УДК 65.262.1

*С. Б. Данилевич*

# ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ DATA MINING ДЛЯ АНАЛИЗА ОБРАЗОВАТЕЛЬНЫХ ПРОЦЕССОВ В ВУЗЕ

**Резюме**

У статті розглядаються питання застосування методів та алгоритмів Data Mining для аналізу даних освітньої статистики і наводяться конкретні приклади, що ілюструють роботу деяких алгоритмів Data Mining.

**Summary**

The article considers the application of methods and algorithms of Data Mining for Education Statistics data analysis and provides examples illustrating the work of some algorithms for Data Mining.

**Ключевые слова:** образование, информационные технологии, интеллектуальный анализ данных (ИАД), дерево решений, кластеризация.

Использование информационных и телекоммуникационных технологий является приоритетным направлением развития сферы образования, что отражено в законодательстве Украины, в частности, в Постановлении Кабинета Министров Украины от 7 декабря 2005 г. № 1153 «Об утверждении Государственной программы «Информационные и коммуникационные технологии в образовании и науке на 2006–2010 годы».

Стремительно развивающиеся информационные технологии открывают новые возможности, позволяют упростить решение различных задач, в частности, для принятия эффективных управленческих решений, но и порождают новые проблемы: необходимость обработки множества электронных документов, повышение уровня требовательности к оперативности и достоверности получаемой информации об образовательных учреждениях.

Императивом развития современных средств коммуникации, поиска информации, вычислений, обработки и анализа данных стала интеллектуализация информационных технологий, что привело к появлению новых инструментов анализа данных – компьютерных

программ, позволяющих выявлять в данных закономерности, исследовать их с различных точек зрения [1]. Повысилась доступность информационных технологий для пользователей, имеющих разные уровни компьютерной подготовки.

Все более широкое распространение приобретают методы интеллектуального анализа данных (ИАД, Data Mining), не отрицая, а дополняя традиционные методы анализа информации.

Суть интеллектуального анализа данных заключается в обнаружении в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [2]. Data Mining впитывает достижения статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др.

Применяемые в интеллектуальных системах методы узнавания, связанные с обучением нахождению решающего правила на множестве положительных и отрицательных, определили исследования в области распознавания образов и систем классификации М. М. Бонгарда [3] и его учеников. Первый нейрокомпьютер, способный обучаться в простейших задачах, был построен на перцептроне, нейронной сети, которую разработал Ф. Розенблатт [4]. База для последующего развития нейрокомпьютинга была заложена работами Уоррена Мак-Каллока и Уолтера Питтса [5]. Иерархическим алгоритмам кластерного анализа посвящены работы Р. Люиса, Е. Фикса и Дж. Ходжеса. Г. С. Лбов [6; 7] является автором научного направления: анализ эмпирической информации в классе логических решающих функций от разнотипных переменных. Использование в системах управления принципов, устройств и способов самонастройки, самообучения, распознавания образов и прогнозирования разработала научная школа А. Г. Ивахненко [8]. Большой вклад в область математической статистики и машинного обучения: алгоритм построения решающих деревьев CART, метод бэггинга и метод случайного леса внес Лео Брейман [9–11]. Б. Д. Рипли [12], М. Н. Стоуну [13], А. Н. Колмогорову [14], В. И. Арнольду [15] принадлежит создание основ строгой теории нейронных сетей.

Большой набор статей по методам добычи данных имеется

в журнале «Proceedings from the American Association of Artificial Intelligence Workshops on Knowledge Discovery in Databases published by AAAI Press».

Методы Data Mining применяются и для анализа данных образовательной статистики [16].

Цель данной статьи – рассмотреть вопросы применения методов и алгоритмов Data Mining для анализа данных образовательной статистики на конкретных примерах, иллюстрирующих работу некоторых алгоритмов Data Mining.

В работе использовалась технологическая платформа Deductor Studio Academic 5.2, предназначенная для образовательных целей и позволяющая создавать законченные аналитические решения. В ней сосредоточены современные методы извлечения, манипулирования, визуализации данных, кластеризации, прогнозирования [2] и реализуются функции импорта, обработки, визуализации и экспорта данных. В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки, используя Мастера обработки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты [1].

В качестве исходных взяты данные по 31 частному вузу за 4 года (2007–2010 гг.). Вся информация была представлена в абсолютных величинах. На основе этих данных сформированы следующие относительные показатели:

- число преподавателей на 100 студентов;
- процент докторов;
- процент кандидатов наук;
- процент выпуска аспирантов с защитой диссертации;
- процент студентов I курса;
- процент студентов, участвующих в НИРС;
- процент студентов, участвующих в олимпиадах и конкурсах.

В качестве примера рассмотрим, какие факторы влияют на изменение переменной «Процент выпуска аспирантов с защитой диссертации».

Первым шагом анализа является загрузка данных с помощью мастера импорта. Учитывая ограниченные возможности работы с информацией версии Deductor Academic, данные должны быть преобразованы в текстовый формат и храниться в файле с расширением .txt. При загрузке информации необходимо правильно установить соответствия между входными данными и значениями по умолчанию.

Для выявления дубликатов, противоречий, аномальных значений в загруженных данных используются обработчики:

– дубликаты и противоречия;
– диаграмма;
– фильтрация данных.

Поскольку объекты довольно сильно разняться, то разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных. Кластеризация – это автоматическое разбиение элементов некоторого множества на группы (кластеры) в зависимости от их схожести. В отличие от классификации, когда каждый объект относится к одной из заранее определенных групп, кластеризация разбивает множество объектов на группы, которые определяются только результатом [17]. Под кластером понимают группу объектов [18], обладающих свойством метрической близости (плотность объектов внутри кластера больше, чем вне него).

Оптимальное число кластеров определено автоматически. В результате работы алгоритма оно оказалось равным трем.

В табл. 1 приведены рейтинги кластеров по показателям, на основе которых проводилась кластеризация.

Среднее значение показателя «Процент выпуска аспирантов с защитой диссертации» первого кластера равно 3%, в то время как второго – 21,864%, а третьего – 26,47%. Первый кластер можно исключить из дальнейшего анализа. Для более детального анализа применим кластеризацию к третьему кластеру. В результате работы алгоритма количество подкластеров оказалось равным двум.

Деревья решений (решающих правил, деревьев классификации и регрессии) – один из методов машинного обучения [19]. Deductor позволяет с помощью алгоритма «Дерево решений» автоматически

*Таблица 1*

**Место, занимаемое кластером по каждому показателю**

| Показатель | 1-й кластер | 2-й кластер | 3-й кластер |
|---|---|---|---|
| Процент студентов, участвующих в НИРС | 3 | 1 | 2 |
| Процент студентов, участвующих в олимпиадах и конкурсах | 3 | 1 | 2 |
| Процент выпуска аспирантов с защитой диссертации | 3 | 2 | 1 |
| Процент докторов | 1 | 2 | 3 |
| Процент кандидатов наук | 3 | 1 | 2 |
| Число преподавателей на 100 студентов | 2 | 1 | 3 |

выявить рейтинг факторов, оказавших влияние на целевой атрибут «Процент выпуска аспирантов с защитой диссертации». Деревья решений способны решать задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

Для второго кластера значащим фактором оказался параметр Процент студентов, участвующих в НИРС (рис. 1).

В аналитической платформе Deductor обработчик «Дерево решений» работает только с дискретными значениями, поэтому было принято решение разбить все данные на две категории, в зависимости от того, увеличилось или уменьшилось количество аспирантов с защищенными диссертациями.

| Целевой атрибут: Изменение к-ва аспирантов с защитой диссертации | | | |
|---|---|---|---|
| Номер | Атрибут | Значимость, % | |
| 1 | Процент студентов, участвующих в НИРС | | 100,000 |
| 4 | Процент кандидатов наук | | 0,000 |
| 5 | Число преподавателей на 100 студентов | | 0,000 |
| 2 | Процент студентов, участвующих в олимпиадах и конкурсах | | 0,000 |
| 3 | Процент докторов | | 0,000 |

Рис. 1. Значащий фактор для атрибута «Изменение количества аспирантов с защитой диссертации» для второго кластера

Введя дискретный параметр «Увеличилось» – «Уменьшилось», присвоив ему назначение: «Выходное», получим правило, описывающее критерий параметра «Процент студентов, участвующих в НИРС», при котором «Процент выпуска аспирантов с защитой диссертации» увеличивается или уменьшается (рис. 2).

Это правило подтверждает довольно очевидный факт, что когда достаточное большое число студентов занимается наукой, то вероятней всего они поступят в аспирантуру и защитят диссертации.

Для второго подкластера третьего кластера значащим фактором оказался параметр «Процент докторов» (рис. 3).

Для первого подкластера третьего кластера значащими факторами оказались параметры (рис. 5): процент студентов, участвующих в НИРС; процент докторов; процент кандидатов наук.

Для более детального анализа применим кластеризацию к первому подкластеру третьего кластера (кластер 31). В результате работы алгоритма количество подкластеров оказалось равным двум.

Первый из них (кластер 311) выделил вузы в те годы, когда процент аспирантов с защитой диссертации уменьшился, а второй (кластер 312) – когда процент аспирантов с защитой диссертации увеличился. Результатом работы Обработчика «Дерево решений» явились правила (см. табл. 2, 3).

Средние значения параметров кластеров 311 и 312 рассчитаны программой Deductor автоматически и представлены в таблице 4.

Из данных таблицы 4 можно сделать вывод, что для рассматриваемых кластеров решающими для параметра «Изменение количества аспирантов с защитой диссертации» оказались факторы «Процент студентов, участвующих в НИРС» и «Процент студентов, участвующих в олимпиадах и конкурсах».

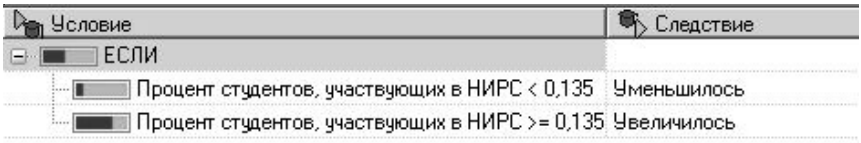


Рис. 2. Правило «Дерева решений»

| Целевой атрибут: Изменение к-ва аспирантов с защитой диссертации | | |
|---|---|---|
| Номер | Атрибут | Значимость, % |
| 3 | Процент докторов | 100,000 |
| 4 | Процент кандидатов наук | 0,000 |
| 5 | Число преподавателей на 100 сту... | 0,000 |
| 2 | Процент студентов, участвующих в... | 0,000 |
| 1 | Процент студентов, участвующих в... | 0,000 |

Рис. 3. Значащий фактор целевого атрибута «Изменение количества аспирантов с защитой диссертации» для второго подкластера третьего кластера

| Номер правила | Условие | | | Следствие |
|---|---|---|---|---|
| | Показатель | Знак | Значение | ab Изменение к-ва аспира |
| 1 | **9.0** Процент докторов | < | 0,07 | Уменьшилось |
| 2 | **9.0** Процент докторов | >= | 0,07 | Увеличилось |
| | **9.0** Процент докторов | < | 0,085 | |
| 3 | **9.0** Процент докторов | >= | 0,07 | Уменьшилось |
| | **9.0** Процент докторов | >= | 0,085 | |
| | **9.0** Процент докторов | < | 0,095 | |
| 4 | **9.0** Процент докторов | >= | 0,07 | Увеличилось |
| | **9.0** Процент докторов | >= | 0,085 | |
| | **9.0** Процент докторов | >= | 0,095 | |

Рис. 4. Решающие правила для целевого атрибута «Изменение количества аспирантов с защитой диссертации»

| Целевой атрибут: Изменение к-ва аспирантов с защитой диссертации | | |
|---|---|---|
| Номер | Атрибут | Значимость, % |
| 1 | Процент студентов, участвующих в... | 49,049 |
| 3 | Процент докторов | 27,829 |
| 4 | Процент кандидатов наук | 23,122 |
| 5 | Число преподавателей на 100 сту... | 0,000 |
| 2 | Процент студентов, участвующих в... | 0,000 |

Рис. 5. Значащие факторы целевого атрибута «Изменение количества аспирантов с защитой диссертации» для первого подкластера третьего кластера

*Таблица 2*

**Правило Обработчика «Дерево решений» для кластера 311**

| № | Условие | Следствие (Изменение к-ва аспирантов с защитой диссертации) | Поддержка | | Достоверность | |
|---|---|---|---|---|---|---|
| | | | % | Кол-во | % | Кол-во |
| 1 | | Уменьшилось | 100,00 | 28 | 67,86 | 19 |

*Таблица 3*

**Правило Обработчика «Дерево решений» для кластера 312**

| № | Условие | Следствие (Изменение к-ва аспирантов с защитой диссертации) | Поддержка | | Достоверность | |
|---|---|---|---|---|---|---|
| | | | % | Кол-во | % | Кол-во |
| 1 | | Увеличилось | 100,00 | 15 | 60,00 | 9 |

*Таблица 4*

**Средние значения кластеров 311, 312**

| Показатель | Средние значения кластера 311 | Средние значения кластера 312 |
|---|---|---|
| Процент студентов, участвующих в НИРС | 6,00% | 6,94% |
| Процент студентов, участвующих в олимпиадах и конкурсах | 0,28% | 1,31% |
| Процент выпуска аспирантов с защитой диссертации | 23,24% | 28,94% |
| Процент докторов | 7,41% | 4,31% |
| Процент кандидатов наук | 47,34% | 34,88% |
| Число преподавателей на 100 студентов | 5,52% | 4,81% |

Рассмотренный пример представляет собой иллюстрацию процесса анализа данных для принятия обоснованных управленческих решений в сфере образования.

Полученные результаты свидетельствуют, что в вузах, где уделяется повышенное внимание организации научной работы студентов, имеющих в составе достаточное количество докторов и кандидатов наук, число аспирантов с защитой диссертации увеличилось.

Deductor Studio является полезным инструментом при обработке многомерных массивов данных и позволяет получать весьма ценную информацию особенно там, где применение классических методов затруднительно. Применение технологий Data Mining помогает совершенствовать прогнозирование образовательного процесса в вузе.

## Список литературы

1. Кислова О. Н. Интеллектуализация информационных технологий как фактор развития интеллектуального анализа социологических данных / О. Н. Кислова // Методологія, теорія та практика соціологічного аналізу сучасного суспільства : зб. наук. пр. – Х. : Вид. центр ХНУ імені В. Н. Каразіна, 2009. – С. 318–324.

2. Deductor Studio Academic [Электронный ресурс]. – Режим доступа: http://www.basegroup.ru/download/deductor/.

3. Бонгард М. М. Проблема узнавания / М. М. Бонгард. – М. : Наука, 1967. – 320 с.

4. Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга / Ф. Розенблатт.– М. : Мир, 1965. – 480 с.

5. McCulloch W. S. A logical calculus of the ideas immanent in nervous activity / W. S. McCulloch and W. Pitts // Bull. Math. Biophys. – 1943. – № 5. – P. 115–133.

6. Лбов Г. С. Логически решающие функции и вопросы статистической устойчивости решений / Г. С. Лбов, Н. Г. Старцева. – Новосибирск : Изд-во Ин-та математики, 1999. – 212 с.

7. Лбов Г. С. Методы обработки разнотипных экспериментальных данных / Г. С. Лбов. – Новосибирск : Наука, 1981. – 160 с.

8. Ивахненко А. Г. Самообучающиеся системы с положительными обратными связями : справ. пособие / А. Г. Ивахненко. – К. : Изд-во АН УССР, 1963.

9. Breiman L. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, and C. Stone. – Wadsworth, Belmont, CA, 1984.

10. Breiman L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – № 24 (2). – P. 123–140.

11. Breiman L. Random Forests / L. Breiman // Machine Learning. – 2001. – № 45 (1). – P. 5–32.

12. Ripley B. D. Pattern Recognition and Neural Networks / B. D. Ripley. – Cambridge University Press. – 1996. – January.

13. Stone M. N. The generalized Weierstrass approximation theorem / M. N. Stone // Mathem. Mag.– 1948.– V. 21.– P. 167–183, P. 237–254.

14. Колмогоров А. Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного / А. Н. Колмогоров // Докл. АН СССР. – 1957. – Т. 111. – № 5. – С. 953–966.

15. Арнольд В. И. О представлении непрерывных функций нескольких переменных в виде суперпозиции функций меньшего числа переменных / В. И. Арнольд // Мат. просвещение. – 1957. – № 19. – С. 41–61.

16. Бершадский А. М. Применение методов Data Mining для анализа данных образовательной статистики / А. М. Бершадский, А. А. Гудков // Труды XIV Всерос. науч.-метод. конф. «Телематика'2007». – СПб., 2007. – С. 382–384.

17. Статистический словарь. – М. : Финансы и статистика, 1989. – 623 с.

18. Дидэ Э. Методы анализа данных: Подход, основанный на методе динамических сгущений : пер. с фр. / Э. Дидэ. – 1985.

19. Тоби Сегаран. Программируем коллективный разум / Тоби Сегаран. – Изд-во: Символ. – 2008. – 368 с.