеристик, формируемых по результатам обработки диагностических данных по методике Струп-М, наряду с данными других методик комплекса ОнтоМАСТЕР составляют базу признаков когнитивно-стилевого потенциала [21]. База знаний включает правила классификации, которые динамически генерируются в ходе обработки полученных параметров КСП с применением методов анализа данных на основе теоремы Байеса, деревьев решений, опорных векторов, нейронных сетей и др. Результаты применения методов анализа данных представлены в табл. 3.

В случае классификации на 3 класса метод AdaBoostM1 оказался лучшим по модифициро-

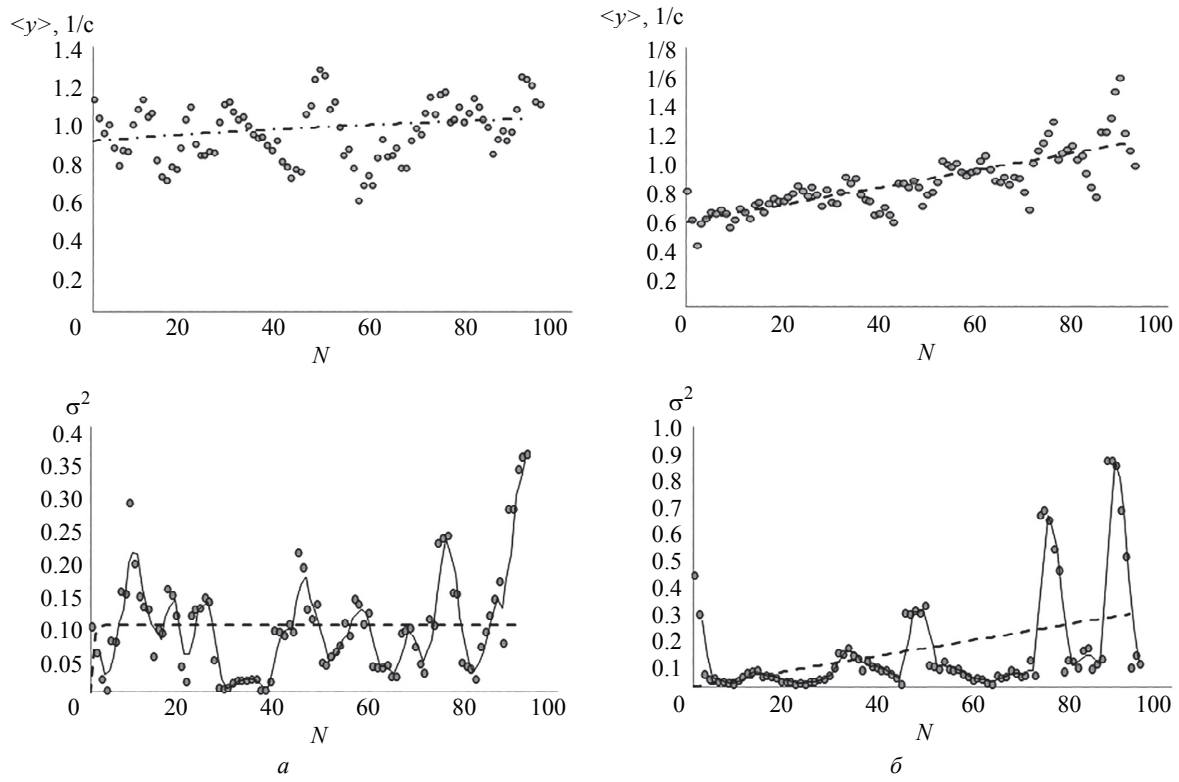


Рис. 4

Таблица 3

Число факторов	Метод	Процент точности классификации метода, test_pc	Число классов	Значения критериев точности					
				Aicc	kappa	mae	rmse	rae	rrse
4	AdaBoostM1	62.5	3	–45.8	0.4269	0.3007	0.4215	68.1951	89.697
45	RandomForest	93.75	2	–282.365	0.8667	0.2259	0.3019	45.5126	60.794
45	BayesNet	90.625	2	–283.433	0.7895	0.1023	0.2969	20.6035	59.7817
3	j48	87.5	2	–62.7919	0.7143	0.1528	0.3368	30.7754	67.8158

ванному критерию Акаике (aicc), т. е. наиболее точным методом, использующим наименьшее число факторов.

В случае классификации на 2 класса лучший результат по критерию точности показал метод RandomForest: 93.75 % (отношение числа студентов, правильно предсказанных по принадлежности к классу, к общему числу студентов). В методе применены 45 факторов КСП. Эти факторы были получены по результатам обработки так называемых сырых данных методик 1, 2 и 3, полученных в ходе экспресс-диагностирования.

Метод байесовских сетей BayesNet показал точность классификации 90.625 %. Этот метод оказался лучшим по критерию наименьшей среднеквадратической ошибки (rmse) 0.2969 и информационному критерию Акаике (aicc) со значением (–283.433).

По методу деревьев решений j48 получена точность классификации 87.5 %. Вопрос числа признаков оказался достаточно интересным для исследования. Из полученных результатов заметно, что для метода j48 при классификации на 2 класса оказалось достаточно только трех признаков. Методом j48 из 45 признаков были определены 3 наиболее значимых, обозначенных в программе errsum2, errsum3, Н3. Они характеризуют важные параметры когнитивно-мыслительной деятельности по восприятию новой информации, а именно число ошибок (errsum2, errsum3 – число ошибок, допущенных испытуемым во второй и третьей сериях задания Струп-М) и Н3 – коэффициент продуктивности, что подтверждает предположение о значимости признаков дифференциации, относящихся к параметрам познавательно-мыслительной сферы КСП.

Полученные на тестовой выборке результаты по точности классификации при использовании признаков КСП превосходят некоторые результаты других исследований. Например, при помощи метода классификации, основанного на моделях, включающих в том числе оценки довузовской и академической успеваемости, получена точность 66 % [22].

На основании полученных признаков в соответствии с описанными процедурами осуществляются дальнейшие этапы: классификация обучаемых, прогнозирование результатов обучения при средних и дополнительных (индивидуальных) дидактических ресурсах.

5. Обсуждение результатов и заключение. Стратегии формирования индивидуальных программ обучения предполагают дифференциацию (классификацию) разнородного контингента учащихся на группы. Задача классификации обучающихся представляет интерес для решения с применением методов интеллектуального анализа данных. Методы EDM позволяют достаточно оперативно получить важную информацию об индивидуальных различиях когнитивных характеристик обучающихся для детального построения индивидуального процесса обучения.

Цель авторов статьи – представить метод, позволяющий классифицировать студентов до первой экзаменационной сессии на группы. Это позволит определить тех студентов, которые нуждаются в помощи и дополнительном внимании к формированию учебных программ, заданий, и предложить им конкретные учебные действия. Ранее в исследованиях авторов было показано, что на основании баллов ЕГЭ невозможно построить прогноз результатов обучения в вузе.

Особенностью анализа данных в сфере образования являются небольшие наборы данных ввиду небольшого количества учащихся в учебных группах.

Возникает 2 важных вопроса: количество групп при классификации и подготовка признаков дифференциации для решения задачи классификации.

Ввиду небольшого количества студентов в группах и разнородности состава (из года в год меняется состав обучающихся) один из вопросов детального исследования касается определения оснований для классификации. В исследовании проводился анализ определения количества групп классификации. Попытки разделить обучающихся по уровню академической успеваемости на 3, 4 и большее число групп показали, что увеличение числа групп не приводит к значимым различиям (например, 3 группы, см. табл. 3). Аналогичные результаты получены и зарубежными исследователями. Возможна группировка учащихся различными способами. Например, предлагается разделять студентов на 2 и на 3 класса (класс «высоких», «средних» и «низких» в зависимости от оценок в баллах) или даже на 9 классов (в зависимости от результатов в баллах классы определяются по интервалам оценок) [23]. Авторы исследований [24] отмечают, что ни один алгоритм не обеспечивает процент правильно классифицированных результатов выше 70 %. При классификации обучающихся на 4 или 3 класса результаты классификации не превышают аналогичных значений.

В исследованиях встречаются различные интересующие факторы (переменные). Например, в качестве переменных используются: возраст учащегося, средний процент занятий в течение одной недели, количество часов по математике, изучаемой на уровне средней школы, или их среднее значение в конце последнего класса и др. [25]. Авторы применяют разные методы для классификации студентов на 3 группы по так называемому уровню риска успешности обучения (определение групп дано авторами статьи [25]). Общий процент правильно классифицированных студентов достигает 40.63 % для алгоритма деревьев решений ID3, 51.78 % для алгоритма случайных деревьев решений CART, 51.88 % для нейронных сетей.

В предложенной статье методы EDM применяются к анализу данных познавательной сферы обучающихся для получения представления о когнитивных моделях (в рассматриваемом случае введено понятие модели когнитивно-стилевого потенциала обучающихся), в которых поведение учащихся отслеживается и архивируется с целью дальнейшего выявления типовых профилей.

Предложен метод нахождения признаков дифференциации на основе параметров КСП обучающихся, полученных по данным экспресс-диагностирования при помощи модифицированной методики Струп-М. Объем и точность этих данных таковы, что даже довольно короткая сессия в компьютерной системе диагностирования (например, методика Струп-М занимает всего около 15…20 мин) может привести к большому количеству данных для анализа. Время получения результата по методам классификации составляет 0.01 с.

Обучение и поведение в целом – это сложная система, в которой многие процессы и подсистемы динамически взаимодействуют, создавая явления, которые мы наблюдаем [26]. Предлагаемый метод основан на гипотезе о допустимости экстраполирования процедуры экспресс-диагностирования на материале решения задач с когнитивной нагрузкой как имитации будущего процесса обучения «в ускоренном времени». В определенной степени гипотеза относится к так называемой супервентности: как процессы, работающие в разных масштабах времени, взаимодействуют, чтобы произвести наблюдаемое и предполагаемое поведение [26].

В статье сравниваются эффективность и полезность различных методов интеллектуального анализа данных для классификации обучающихся с целью получения полезных данных с использованием среды интеллектуальной поддержки процесса обучения ОнтоМАСТЕР, в которой реализованы 50 методов анализа данных для разных задач (в частности, применен метод ансамбля). В итоге исследованы различные методы и получены результаты классификации на трех и двух классах.

Основные функции комплекса ОнтоМАСТЕР заключаются в диагностике когнитивно-стилевого потенциала обучающихся; выявлении признаков дифференциации, классификации обучающихся на группы; определении типовых профилей, построении прогноза траекторий обучения; распределении персональных дидактических ресурсов. Такая информация даст представление о дизайне учебной среды, которая позволяет студентам, преподавателям, методистам, администраторам принимать обоснованные решения по взаимодействию, предоставлению образовательных ресурсов и управлению ими, учитывая индивидуальные особенности и предпочтения студентов. Экспресс-диагностирование КСП может проводиться дистанционно для нескольких групп обучающихся или индивидуально. Время отклика системы для получения результата построения модели составляет около 10…15 с в зависимости от числа пользователей, одновременно участвующих в сеансе диагностирования.

В дальнейшем целесообразно решить следующие вопросы: будут ли факторы, влияющие на успехи в учебе, стабильными из года в год в одном и том же университете, например, при подготовке по одному направлению? Возможно ли найти факторы, общие для разных университетов при прогнозировании успешности обучения? Может ли сочетание разных методов прогнозирования привести к улучшению общего результата?

Задачи прогнозирования результатов обучения предполагается рассмотреть в последующих публикациях.

Авторы статьи благодарят д-ра техн. наук, проф. Имаева Д. Х. за ценные замечания и поправки к статье, которые были учтены при подготовке окончательной версии.

СПИСОК ЛИТЕРАТУРЫ

1. Стиль человека. Психологический анализ / под ред. А. В. Либина. М.: Смысл, 1998. 310 с.

2. Холодная М. А. Когнитивные стили: О природе индивидуального ума: учеб. пособие. М.: ПЕР СЭ, 2002. 304 с.

3. Thurstone L. L. A factorial study of perception. Chicago: University of Chicago Press, 1944. 148 p.

4. Kagan J. Reflection-impulsivity: The generality and dynamics of conceptual tempo // J. of abnormal psychology. 1966. Vol. 71, №. 1. P. 17–24.

5. Stroop J. R. Studies of interference in serial verbal reactions // J. of Exper. Psychology. 1935. Vol. 18. P. 643–662.

6. Когнитивные стили. Тезисы науч.-практ. семинара / под ред. В. Колги; Таллинский пед. ин-т им. Э. Вильде. Таллин, 1986. 250 с.

7. Аверин В. А., Киреева Н. Н., Котова Е. Е. Интеллектуально-стилевая организация человека: учеб. пособие для преподавателей и студентов. СПб.: Изд-во СПбГПМУ, 2014. 36 с.

8. Котова Е. Е., Падерно П. И. Экспресс-диагностика когнитивно-стилевого потенциала обучающихся в интегрированной образовательной среде // Образовательные технологии и общество. 2015. Т. 18, № 1. С. 561–576.

9. Van Geert P. Dynamic Systems Model of Cognitive and Language Growth // Psychological Review, 0033-295X. Jan. 1, 1991. Vol. 98, № 1. URL: http://www.paulvangeert.nl/publications_files/psychological%20review%201991.htm.

10. Майер Р. В. Кибернетическая педагогика: имитационное моделирование процесса обучения. Глазов: Изд-во ГГПИ, 2013.

11. Имаев Д. Х., Котова Е. Е. Модели и алгоритмы принятия решений о распределении дидактических ресурсов в среде обучения // Изв. СПбГЭТУ «ЛЭТИ». 2013. № 8. С. 79–85.

12. Имаев Д. Х., Котова Е. Е. Система управления процессом обучения с логическими алгоритмами принятия решений // Изв. СПбГЭТУ «ЛЭТИ». 2013. № 10. С. 84–90.

13. Имаев Д. Х., Котова Е. Е. Компьютерная имитация процесса обучения в условиях периодического контроля успеваемости // Изв. СПбГЭТУ «ЛЭТИ». 2014. № 1. С. 74–79.

14. Имаев Д. Х., Котова Е. Е. Моделирование и имитация процессов обучения с разделением дидактических ресурсов. Динамический подход. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2014. 111 с.

15. Имаев Д. Х., Котова Е. Е. Дифференциация учащихся по показателям экспресс-диагностирования // Изв. СПбГЭТУ «ЛЭТИ». Современные технологии образования. 2014. № 10. С. 71–77.

16. MATLAB. URL: http://www.mathworks.com/products/matlab.

17. Оксендаль Б. Стохастические дифференциальные уравнения. Введение в теорию и приложения / пер. с англ. М.: Мир; ООО «Издательство АСТ», 2003. 408 с.

18. Iacus S. M. Simulation and inference for stochastic differential equations: with R examples. Springer Science & Business Media, 2009. 300 p.

19. Котова Е. Е., Писарев И. А. Применение многоагентных технологий и эвристических методов в on line системе управления обучением студентов // Всерос. науч. конф. по проблемам управления в технических системах. Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет ЛЭТИ им. В. И. Ульянова (Ленина). 2015. № 1. С. 182–185.

20. Котова Е. Е., Писарев А. С. Адаптивное прогнозирование результатов обучения студентов в режиме online // Всерос. науч. конф. по проблемам управления в технических системах. Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет ЛЭТИ им. В. И. Ульянова (Ленина). 2017. № 1. С. 143–146.

21. Котова Е. Е., Печников А. Н., Писарев А. С. Программный комплекс диагностики когнитивных параметров специалиста (ОнтоМАСТЕР-Диагностика). Свидетельство о государственной регистрации программы для ЭВМ № 2009615001 от 14 сент. 2009 г.

22. Kabakchieva D. Predicting student performance by using data mining methods for classification // Cybernetics and information technologies. 2013. Vol. 13, № 1. P. 61–72.

23. Minaei-Bidgoli B. Predicting student performance: an application of data mining methods with an educational web-based system // IEEE. 2003. Vol. 1. P. T2A–13.

24. Romero C. Data mining algorithms to classify students // Educational data mining. The 1st Intern. Conf. on Educational Data Mining. Montréal, Québec, Canada, 2008. P. 8–17.

25. Superby J. F., Vandamme J. P., Meskens N. Determination of factors influencing the achievement of the first-year university students using data mining methods // Workshop on Educational Data Mining. 2006. Vol. 32. P. 234.

26. Jacobson M. J. et al. Rising above? Implications of complexity for theories of learning // Proc. of Intern. Conf. of the Learning Sciences, ICLS. Intern. Society of the Learning Sciences. London: United Kingdom, 2018. Vol. 2. P. 1328–1333.

E. E. Kotova, A. S. Pisarev
Saint Petersburg Electrotechnical University «LETI»

THE PROBLEM OF CLASSIFICATION OF STUDENTS USING THE METHODS OF INTELLECTUAL DATA ANALYSIS

Technologically supported learning environments generate a large amount of data that can be collected and analyzed using relevant algorithms. Learning analytics functions are necessary for planning and introducing changes in the organization of learning processes, providing adaptive recommendations and personalized analysis of learning activities. There are several different classification methods used in knowledge discovery and data mining (Knowledge Discovery and Data Mining). Each method or technique has its advantages. Data analysis methods are applied to the task of classifying students. One of the questions is the definition of signs of differentiation of students. The signs of differentiation, characterizing the individual parameters of the cognitive sphere, and components of the model of the cognitive-style potential of students are proposed. These methods are integrated into the web environment of intellectual support of learning processes. The results obtained with the help of several classifiers are analyzed. The classification according to the method proposed in the article gives more accurate results than those obtained in other studies and published in the available sources.

Data analysis, student classification, cognitive-style potential, intellectual support web environment