УДК 004.048:378.14.015.62

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ОБРАЗОВАТЕЛЬНЫХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ УСПЕШНОСТИ УЧЕБНОЙ ДЕЯТЕЛЬНОСТИ

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

Рассмотрены цели и задачи интеллектуального анализа образовательных данных. Приведены результаты разбора итогов первой экзаменационной сессии, полученные с помощью кластерного анализа, выполненного в среде системы интеллектуального анализа данных WEKA. На основе полученных результатов прогнозируется успешность дальнейшей учебной деятельности. Качество прогноза проверяется путем построения ROC-кривой, выполненного с использованием надстройки AtteStat для табличного процессора MS Excel.

Ключевые слова: *интеллектуальный анализ образовательных данных, прогноз успешности обучения, ROC-анализ.*

Введение. Интеллектуальный анализ образовательных данных (от англ. Educational Data Mining, далее – EDM) – совокупность методов и алгоритмов анализа данных, накапливаемых в учебном заведении в процессе его деятельности с целью выявления скрытых, неочевидных, практически полезных и интерпретируемых знаний об учебном процессе и его участниках для поддержки и принятия решений.

Источниками данных для EDM становятся базы данных университетских систем управления обучением, результаты промежуточных и итоговых аттестаций по дисциплинам, письменные работы студентов, учебная документация, ведущаяся на кафедрах и в деканатах, демографические данные, результаты опросов и анкетирований, социальные сети и т.д.

Интеллектуальный анализ образовательных данных (далее – ИАОД) сравнительно молодое направление научных исследований. Первая международная конференция по EDM прошла в 2008 году в Монреале. С тех пор конференции стали проводится ежегодно. Конференции проходили в Испании, Великобритании, США, Греции. Последняя, 8-я конференция, EDM 2015, была проведена в июне 2015 года в Мадриде, на базе Национального университета дистанционного образования (UNED). С 2010 года издается международный журнал «Educational data mining». С октября по декабрь 2013 года на Интернет-ресурсе «Coursera» (https://www.coursera.org) профессор Колумбийского университета Райан Бейкер (Ryan Baker), один из ведущих специалистов в области ИАОД, провел курс под названием «Big Data in Eduation» [1]. В курсе рассматривались вопросы использования методов математической статистики, машинного обучения и интеллектуального анализа данных в образовании.

Весьма актуальным это научное направление становится для Республики Беларусь, в высшей школе которой идут серьезные реформы.

Цели и задачи ИАОД. Главной целью ИАОД является повышение качества подготовки специалистов. В последнее 10-летие появился ряд исследований, конкретизирующих эту глобальную цель. Так, авторы работы [2] предлагают определять цели применения ИАОД в зависимости от точки зрения конечного пользователя. Они выделяют четыре категории конечных пользователей: обучающиеся, преподаватели, исследователи и администраторы. Цели каждой из этих категорий сведены в таблицу 1.

Таблица 1 – Цели конечных пользователей систем ИАОД

Пользователи	Цели
Обучающиеся	Получить рекомендации по индивидуализации образовательной траектории. Получить более качественную обратную связь с преподавателем. Улучшить успеваемость
Преподаватели	Применять технологии и методы обучения, наиболее подходящие для данной, конкретной группы обучающихся. Улучшить понимание социальных, поведенческих и когнитивных аспектов учебного процесса
Исследователи	Развивать и сравнивать между собой различные методы и алгоритмы ИАОД. Оценивать эффективность и результативность учебного процесса
Администраторы	Принимать обоснованные управленческие решения. Оптимизировать распределение ресурсов учебного заведения

В соответствии с перечисленными целями можно сформировать следующие типовые задачи, решаемые средствами ИАОД.

Для обучающихся. Осознанное формирование индивидуальной образовательной траектории. Правильный выбор факультативных дисциплин и дисциплин по выбору. Профессиональная ориентация и точный выбор сферы будущей профессиональной деятельности.

Для преподавателей. Разделение студентов на кластеры и подбор для каждого кластера оптимальной технологии и наиболее эффективных методов обучения. Оптимизация структуры и содержания лекционного курса. Прогнозирование успешности учебной деятельности.

Для исследователей. Разработка методов объективной оценки эффективности и результативности учебного процесса. Разработка новых технологий и методов обучения. Совершенствование существующих и разработка новых методов и алгоритмов ИАОД.

Для администраторов. Поддержка принятия научно обоснованных управленческих решений.

Продемонстрируем возможности ИАОД на примере решения задачи предсказания успешности учебной деятельности.

Постановка задачи. Имеются результаты сдачи первой экзаменационной сессии некоторой группой студентов. На основе этих данных требуется построить прогноз успешности завершения обучения для этих студентов.

Методика построения прогноза. Проведенный нами анализ базировался на результатах работы со студентами первого курса специальности 1-40 01 01 «Программное обеспечение информационных технологий» 2007 года приема. Кафедра технологий программирования учреждения образования «Полоцкий государственный университет», на которой было выполнено настоящее исследование, является выпускающей по данной специальности. Как видно из таблицы 2, в 2007 году на первом курсе обучалось 46 человек, причем только 15 из них были приняты в счет плана приема, остальные обучались за счет личных средств или на основании трехсторонних договоров с предприятиями.

Итоги первой экзаменационной сессии представлены в таблице 2.

Студенты сдавали экзамены по четырем дисциплинам: «Истории Беларуси», «Высшей математике», «Основам алгоритмизации и программирования» (ОА и П), «Начертательной геометрии и графике» (НГ и Г). Последний столбец таблицы содержит признак успешного завершения обучения в плановый срок в 2012 году («У» – успешное завершение, «ОТЧ» – отчисление в процессе обучения).

Для проведения анализа и построения прогноза мы использовали Data Mining систему WEKA и надстройку AtteStat для табличного процессора MS Excel. Оба программных продукта являются свободно распространяемым программным обеспечением и могут быть загружены из сети Интернет.

Система WEKA позволяет выполнить три вида поиска закономерностей: классификацию, кластеризацию и поиск ассоциаций. Мы остановились на кластеризации, позволяющей разбить исследуемое множество объектов на группы, без каких бы то ни было предварительных условий. При этом при проведении анализа мы учитывали только оценки, полученные студентами в ходе экзаменационной сессии.

Система выделила два кластера: «0», в который вошло двадцать два объекта, и «1», с двадцатью четырьмя объектами. На этом этапе анализа мы решили посмотреть, в какой из кластеров попали студенты, обучающиеся на бюджетной основе. Картина получилась весьма впечатляющая (табл. 3 и 4).

Все 100% студентов-бюджетников оказались в кластере «0». Успеваемость в кластере – 100%, средний балл – 7,25.

Совершенно иначе выглядит ситуация в кластере «1». Сюда попали все студенты, обучающиеся на договорной основе. Семнадцать из них не смогли пройти сессию без двоек, а семь человек получили две и более неудовлетворительные оценки. Средняя успеваемость по кластеру – 29%, средний балл – 4,25.

Основываясь на визуальном анализе полученных результатов, можно сделать вывод о том, что нахождение в кластер «0» позволяет с высокой вероятностью прогнозировать успешное завершение обучения в плановый срок, а кластер «1» является зоной риска и нахождение в нем свидетельствует о большой вероятности отчисления из вуза.

Для проверки этой гипотезы выполним ROC-анализ полученных результатов.

ROC-кривая (от англ. Receiver Operating Characteristic – рабочая характеристика приемника) – график, позволяющий оценить качество бинарной классификации. Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В терминологии ROC-анализа первые называются истинно положительным множеством, вторые – ложно отрицательным. При этом предполагается, что

у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом или точкой отсечения.

Таблица 2 – Итоги первой экзаменационной сессии

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности
200740010100	*4*	*4*	*2*	*5*	***ОТЧ***
200740010101	*4*	*2*	*2*	*2*	***ОТЧ***
200740010102	*4*	*4*	*2*	*4*	***ОТЧ***
200740010103	*9*	*10*	*7*	*5*	***У***
200740010104	*2*	*6*	*2*	*4*	***ОТЧ***
200740010105	*7*	*6*	*7*	*6*	***У***
200740010106	*9*	*9*	*10*	*9*	***У***
200740010107	*4*	*7*	*4*	*6*	***ОТЧ***
200740010108	*6*	*4*	*5*	*4*	***ОТЧ***
200740010109	*6*	*4*	*2*	*2*	***ОТЧ***
200740010110	*4*	*6*	*5*	*4*	***ОТЧ***
200740010111	*4*	*2*	*4*	*5*	***ОТЧ***
200740010112	*9*	*7*	*7*	*4*	***У***
200740010113	*7*	*4*	*5*	*4*	***ОТЧ***
200740010114	*2*	*2*	*4*	*4*	***ОТЧ***
200740010115	*5*	*7*	*10*	*5*	***ОТЧ***
200740010116	*8*	*9*	*5*	*4*	***У***
200740010117	*8*	*6*	*5*	*6*	***У***
200740010118	*9*	*9*	*7*	*6*	***У***
200740010119	*4*	*2*	*6*	*6*	***У***
200740010120	*8*	*8*	*6*	*7*	***У***
200740010121	*6*	*7*	*4*	*4*	***ОТЧ***
200740010122	*7*	*9*	*10*	*5*	***У***
200740010123	*4*	*9*	*8*	*7*	***У***
200740010124	*8*	*9*	*9*	*7*	***У***
200740010125	*5*	*2*	*5*	*4*	***ОТЧ***
200740010126	*8*	*9*	*8*	*9*	***У***
200740010127	*6*	*6*	*6*	*5*	***У***
200740010128	*4*	*6*	*2*	*5*	***У***
200740010129	*9*	*7*	*6*	*6*	***У***
200740010130	*7*	*5*	*4*	*5*	***ОТЧ***
200740010131	*4*	*4*	*2*	*4*	***ОТЧ***
200740010132	*4*	*4*	*4*	*4*	***ОТЧ***
200740010133	*7*	*6*	*7*	*5*	***У***
200740010134	*8*	*8*	*7*	*9*	***У***
200740010135	*4*	*6*	*5*	*6*	***У***
200740010136	*5*	*2*	*4*	*4*	***У***
200740010137	*4*	*2*	*4*	*5*	***ОТЧ***
200740010138	*6*	*2*	*2*	*6*	***У***
200740010139	*6*	*4*	*2*	*6*	***У***
200740010140	*4*	*2*	*4*	*4*	***ОТЧ***
200740010141	*8*	*7*	*8*	*8*	***У***
200740010142	*6*	*6*	*8*	*8*	***ОТЧ***
200740010143	*9*	*9*	*9*	*6*	***У***
200740010144	*8*	*6*	*6*	*7*	***У***
200740010145	*9*	*6*	*4*	*7*	***У***

Таблица 3 – Кластер «0»

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Средний балл	Признак
200740010106	*9*	*9*	*10*	*9*	*9,25*	*У*
200740010126	*8*	*9*	*8*	*9*	*8,5*	*У*
200740010124	*8*	*9*	*9*	*7*	*8,25*	*У*
200740010143	*9*	*9*	*9*	*6*	*8,25*	*У*
200740010134	*8*	*8*	*7*	*9*	*8*	*У*
200740010103	*9*	*10*	*7*	*5*	*7,75*	*У*
200740010118	*9*	*9*	*7*	*6*	*7,75*	*У*
200740010122	*7*	*9*	*10*	*5*	*7,75*	*У*
200740010141	*8*	*7*	*8*	*8*	*7,75*	*У*
200740010120	*8*	*8*	*6*	*7*	*7,25*	*У*
200740010123	*4*	*9*	*8*	*7*	*7*	*У*
200740010129	*9*	*7*	*6*	*6*	*7*	*У*
200740010142	*6*	*6*	*8*	*8*	*7*	*ОТЧ*
200740010112	*9*	*7*	*7*	*4*	*6,75*	*У*
200740010115	*5*	*7*	*10*	*5*	*6,75*	*ОТЧ*
200740010144	*8*	*6*	*6*	*7*	*6,75*	*У*
200740010105	*7*	*6*	*7*	*6*	*6,5*	*У*
200740010116	*8*	*9*	*5*	*4*	*6,5*	*У*
200740010145	*9*	*6*	*4*	*7*	*6,5*	*У*
200740010117	*8*	*6*	*5*	*6*	*6,25*	*У*
200740010133	*7*	*6*	*7*	*5*	*6,25*	*У*
200740010127	*6*	*6*	*6*	*5*	*5,75*	*У*

Таблица 4 – Кластер «1»

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Средний балл	Признак
200740010107	*4*	*7*	*4*	*6*	*5,25*	*ОТЧ*
200740010121	*6*	*7*	*4*	*4*	*5,25*	*ОТЧ*
200740010130	*7*	*5*	*4*	*5*	*5,25*	*ОТЧ*
200740010135	*4*	*6*	*5*	*6*	*5,25*	*У*
200740010113	*7*	*4*	*5*	*4*	*5*	*ОТЧ*
200740010108	*6*	*4*	*5*	*4*	*4,75*	*ОТЧ*
200740010110	*4*	*6*	*5*	*4*	*4,75*	*ОТЧ*
200740010119	*4*	*2*	*6*	*6*	*4,5*	*У*
200740010139	*6*	*4*	*2*	*6*	*4,5*	*У*
200740010128	*4*	*6*	*2*	*5*	*4,25*	*У*
200740010125	*5*	*2*	*5*	*4*	*4*	*ОТЧ*
200740010132	*4*	*4*	*4*	*4*	*4*	*ОТЧ*
200740010138	*6*	*2*	*2*	*6*	*4*	*У*
200740010100	*4*	*4*	*2*	*5*	*3,75*	*ОТЧ*
200740010111	*4*	*2*	*4*	*5*	*3,75*	*ОТЧ*
200740010136	*5*	*2*	*4*	*4*	*3,75*	*У*
200740010137	*4*	*2*	*4*	*5*	*3,75*	*ОТЧ*
200740010102	*4*	*4*	*2*	*4*	*3,5*	*ОТЧ*
200740010104	*2*	*6*	*2*	*4*	*3,5*	*ОТЧ*
200740010109	*6*	*4*	*2*	*2*	*3,5*	*ОТЧ*
200740010131	*4*	*4*	*2*	*4*	*3,5*	*ОТЧ*
200740010140	*4*	*2*	*4*	*4*	*3,5*	*ОТЧ*
200740010114	*2*	*2*	*4*	*4*	*3*	*ОТЧ*
200740010101	*4*	*2*	*2*	*2*	*2,5*	*ОТЧ*

В нашем случае положительным исходом будем считать успешное окончание высшего учебного заведения в установленный срок, а отрицательным исходом – отчисление из вуза в процессе обучения. Количественную интерпретацию ROC-анализа дает показатель AUC (от англ. Area Under ROC Curve – площадь под ROC-кривой) – площадь, ограниченная ROC-кривой и осью доли неверно классифициро-

ванных отрицательных примеров. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию).

В качестве классификатора в нашем анализе был использован средний балл студента. Из таблиц 3 и 4 нетрудно видеть, что пороговым значением является значение 5,75. Студенты, имеющие средний балл больше или равный 5,75, относятся к кластеру «0», а студенты, у которых средний балл ниже 5,75, – к кластеру «1».

ROC-анализ выполнялся в табличном процессоре MS Excel с использованием надстройки AtteStat. Результаты представлены на рисунке.

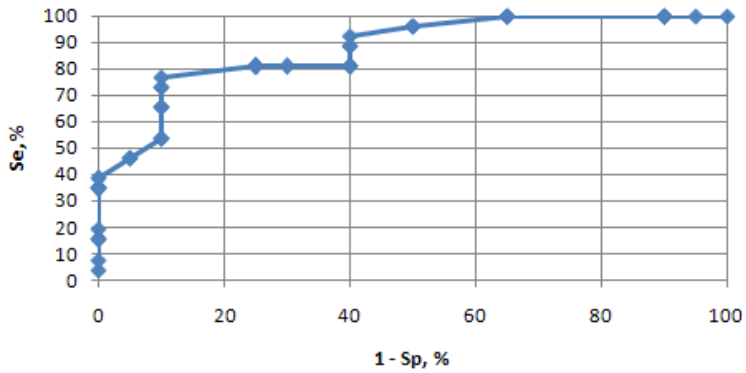

Рисунок – Результаты ROC-анализа

На рисунке **Se –** чувствительность, доля истинно положительных случаев, **Sp –** специфичность, доля истинно отрицательных случаев, правильно идентифицированных моделью.

Показатель AUC = 0,88, что является очень хорошим результатом и убедительно подтверждает выдвинутую нами гипотезу.

Выводы. Результаты первой экзаменационной сессии являются хорошим индикатором учебной деятельности.

Совместное применение ROC- и кластерного анализа позволяет строить эффективный прогноз успешности учебной деятельности.

Для успешного обучения студентов, попадающих в кластер «1», представляется целесообразным:

- выделение студентов, вошедших в этот кластер, в отдельную группу;
- закрепление за этой группой наиболее опытных преподавателей;
- введение в группе обязательной контролируемой самостоятельной работы;
- выполнение анализа и индивидуального планирования самостоятельной работы для каждого студента этой группы с учетом личных склонностей, способностей и возможностей.

Совершенно очевидно, что обязательным условием успешности обучения студентов, вошедших в кластер «1», становятся высокий уровень самодисциплины и высокая мотивация к получению соответствующей профессии.

ЛИТЕРАТУРА

1. Big Data in Education [Электронный ресурс]. – Режим доступа: https://www.coursera.org/course/bigdata-edu. – Дата доступа: 12.01.2016.
2. Romero, C. Data mining in education / C. Romero, S. Ventura // Wiley interdisciplinary reviews. Data mining and knowledge discovery. – 2013. – № 3(1). – P. 12–27.
3. Оськин, А.Ф. Информационно-образовательная среда поддержки управляемой самостоятельной работы студентов / А.Ф. Оськин // Вышэйшая школа. – 2007. – № 5. – С. 67–72.

Поступила 16.03.2016

APPLICATION OF EDUCATINAL DATA MINING FOR PREDICTING OF ACADEMIC SUCCESS

A. OSKIN, D. OSKIN

The goals and objectives of educational data mining are considered. The results of the first examination session, obtained by cluster analysis carried out in the environment of the system of data mining WEKA, are analysed. Further academic success is predicted on the basis of the acquired results. The quality of the prediction is checked by constructing ROC-curve with the help of AtteStat add-in for MS Excel spreadsheet processor.

Keywords: *educational data mining, predicting of academic success, ROC-analysis*.