УДК 004.89, 004.048

И.В. Шуба

Национальный технический университет "ХПИ", Харьков

ИСПОЛЬЗОВАНИЕ МЕТОДОВ DATA MINING ПРИ АНАЛИЗЕ СОЦИАЛЬНЫХ ЯВЛЕНИЙ

В статье проанализирована возможность применения интеллектуальных методов обработки данных при анализе социальных явлений. Приведен пример обработки данных анкет студентов пятого курса различных специальностей, используя аналитическую платформу Deductor. Были выявлены факторы, оказывающие влияние на «удовлетворенность» выбранной специальностью пятикурсников.

Ключевые слова: *интеллектуальные методы анализа данных, нейронная сеть, самоорганизующиеся карты Кохонена, прогнозирование, выбор специальности*

Введение

Постановка проблемы. Одним из важных выборов человека, который он принимает в жизни, является выбор профессии, т.к. за ним за ним стоит ответственность за судьбу и способность принимать судьбоносные решения [1]. Такой выбор возникает перед школьниками-выпускниками при выборе ВУЗа и специальности для поступления, и часто они оказываются не готовыми психологически и социально к принятию решения. Таким образом, как по объективным, так и по субъективным причинам, возникает проблема «удовлетворенности» выбранным ВУЗом и специальностью. Выявление этих причин и зависимостей с помощью интеллектуальных средств обработки данных легли в основу данных исследований.

Анализ последних исследований и публикаций. Проблема анализа социальной информации с целью выявления закономерностей, построения моделей и прогнозов развития общества были ориентиром и двигающей силой при создании специальных механизмов обработки и анализа данных в виде программных продуктов. Так, методы Data Mining успешно используются для решения бизнес-задач в банковском деловодстве, страховании, промышленном производстве, маркетинговых исследованиях, торговле, в научных исследованиях и других.

Достаточно ёмкое определение Data Mining предложено одним из основателей направления Г. Пиатецки-Шапиро – это процесс обнаружения в сырых данных ранее не известных, нетривиальных, практически полезных и доступных интерпретаций знаний, необходимых для принятия решений в различных сферах.

К задачам Data Mining относятся [2 – 5]:

1. Классификация. Заключается в упорядочении по некоторому принципу множества объектов, у которых сходны квалификационные признаки, которые выбраны для определения сходства либо различия между объектами сравнения.

2. Прогнозирование. Метод, при котором используются как накопленный в прошлом опыт, так и текущие допущения насчет будущего с целью его определения.

3. Кластеризация. Заключается в выявлении классов объектов изначально не предопределенных и в конечном результате разбиении объектов на классы.

4. Последовательность – это установление закономерностей между событиями, которые связаны во времени.

5. Оценивание – это предсказание непрерывных значений признака.

6. Анализ связей – это нахождение зависимостей в наборе данных.

7. Визуализация – использование графических методов, показывающих наличие закономерностей.

Рынок программного продукта в настоящее время представлен универсальными статистическими пакетами, которые оснащены набором методов интеллектуального анализа данных, например, SPSS (SPSS, Clementine), Statistica (StatSoft), SAS Institute (SAS Enterprise Miner), Deductor (Base Group Labs). В работе для проведения анализа использована аналитическая платформа Deductor.

Основные алгоритмы Data Mining в Deductor представлены набором:

а) нейронные сети;

б) самоорганизующиеся карты Кохонена;

в) автокорреляция;

г) деревья решений;

д) ассоциативные правила.

Нейронные сети относятся к классу нелинейных адаптивных систем с архитектурой, условно имитирующей нервную ткань, состоящую из нейронов [6].

Самоорганизующиеся карты Кохонена – одна из разновидностей нейронных сетей, отличием которых, является используемое неконтролируемое обучение.

Деревья решений — метод структурирования задачи в виде древовидного графа, вершины которого соответствуют продукционным правилам, позволяющим классифицировать данные или осуществлять анализ последствий решений.

Автокорреляция - заключается в расчете выборочной корреляции.

Ассоциативные правила – заключается в нахождении закономерностей между связанными событиями, например, часто покупатель приобретает не один товар, а несколько сопутствующих, между которыми в большинстве случаев есть взаимосвязь.

Целью статьи является анализ возможности применения базовых существующих алгоритмов интеллектуальной обработки данных при выявлении причин «неудовлетворенности» выбором специальности и взаимосвязи между причинами и следствием.

Основная часть

В качестве исходной базы данных использовалась информация, полученная с помощью анкет, разработанных автором статьи, которые заполняли студенты пятого курса разных факультетов и специальностей НТУ «ХПИ». Анкета состояла из 32 вопросов, в опросе приняли участие 200 студентов, собранная информация анализировалась с помощью Deductor 5.2, используя нейронные сети и самоорганизующиеся карты Кохонена. Алгоритм проведения анализа приведен на рис. 1.

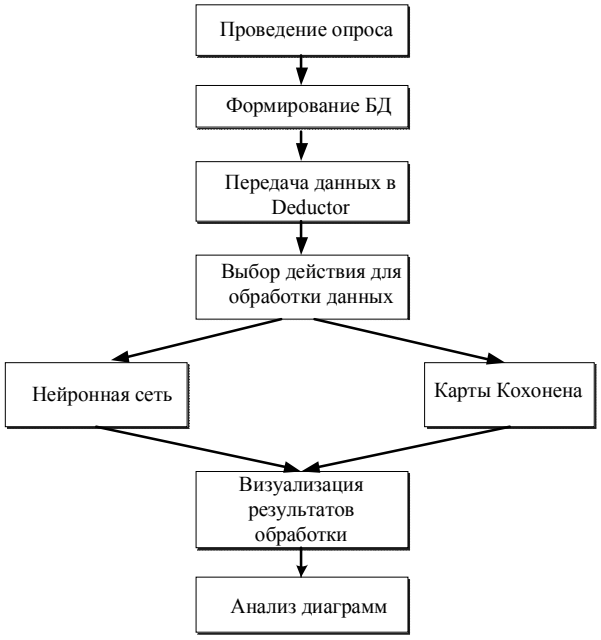

Рис. 1. Алгоритм проведения исследования

Прежде чем приступить к использованию алгоритмов в Deductor, необходимо произвести подготовку набора анализируемых данных. Поскольку такой метод анализа может обнаружить только присутствующие в данных закономерности, исходные данные с одной стороны должны иметь достаточный объем, чтобы эти закономерности в них присутствовали, а с другой — быть достаточно компактными, чтобы анализ выполнялся оперативно. Сначала осуществляют очистку данных, которые после обработки сводятся к наборам признаков (или векторам, если алгоритм может работать только с векторами фиксированной размерности).

Из всех вопросов анкеты для определения «удовлетворенности» выбором специальности и ВУЗа были взяты восемь (рис. 2):

1) пол;
2) возраст;
3) факторы, оказывающие влияние на выбор работы;
4) причина поступления в ВУЗ;
5) цель поступления в ВУЗ;
6) количество рассматриваемых ВУЗов при поступлении;
7) вероятность найти работу по специальности;
8) причина выбора специальности.

В качестве вопроса определяющего удовлетворенность, был взят вопрос анкеты «хотели бы Вы учиться в другом ВУЗе или поменять специальность». Предложено три варианта ответа: 1 - да; 2 - нет; 3 - затрудняюсь ответить. Если респондент отвечал, что хочет поменять ВУЗ и/или специальность, значит, он не доволен своим выбором пять лет назад при поступлении в ВУЗ. Все данные были оцифрованы для корректной работы программы.

Дальнейшим этапом является построение модели с целью исследования моделируемого объекта и получения новых знаний, необходимых для принятия решений. Аналитик создает модель, так как аналитическая платформа не может заменить человека и не может ответить на вопросы, которые н были заданы, и, как отмечалось в работе [6], она всего лишь дает ему мощный инструмент для облегчения и улучшения его работы. Deductor позволяет сконструировать нейронную сеть с заданной структурой, изучить ее параметры и обучить с помощью алгоритма обучения. В результате будет получен эмулятор нейронной сети, который будет использоваться для решения задач анализа и прогнозирования, классификации, поиска скрытых закономерностей и др. В работе использовалась нейронная сеть (рис. 3), где в качестве исходящего поля выбрано «хотели бы поменять ВУЗ и/или специальность», остальные поля входящие.

Анализ нейронной сети позволил выявить зависимости.

Так, например, при анализе фактора «Причина поступления в ВУЗ», где варианты ответов соответствуют цифровым значениям:

Рис. 2. Логическая модель построения БД

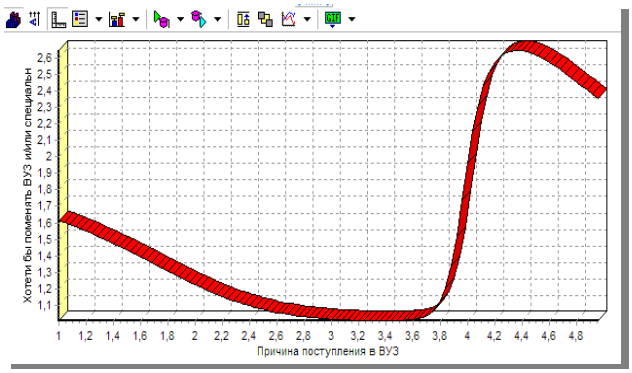

Рис. 3. Экран модели социального явления – выбор профессии (нейронные сети). Зависимость «удовлетворенности» выбором специальности от причин поступления в ВУЗ

1 - пример родителей;
2 - пример товарищей;
3 - настоятельные требования родителей;
4 – самостоятельное решение;
5 – другое,

четко видно зависимость «удовлетворенности» студента выбором специальности от этого показателя (рис. 3).

«Неудовлетворенность» присуща тем студентам, которые пошли получать высшее образование и/или выбранную специальность по «настоятельным требованиям родителей», а те которые «самостоятельно принимали решение» довольны своим выбором и не хотят ничего менять.

Это свидетельствует о том, что родители, опираясь на свой жизненный опыт, а также пользуясь авторитетом и влиянием на детей, выбирают ВУЗ и специальность на свое усмотрение, не особо учитывая пожелание и способности детей. Вследствие чего студент остается «нереализованным», у него отсутствует желание, интерес и стремление к учебе.

В настоящее время у абитуриентов есть возможность подавать документы на 15 специальностей (в 5 ВУЗов на 3 специальности в каждом) [7]. С одной стороны это дает возможность снизить риск не поступления с первого раза в ВУЗ, но с другой абитуриенты, пользуясь этой возможностью, подходят не осознано к важному решению и, пройдя по конкурсу в различных ВУЗах, останавливаются на более престижном ВУЗе/специальности, пренебрегая возможностью самореализации в той профессии, к которой у него способности. Это подтверждают данные, представленные на рис. 4.

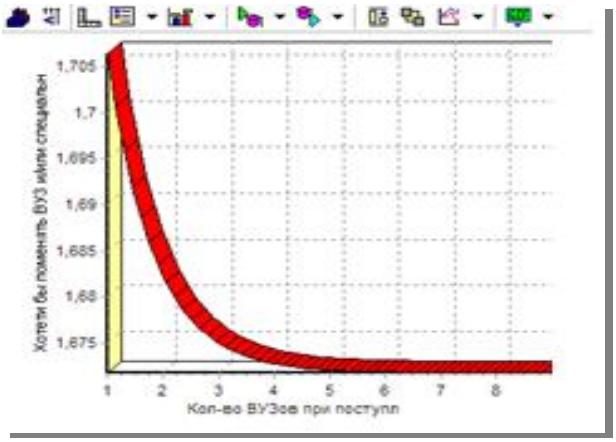

Рис. 4. Зависимость «удовлетворенности» выбором ВУЗа/специальности от количества ВУЗов при поступлении

К факторам, оказывающим влияние на осознанность в подходе к выбору будущей профессии, можно отнести возраст абитуриента при поступлении, что подтверждается результатами изучения нейронной сети (рис. 6).

Эти три фактора можно отнести к неосознанным или сделанным под влиянием.

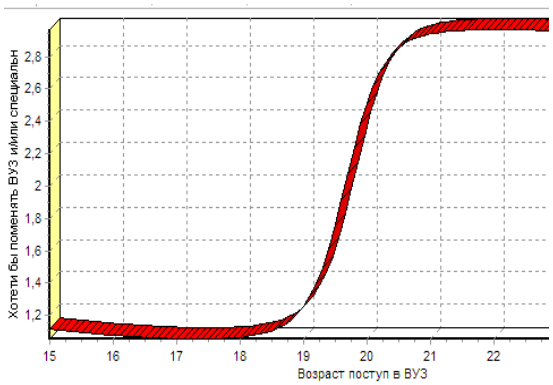

Рис. 5 Влияние возраста абитуриента на правильность выбора ВУЗа/специальности

В работе рассматривалось также влияние факторов «почему выбрали данную специальность», «вероятность найти работу по специальности» и «факторы, оказывающие влияние при выборе места работы».

Результаты анализа представлены на рис. 6 - 8.

Так, неосведомленность о будущей специальности приводит к ошибочному принятию решения относительно поступления, не всегда название специальности раскрывает суть получаемой квалификации и области знаний, возможности дальнейшего трудоустройства (рис.6). Хотели бы поменять ВУЗ/специальность те кто «ничего не знал о специальности до поступления», что соответствует числовому значению 5 и те, кто руководствовался «советом родителей» - 4.

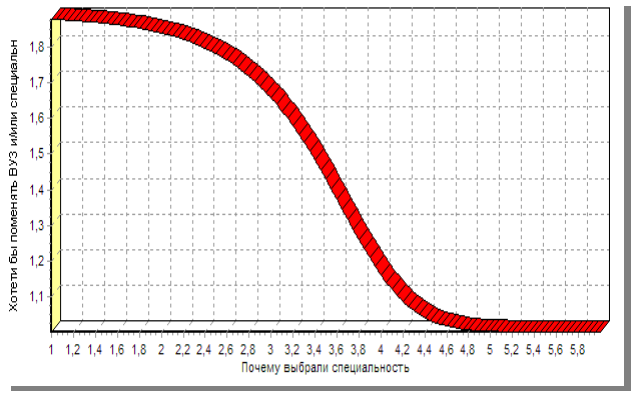

Рис. 6 Влияние причин выбора специальности на «удовлетворенность» специальностью

В анкете были предложены факторы, играющие роль при выборе работы:

1 – высокая заработная плата;

2 - интересная работа;

3 – руководящая должность;

4 – близкое расположение к дому;

5 – затрудняюсь ответить.

При анализе данных было выявлено, что те студенты, которые не могут сформулировать требования к будущей работе, т.е. «затрудняются ответить» (рис. 7) не удовлетворены своей специальностью.

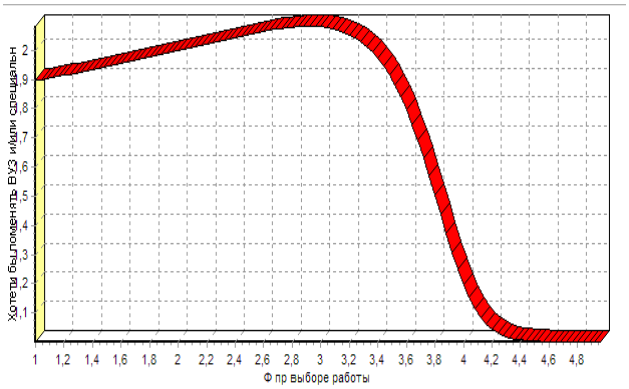

Рис. 7 Зависимость «удовлетворенности» выбора ВУЗа/специальности от требований, предъявляемых к будущей работе

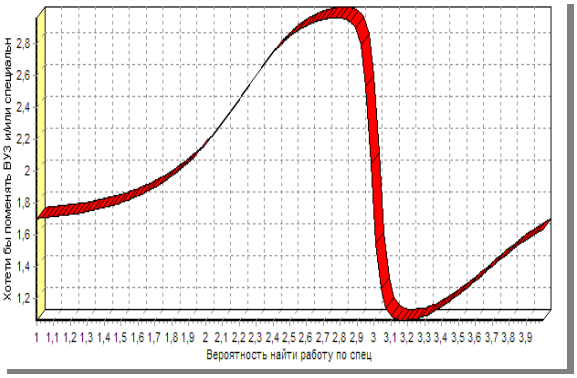

Рис. 8 Влияние на «удовлетворенность» выбора ВУЗа/специальности

На рассматриваемый фактор также оказывает существенное влияние «вероятность найти работу по специальности».

К сожалению, в настоящее время большинство отраслей промышленности находятся на грани выживания, поэтому и востребованность специалистов с техническим образованием значительно снизилась. Это не зависящий от студента фактор, но оказывающий влияние на «удовлетворенность» выбором при поступлении (рис.8), т.к. хотели бы его изменить те, кто считает, что нет «никакой» (соответствует цифре 3 на графике) возможности найти работу по специальности. Как показывает исследование, большинство студентов планируют получать второе высшее образование экономическое, юридическое, что расширит возможности трудоустройства, имея базовое техническое образование.

Данные обработки БД с помощью нейронной сети подтверждаются и алгоритмом самоорганизующиеся карты Кохонена (рис.9) из пяти кластеров, которые характеризуют входящие поля:

возраст поступления в ВУЗ;

причина поступления в ВУЗ;

кКоличество ВУЗов при поступлении;

причина выбора специальности;

возможность найти работу по специальности.

Рис. 9. Карты Коханена для анализа выбора ВУЗа/специальности

Таким образом, полученные результаты исследований, посвященных проблеме «удовлетвореннности» студентов выбранным ВУЗом/специальностью с помощью применения интеллектуального метода обработки данных позволяют сделать вывод о том, что большинство молодых людей при выборе будущей профессии руководствуются не собственными желаниями и предрасположенностями, а осуществляют выбор под влиянием сторонних факторов.

Выводы

В статье рассмотрена возможность применения методов Data Mining при изучении социальных явлений на примере выбора ВУЗа/специальности абитуриентом при поступлении. Приведен пример обработки данных с помощью нейронных сетей и самоорганизующихся карт Кохонена, выявления факторов, оказывающих существенное влияние на «удовлетворенность» выбором.

Темой дальнейших исследований является применение описанных методов для выявления факторов, влияющих на успеваемость, желание учиться студентов с целью последующей их мотивацией к качественному получению знаний.

Список литературы

1. Смирнов С.М. Педагогика и психология высшего образования от деятельности к личности / С.М. Смирнов. – М.: Академия, 2005. – 400 с.

2. Wang J. Data Mining: Opportunities and Challenges / J. Wang (Ed.). Hershey: Idea Group Publishing, 2003. – 468 p.

3. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б. Паклин, В.И. Орешков. – СПб.: Питер, 2009. – 624 с.

4. Шумейко А.А., Сотник С.Л. Интеллектуальный анализ данных / А.А. Шумейко, С.Л. Сотник. – Днепропетровск: Белая, 2012. – 212 с.

5. Макленнен Дж. Microsoft SQL Server 2008: Data mining – Интеллектуальный анализ данных. Пер. с англ. / Дж. Макленнен, Чж. Танг, Б. Криват. – БХВ-Петербург, 2009. – 720 с.

6. Клименко А.Н. Использование интеллектуальных методов анализа при обосновании разработки суперкомпьютера / А.Н. Клименко, Н.Ю. Любченко, А.А. Подорожняк // Системи обробки інформації, - Х. : ХУПС, 2012. – Вип. 4(27). – С. 105-109.

7. Сайт «Вступна компанія» [Электронный ресурс]. – Режим доступа к материалам сайта : http://osvita.ua/questions/vnz/consultations/25.

Надійшла до редколегії 19.05.2014

Рецензент: д-р техн. наук, ст. наук. співр. С.Г. Семенов, Національний технічний університет "ХПІ", Харків.

ВИКОРИСТАННЯ МЕТОДІВ DATA MINING ПРИ АНАЛІЗІ СОЦІАЛЬНИХ ЯВИЩ

І.В. Шуба

У статті проаналізовано можливість використання інтелектуальних методів обробки даних при аналізі соціальних явищ. Наведено приклад обробки даних анкет студентів п'ятого курсу різних спеціальностей, використовуючи аналітичну платформу Deductor. Були виявлені чинники, що впливають на «задоволеність» вибраною спеціальністю п'ятикурсників.

Ключові слова: *інтелектуальні методи аналізу даних, нейронна мережа, самоорганізовані карти Кохонена, прогнозування, вибір спеціальності.*

USE OF METHODS OF DATA MINING IS FOR ANALYSIS OF THE SOCIAL PHENOMENA

I.V. Shuba

In the article possibility of application of intellectual methods of processing of data is analysed at the analysis of the social phenomena. The example of treatment of these questionnaires of students of fifth course of different specialities is resulted, using the analytical platform of Deductor. Factors were exposed having influence on «satisfaction» by the chosen speciality of 5-years education.

Keywords: *intellectual methods of analysis of data, neuron network, selfgettings organized karty Kokhonena, prognostication, choice of speciality.*