УДК 004.8


Т.В. Ящун, Е.В. Громов
Вологодский государственный университет

ПРОГНОЗИРОВАНИЕ КАЧЕСТВА УЧЕБНОЙ ДЕЯТЕЛЬНОСТИ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Проанализированы подходы к решению проблемы прогнозирования качества учебно-познавательной деятельности (УПД) студентов на основе методов интеллектуального анализа данных (EDM). Отмечено, что учебная деятельность будущих специалистов компьютерных наук имеет свою специфику, так как неразрывно связана с алгоритмической деятельностью и взаимодействием с техническими устройствами. Это непосредственно влияет на подбор входных параметров модели. Предложена программная реализация одного из методов машинного обучения – прогнозирования на основе построения многофакторной регрессионной модели на базе метода группового учета аргументов (МГУА) – в приложении к УПД студентов компьютерных специальностей. В целях исследования предложено расширение метода EDM механизмом саморегуляции учебной деятельности.

Интеллектуальный анализ данных, машинное обучение, многофакторная регрессионная модель, учебно-познавательная деятельность, прогнозирование, саморегуляция.

Среди современных педагогических тенденций можно выделить индивидуализацию обучения. Однако реализация такого направления в рамках коллективного обучения, принятого в классических высших учебных заведениях, довольно сложна и не всегда продуктивна. Повышенное внимание к отдельному индивидууму влечет за собой снижение качества обучения основной группы. Решение подобного противоречия возможно средствами прогнозирования учебно-познавательной деятельности с определением значений как личных факторов студента, так и факторов, отвечающих за методику преподавания, способствующих оптимизации процесса в целом.

Прогнозирование собственной учебной деятельности позволит студенту выполнять ее саморегуляцию с целью достижения желаемого результата. Подбор факторов методики преподавания для каждого студента, в свою очередь, поможет разрешить проблему индивидуализации обучения.

Существующие подходы к решению поставленных задач

Прогнозирование УПД является одним из направлений интеллектуального анализа данных в образовании (EDM – Educational Data Mining), связанного с разработкой методов, позволяющих извлекать знания из данных, поступающих из образовательной среды [1]. Анализ решения проблемы прогнозирования успеваемости учащихся проводится в соответствии со следующими логическими этапами [2]:

- формирование наборов данных – выделение множества параметров обучаемого и сбор информации о студентах;
- разметка набора данных для обучающей последовательности модели – отсев несущественной информации и занесение собранной статистической информации в базу данных на основе ее предварительного измерения;
- обучение модели – применение алгоритмов моделирования к наборам данных и формирование модели прогнозирования;
- апробация построенной модели на тестовых данных, определение точности прогнозирования.

Входной набор данных в виде множества параметров обучаемых зависит от целей исследования, но имеет общие логические закономерности определения. Чаще всего это начальные или предварительные результаты обучения студентов, личностные характеристики, показатели работы в системах дистанционного управления обучением (LMS) (например, Moodle) [3–5], статистика просмотра учебной информации в обучающих средах [3], баллы на вступительном экзамене, результаты внешнего независимого тестирования [2, 7], результаты обучения по профилирующим предметам в выпускных классах [2, 6], пол [2, 7], посещаемость на начальном этапе обучения [8], результаты входного тестового контроля [8], мотивационные убеждения (ориентация на цель, ценностные ориентации, тестовая тревожность и т.п.) [7], саморегулируемые компоненты обучения (использование когнитивных стратегий, саморегуляция) [7].

Составляющие входного набора данных могут иметь разную информативность и, следовательно, различное влияние на конечный результат прогнозирования. Большинство исследователей уделяет отдельное внимание отсеву несущественных параметров на основе вычисления энтропии по известным формулам из области математической статистики [2, 8, 9]. Для разметки входных наборов данных с целью

формирования обучающей последовательности модели предлагается использование как количественных, так и качественных методов (онлайн-анкеты, опросники стиля обучения и мотивированных стратегий для обучения, полуструктурированные интервью и т.п.) [7].

Самым исследуемым и разносторонне реализуемым является этап предсказания результатов учебной деятельности на основе методов машинного обучения:

- классификация, деревья решений [2, 8, 9];
- регрессионный анализ [3, 6, 7];
- нейронные сети [4];
- кластеризация, искусственный интеллект, правила ассоциации, генетический алгоритм, метод ближайшего соседа и т.д.

По результатам выполненного обзора для решения задач исследования целесообразно применение регрессионного анализа на основе самоорганизующихся моделей. Данная технология положена в основу метода группового учета аргументов (МГУА) [10–12], используемого при моделировании небольшого количества статистической информации и выполняющего автоматический отсев несущественных параметров. На сегодняшний день развито применение метода МГУА в различных областях человеческой деятельности, математических исследованиях [13, 14]. Использование МГУА для прогнозирования показателей качества УПД студентов является мало исследованной, хотя и перспективной, областью.

Анализ результатов исследований позволяет сделать вывод не только об актуальности прогнозирования УПД студентов как одной из областей EDM, но и об эффективности использования методов машинного обучения [8, 15] в этой сфере. Результаты обзора способствуют определению структуры множества входных параметров модели УПД студентов и возможных методов их измерения на основе разнообразных анкет и опросов. Проведенный анализ позволяет заявить о новизне МГУА как основного метода моделирования УПД. Применение МГУА в других областях человеческой деятельности подтверждает его эффективность.

Целью исследования является решение задачи прогнозирования показателей качества УПД студентов компьютерных специальностей с использованием одного из методов машинного обучения – регрессионного анализа на основе метода группового учета аргументов в соответствии с логическим этапам решения этой проблемы.

Применяемые методы научного исследования

Для построения модели УПД студентов необходимо решить две обобщенные задачи: сформировать набор данных – множество входных наиболее информативных параметров, выбрать и формализовать метод моделирования.

Формирование множества информативных параметров основывается на измерении меры информативности и отборе параметров с наибольшими значениями такой меры. Для измерения параметров УПД, как было предложено в работе [7], можно использовать унифицированные анкеты-опросники. Для выбора множества информативных параметров УПД используются методы измерения меры информативности на основе энтропийного подхода. Обозначим меру информативности фактора (параметра) X об определяющем показателе качества УПД (оценке обучаемого) Z через $I(Z/X)$. Тогда на основе информационных статистик Шеннона [16] мера информативности вычисляется по формуле:

$$I(Z/X) = H(Z) + H(X) - H(Z,X), \quad (1)$$

где $H(Z)$, $H(X)$, $H(Z,X)$ – соответственно энтропия явления, фактора и совместная (явления и фактора).

Для формирования модели УПД студентов, как указывалось ранее, предлагается использовать метод группового учета аргументов, представляемый в виде функциональной зависимости:

$$Z = F(x_1, x_2, \dots, x_n;\ y_1, y_2, \dots, y_m), \quad (2)$$

где Z – показатель качества УПД; F – некоторая функция; x_i – характеристики обучаемых, $i = \overline{1,n}$, n – их количество; y_j – характеристики методики обучения, $j = \overline{1,m}$, m – их количество.

МГУА – метод математического моделирования сложных систем, основанный на принципе самоорганизации моделей. Согласно этому принципу осуществляется целенаправленный перебор постепенно усложняющихся структур моделей и их отбор по ряду критериев. Автор модели указывает только общие критерии выбора и список возможных переменных, взятый с большим запасом. Метод выбирает наиболее эффективное множество входных переменных. Согласно принципу самоорганизации при постепенном усложнении структуры модели значение некоторого заданного внешнего критерия сначала уменьшается, а затем возрастает, т.е. имеется минимум критерия, соответствующий модели оптимальной сложности. Построение модели УПД студентов заключается в нахождении вида неизвестной функции, соответствующей минимальному значению некоторого внешнего критерия:

$$f^* = \underset{f \in F}{arg\,\min}\, CR(f), \quad (3)$$

где f^* – оптимальная модель, F – множество рассматриваемых моделей, CR – внешний критерий качества модели f из этого множества.

Процесс решения задачи (3) включает в себя следующие этапы [10]: получение статистических данных, выбор способа разбиения данных на обучающую и проверочную последовательности, выбор класса базисных функций, оценивание параметров генерируемых структур и формирование множества F, выбор внешнего критерия, минимизация внешнего критерия CR и выбор оптимальной модели. Формализация данного процесса может быть реализована с помощью алгоритма, представленного на рисунке 1, где ЧО – это частные описания модели, коэффициенты которых оцениваются во время моделирования.

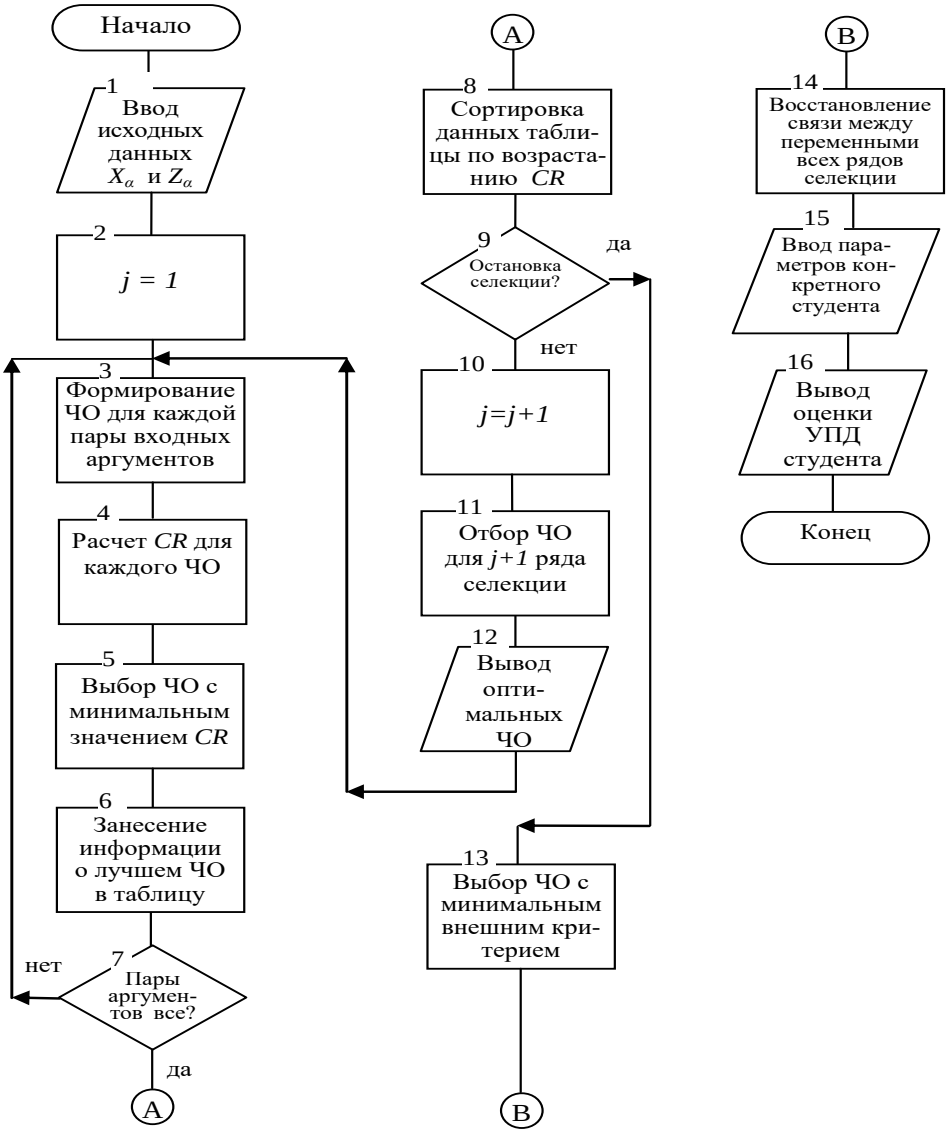

Рис. 1. Схема алгоритма моделирования УПД на основе МГУА

Основная часть

Приведенный выше теоретический материал был конкретизирован и экспериментально апробирован на контрольной группе студентов.

На начальном этапе входные параметры модели УПД были отранжированы в порядке убывания их информативности: средний балл в аттестате – 0,855, оценка по математике – 0,48, оценка по информатике – 0,48, мотивация обучения – 0,375, внимание – 0,287, исполнительность – 0,193 и т.д. Всего исследовано 15 характеристик обучаемых, выделенных с учетом специфики обучения студентов компьютерного профиля, и 15 характеристик методов обучения. Результаты вычисления информативности позволили сформировать множество входных параметров модели взаимодействия, состоящее из 20 параметров.

На следующем этапе исследования метод группового учета аргументов претерпел формализацию и программную реализацию. В результате работы программного средства сформирована цепочка регрессионных зависимостей, отражающих связь параметров УПД и оценок студентов (4).

В формулах (4) цепочка регрессионных зависимостей имеет 7 уровней вложенности. Последний блок зависимостей представляет собой формулы, выражающие зависимость только от переменных x_i, т.е. от параметров УПД. Данные зависимости позволяют также судить о множестве наиболее существенных факторов УПД. Цепочка регрессионных зависимостей (4) имеет довольно громоздкий вид. Задача восстановления итоговой регрессионной зависимости является не только сложной, но и нецелесообразной. Действительно, в результате должна получиться формула, имеющая порядка 10^{45} слагаемых. Для использования формул (4) можно воспользоваться, например, табличным процессором Microsoft Excel.

$$
\begin{aligned}
&r = 0.002 + 1.49t_7 - 0.47t_8; && z_1 = 1.3457 - 0.0323y_{10} + 0.1716y_5y_{10}, \\
&t_7 = 0.46 + 0.75u_5 + 0.033u_5u_8, && \dots \\
&t_8 = 0.178 + 0.9u_5 + 0.0126u_5u_{10} && y_2 = -17.1874 + 4.1903x_1 + 3.2291x_{11} - 0.6183x_1x_{11} \\
&u_5 = 1.845 + 0.1324w_2w_5, && y_3 = -7.7721 + 2.3161x_1 + 1.0793x_8 - 0.1912x_1x_8, \\
&\dots && y_5 = -3.3736 + 1.3985x_1 + 0.0306x_1x_6, \\
&w_2 = 0.0495 + 3.1v_1 - 2.1v_6, && y_7 = 3.9952 - 2.2411x_5 + 0.4662x_1x_5, \\
&\dots && y_9 = -3.4786 + 1.5532x_1, \\
&v_1 = 1.7684 + 0.1372z_1z_2, && y_{10} = 3.6227 - 1.5626x_4 + 0.3471x_3x_4. \\
&\dots &&
\end{aligned}
\tag{4}
$$

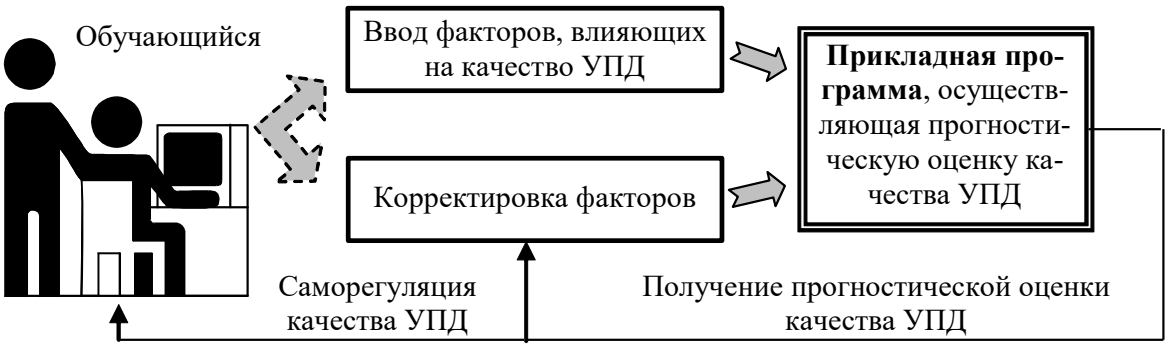

Рис. 2. Схема саморегуляции качества УПД

Предлагаемый метод имеет возможность применения в области саморегуляции обучаемыми качества своей УПД, что особо актуально в сфере современных тенденций дистанционного и онлайн-обучения. Действительно, имея в арсенале представленные регрессионные зависимости и оценив с помощью анкетопросников уровни своих личностных характеристик, каждый студент сможет провести саморегуляцию своей учебной деятельности (рис. 2).

Таким образом, в результате исследования на основе одного из методов машинного обучения – многофакторного регрессионного анализа – построена прогностическая модель УПД обучаемого. Данный метод позволяет не только получить прогностическую оценку студента в конце некоторого периода обучения, но и расширить методы интеллектуального анализа данных (EDM) предложенной концепцией саморегуляции УПД.

Выводы

Предложен метод прогнозирования учебно-познавательной деятельности студентов компьютерных специальностей на базе одной из технологий машинного обучения – регрессионного анализа методом группового учета аргументов. Метод реализован с помощью программного средства, позволяющего определять прогностическую оценку студента на некотором этапе обучения и цепочку регрессионных зависимостей, отражающих связь входных параметров модели (характеристик обучаемого и методики преподавания) с определяющим показателем качества УПД. Определение прогностических оценок студентов позволяет не только выполнять предварительную коррекцию учебного процесса со стороны преподавателя, но и реализовывать концепцию саморегуляции обучаемыми своей учебной деятельности, что в свою очередь способствует индивидуализации обучения и повышению эффективности дистанционного и онлайн-обучения.

Литература

1. Han, J. Data Mining: Concepts and Techniques / J. Han, M. Kamber, J. Pe. – 3th edn. – 2012. – p. 703 – URL: https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques?via=ihub= (дата обращения: 02.01.2024). – Текст : электронный.

2. Predicting students' performance using ID3 and C4.5 classification algorithms / K. Adhatrao, A. Gaykar, A. Dhawan [et al.] // International Journal of Data Mining & Knowledge Management Process (IJDKP). – 2013. – Vol. 3, № 5. – P. 39–52.

3. A neural network approach for students' performance prediction / F. Okubo, T. Yamashita, A. Shimada, H. Ogata // Proceedings of the Seventh International Learning Analytics & Knowledge Conference. – 2017. – P. 598–599.

4. WonYou, J. Identifying significant indicators using LMS data to predict course achievement in online learning / J. WonYou // The Internet and Higher Education. – 2016. – Vol. 29. – P. 23–30.

5. Канаш, А. В. Интеллектуальный анализ данных для построения моделей машинного обучения в образовании / А. В. Канаш, А. С. Мезина, Т. А. Ткалич // Цифровая трансформация – шаг в будущее : материалы II Международной научно-практической конфе-

ренции молодых ученых, посвященные 100-летию Белорусского государственного университета (Минск, 27 октября 2021 г.). – Минск : БГУ, 2021. – С. 135–139.

6. Huang, Sh. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models / Sh. Huang, N. Fang // Computers & Education. – 2013. – Vol. 61. – P. 133–145.

7. Yukselturk, E. Predictors for Student Success in an Online Course / E. Yukselturk, S. Bulut // Educational Technology & Society. – 2007. – Vol. 10(2). – P. 71–83.

8. Salal, Y. K. Educational Data Mining : Student Performance Prediction in Academic / Y. K. Salal, S. Abdullaev, Mukesh Kumar // Computer Science, Education. – 2019 – URL: https://www.semanticscholar.org/paper/Educational-Data-Mining-%3A-Student-Performance-in-Salal-Abdullaev/b21fa7245581c3baad2d468cb9d706940de7e010 (дата обращения: 02.01.2024). – Текст : электронный.

9. Karishma B. Bhegade, Student Performance Prediction System with Educational Data Mining / Karishma B. Bhegade, Swati V. Shinde // International Journal of Computer Applications. – 2016. – Vol. 146(5). – P. 32–35.

10. Ивахненко, А. Г. Моделирование сложных систем по экспериментальным данным / А. Г. Ивахненко, Ю. П. Юрачковский. – Москва : Радио и связь, 1987. – 120 с.

11. Ивахненко, А. Г. Самоорганизация прогнозирующих моделей / А. Г. Ивахненко, Й. А. Мюллер. – Киев : Техника, 1985 ; Берлин : ФЕБ Ферлаг Техник, 1984. – 223 с.

12. Ивахненко, А. Г. Индуктивный метод самоорганизации моделей сложных систем / А. Г. Ивахненко. – Киев : Наук. думка, 1981. – 296 с.

13. Комаров, В. С. Модифицированный метод группового учета аргументов и опыт его применения в задачах трехмерной пространственной экстраполяции мезометеорологических полей / В. С. Комаров, А. В. Креминский, Ю. Б. Попов // Метеорология и гидрология. – 1999. – № 8. – С. 235–241.

14. Главчев, Д. М. Программная компонента для поиска решений системы уравнений в частных производных в ГТУ методом группового учета аргументов / В. Д. Дмитрієнко, О. Ю. Заковоротний, С. Ю. Леонов, Д. М. Главчев // Вісник НТУ «ХПІ». – Харків : НТУ «ХПІ», 2019. – №13 (1338). – С. 61–72.

15. Kotsiantis, S. B. Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades / S. B. Kotsiantis // Artificial Intelligence Review. – 2012. – Vol. 37, № 4. – С. 331–344.

16. Ladvanszky, J. A Modification to the Shannon Formula / J. Ladvanszky // Network and Communication Technologies. – 2020. – Vol. 5, № 2. – P. 1–6.

T.V. Yaschun, Y.V. Gromov
Vologda State University

PREDICTING LEARNING ACTIVITY QUALITY USING MACHINE LEARNING METHODS

Approaches to solving the problem of predicting the quality of educational and cognitive activity (ECA) of students based on Educational Data Mining (EDM) methods are analyzed. It is noted that the learning activities of future computer science specialists have their own specifics, since they are inextricably linked with algorithmic activities and interaction with technical devices. It directly affects the selection input parameters of model. A software implementation of one of the machine learning methods – predicting based on the construction of a multifactor regression model based on the method of group accounting of arguments (MGAA) – in the application to the educational program of students of computer specialties is proposed. For the purposes of the research, the expansion of the EDM method with a mechanism of self-regulation of learning activities is proposed.

Educational Data Mining, machine learning, multifactor regression model, educational and cognitive activity, predicting, self-regulation.