ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ ПРИ ИЗУЧЕНИИ ДИСЦИПЛИН ЕСТЕСТВЕННО-МАТЕМАТИЧЕСКОГО ЦИКЛА СТУДЕНТАМИ МЕДИЦИНСКОГО ВУЗА

Intellectual data processing technologies in the study of natural-mathematical cycle disciplines by students of a medical university

Дорошина Наталья Владимировна, *старший преподаватель кафедры математики, физики и медицинской информатики Рязанского государственного медицинского университета имени академика И.П. Павлова, ФГБОУ ВО РязГМУ Минздрава России.*

ndoroshina@mail.ru

Дмитриева Мария Николаевна, *кандидат педагогических наук, доцент кафедры математики, физики и медицинской информатики Рязанского государственного медицинского университета имени академика И.П. Павлова, ФГБОУ ВО РязГМУ Минздрава России.*

dmitrm05@mail.ru

Кабанов Анатолий Николаевич, *кандидат технических наук, доцент кафедры АСУ Рязанского государственного радиотехнического университета, ФГБОУ ВО РГРТУ.*

info@schoolfut.ru

В статье показана необходимость преподавания технологии интеллектуальной обработки экспериментальных данных в курсе «Информатика, медицинская информатика и статистика» для студентов-первокурсников медицинского вуза с применением современных технологий Data Mining. Выявляются сложности в усвоении статистических методов студентов-медиков. Показано, что использование реальных данных, полученных экспериментальным путем, повышает интерес студентов к новым технологиях.

The paper presents the intellectual data processing sequential algorithm in teaching of course "Informatics, medical informatics and statistics" to first-year students of medical university with using of modern Data Mining technologies. Difficulties in statistical methods assimilation of medical students are revealed. Using real data from experiments increases students' interest in new technologies.

Ключевые слова: **студенты-гуманитарии, аналитический пакет Deductor, технологии Data Mining, интеллектуальный анализ данных, визуализация результатов.**

Keywords: **students in the humanities, analytical programs, Data Mining technologies, intellectual data analysis, visualization of results.**

Традиционные методы математической статистики долгое время претендовали на роль основного инструмента анализа данных. Но они оказались полезными главным образом для проверки заранее сформулированных гипотез и для “грубого” разведочного анализа, составляющего основу оперативной аналитической обработки данных. Сейчас проблема устойчивости анализа данных в любой отрасли выдвигается в число важнейших. Это связано с развитием исследований в условиях неопределенности и наличия сбойных результатов экспериментов. Результаты таких расчетов во многом зависят от выбора метода обработки данных, характеристик устойчивости методов по отношению к нарушениям исходных предпосылок метода. А реально исходные предпосылки метода могут соблюдаться лишь приближенно. В связи с этим возникла проблема огрубления, стабилизации алгоритмов обработки данных при их реализации на ЭВМ [1], а значит, применения новых технологий анализа данных.

Говоря о преподавании методов математической статистики в медицинских вузах, следует отметить, что статистическая обработка результатов медико-биологических экспериментов и данных повседневной медицинской практики сложна, многокомпонентна в силу различных факторов: многомерность данных, неполнота, разнородность, наличие аномальных результатов и т.д., а также их возможная конфиденциальность и неоднозначность. Она тяжела для понимания студентами медицинских вузов, особенно первокурсниками. Люди с гуманитарным образованием испытывают серьезные затруднения при освоении математической статистики, являющейся отраслью высшей математики, и основ теории вероятности, идущей рука об руку с математической статистикой. Для исправления положения необходимо предложить некоторый инструмент, который позволит верно выполнить статистические расчеты [2]. Под инструментом здесь понимается метод обработки данных и программный пакет для проведения расчетов. Поэтому при построении курса «Информатика, медицинская информатика и статистика» для студентов 1 курса специальности 32.05.01 «Медико-профилактическое дело» выбор такого инструмента выдвигалось в число первоочередных задач. Обучение дисциплине происходит по соответствующим модулям: «Информатика», «Медицинская информатика» и «Статистика». Последний включает в себя следующие традиционные разделы: выборочная и генеральная совокупность, оценки числовых и интервальных характеристик выборки, корреляционный анализ, регрессионный анализ, проверка статистических гипотез, анализ временных рядов и прогнозирование, кластерный анализ, факторный анализ, представление многомерных данных.

Основная часть практических занятий в рамках модуля «Статистика» проходит в виде выполнения лабораторных работ в программе MS Excel с использованием «Пакета анализа» и инструментария формул. Студентам объясняется преимущество использования MS Excel при решении статистических задач: простота использования, доступность программного обеспечения, частичная визуализация результатов, экономия времени [3].

Но данная программа не обладает возможностями по реализации многих современных алгоритмов обработки данных. Поэтому изучение некоторых тем, например, кластерного анализа, происходит в контексте параллельного знакомства обучающихся с современными технологиями Data Mining, расширяющими возможности классических статистических методов.

Data Mining (или интеллектуальный анализ данных) – это современная концепция, включающая совокупность различных методов обнаружения полезных знаний из большой совокупности данных. Вот некоторые из них: классификация событий и ситуаций по совокупностям признаков, кластеризация, нейронные сети, выявление взаимозависимостей, причинно-следственных связей, ассоциаций и аналогий, прогнозирование и т.д. Потребность практического использования в гуманитарной сфере методов интеллектуальной обработки данных приводит к необходимости построения и соответствующей адаптации их обобщенных математических и алгоритмических моделей и созданию оригинальных методик их применения и обучения. Создание механизма повышения эффективности отдельных методов интеллектуальной обработки данных выдвигает задачу разработки процедуры, обеспечивающей их адаптацию к различным отраслям гуманитарного применения [4].

Широкое распространение эти методы получили в медицинских исследованиях [5]. [2] считает, что целесообразно развить курс статистической обработки данных в направлении представлений о классификационных методах и кластерном анализе. Рассмотрим задачу кластерного анализа многомерных данных подробнее.

Формулировка задачи. Данные медицинских исследований являются в большинстве случаев многомерными, например, когда состояние каждого пациента характеризуется множеством параметров и сведены в таблицу «Объект-свойство» («Пациент-симптомы»). На первом этапе анализа данных формирование отношений между объектами (пациентами) практически невозможно. Определение связей между объектами сильно облегчается, если исходное множество всех объектов удается описать более кратким способом, чем перечисление всех

объектов со всеми их свойствами. Наиболее распространенный способ сокращения описания связан с разделением множества M объектов таблицы на небольшое число групп, связанных друг с другом каким-нибудь закономерным свойством. Обычно в качестве такого свойства используется «похожесть» объектов одной группы. Закономерности «групповой похожести» позволяют намного сократить описание таблиц «Объект-свойство» при малой потере информации. Вместо перечисления всех объектов можно дать список «типичных» или «эталонных» представителей групп, указать номера (имена) объектов, входящих в состав каждой группы. При небольшом числе групп описание данных становится обозримым и легко интерпретируемым. В этом состоит типичная задача кластерного анализа – разбиение объектов на однородные группы (кластеры) и определение их центров.

Студентам-первокурсникам медицинского вуза целесообразно предлагать решение подобных задач непосредственно на данных, хорошо знакомых и понятых студентами, а также использовать специальные аналитические пакеты, которые имеют мощные алгоритмы обработки данных, наглядны и несложны в использовании [6, 7].

В качестве одного из них выбран Deductor Academic – аналитическая платформа, основа для создания законченных прикладных решений в области анализа информации на основе современных технологий Data Mining and Knowledge Discovery (DM&KD) – «добыча» данных и обнаружение знаний. Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от консолидации данных до построения моделей и визуализации полученных результатов. Эта программа распространяется бесплатно, не нуждается в профессиональной установке и имеет небольшой объем. Основное внимание в нашей работе направляется на последовательное применение методов анализа данных на основе сценарного подхода, реализованного в Deductor.

Следует отметить, что Deductor имеет колоссальное применение в преподавании специальных дисциплин для студентов экономических специальностей [8].

Имеются однородные данные реальных исследований, представленных многомерным массивом данных, включающих результаты тестовых испытаний группы студентов на скорость сенсомоторной реакции по шести параметрам (рис.1). Здесь количество строк равно количеству тестируемых (10), а количество столбцов – количеству параметров (6). Для получения данных использовалась программа «Исследования сенсомоторной реакции пользователя ПЭВМ», в которой фиксируется:

– скорость простой сенсомоторной реакции - при появлении области с цифрой «1» на экране монитора нужно нажать «1» на клавиатуре (x1); при появлении области «1» на экране нужно кликнуть мышью на эту область (x2);
– скорость сложной сенсомоторной реакции – при появлении области «1» или «2» нужно нажать соответствующий номер на клавиатуре (x3); при появлении области синего или красного цвета нужно нажать «1» на клавиатуре, если это область синего цвета и «2», если это область красного цвета (x4); при появлении области «1» или «2» нужно кликнуть мышью на эту область (x5); при появлении области синего или красного цвета нужно кликнуть мышью на «1», если это область синего цвета и «2», если это область красного цвета (x6).

	A	B	C	D	E	F	G
1		x1	x2	x3	x4	x5	x6
2	№1	0,25	0,29	1,69	1,31	1,88	1,96
3	№2	0,64	0,68	1,26	1,39	2,33	2,8
4	№3	0,67	0,77	1,94	1,39	2,82	2,83
5	№4	0,6	0,67	2,15	1,59	2,17	2,53
6	№5	0,64	0,76	1,18	1,69	2,34	1,81
7	№6	0,64	0,64	1,25	1,78	2,3	2,46
8	№7	0,7	0,66	1,64	0,98	2,1	1,66
9	№8	0,71	0,91	1,61	1,22	2,6	2,65
10	№9	0,59	0,78	1,56	1,41	2,51	2,85
11	№10	0,59	0,71	1,48	1,63	2,39	2,89

Рис. 1. Исходные данные.

Для реализации поставленной выше задачи авторами разработан алгоритм интеллектуальной обработки многомерных экспериментальных данных:

1) очистка данных;
2) кластеризация;
3) уменьшение размерности (метод главных компонент);
4) формирование области кластера;
5) нахождение центра кластера.

Реализация этапов:

1. *Очистка данных*. Это необходимая процедура предварительной обработки данных по исключению сбойных результатов, даже если таких нет. Программа просто не будет изменять данные. Сначала устанавливается критерий аномальности экспериментальных

данных (величина среднего и среднего квадратического отклонения). Процедура осуществляется в программе Deductor в разделе *Фильтрация-Редактирование выбросов и экстремальных значений*.

2. *Кластеризация данных.* Исходный набор данных разбиваем на 2 класса методом Прима - построением кратчайшего остовного дерева в программе MS Excel. По заданному графу заполняется матрица весов W(N, N). Веса несуществующих ребер предполагаются сколь угодно большими. Образуется массив P(N) меток вершин графа (столбцов матрицы весов). Алгоритм решения задачи заключается в последовательном заполнении массива меток столбцов и состоит из следующих этапов.

Предварительный этап. Обнуляется массив P(N) меток столбцов таблицы. Произвольно выбранному столбцу присваивается значение метки, равная его номеру.

Этап, повторяющийся N-1 раз (общий этап). В строках, номера которых равны номерам помеченных столбцов, находится минимальный элемент среди элементов непомеченных столбцов. Столбец, в котором находится минимальный элемент, помечается меткой, номер которой равен номеру его строки. В случае если минимальных элементов несколько, то выбирается любой. После отметки очередного столбца элементу, симметричному относительно главной диагонали (для многомерного графа – с «транспонированными индексами»), присваивается сколь угодно большое значение.

Заключительный этап. Ребра, включенные в минимальное остовное дерево, определяются по меткам столбцов. Вес остовного дерева задается суммой весов, входящих в него ребер.

Адаптивная кластеризация множества элементов производится путем удаления части ребер графа по критерию минимальной суммарной дисперсии классов. Для разбиения множества элементов на «К» классов удаляются «К-1» ребер [1]. В результате разбиения получены 2 группы студентов: 1-ый кластер - № 1, 5, 7, 2-й кластер - № 2, 3, 4, 6, 8, 9, 10.

Параллельно проводим кластеризацию методом k-средних в программе Deductor, результатом которой является выявление принадлежности студентов одному из кластеров и расстояние до центра. Также используем кластеризацию с помощью *Самоорганизующейся карты Кохонена* (рис. 2).

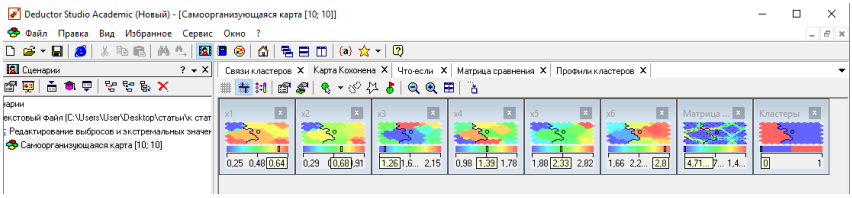

Рис. 2.
Кластеризация с помощью самоорганизующейся карты Кохонена.

Результаты кластеризации разными методами идентичны. Анализ журнала успеваемости студентов позволяет провести полную аналогию с разбиением на кластеры и успеваемостью студентов – в первый кластер попали студенты с лучшей успеваемостью.

3. *Уменьшение размерности исходных данных (сжатие данных).* Исходные многомерные данные невозможно представить графически, поэтому применим метод главных компонент для уменьшения размерности данных. В программе MS Excel с используем формулу матричных преобразований $Y = W*X$, где X – матрица исходных данных с 6-ю параметрами, Y – матрица сжатых данных с 2-мя параметрами, W – вспомогательная матрица весов. Изначально веса берутся любыми, предположим, все единицы. Применяя формулу обратного преобразования $X = W^T*Y$, получаем новые значения X. Возьмем сумму квадратов отклонений соответствующих значений xij из начальной и полученной матриц за целевую функцию и минимизируем ее с помощью надстройки *«Поиск решения»,* изменяя при этом диапазон весов W. В итоге получим сжатые исходные данные (рис. 3).

		№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9	№ 10	
22												
23	x1	0,83	1,01	1,15	1,05	0,89	0,99	0,82	1,06	1,07	1,07	y=wx
24	x2	1,66	2,01	2,27	2,11	1,80	1,99	1,61	2,10	2,14	2,15	

Рис. 3. Снижение размерности данных матричным методом.

Мы видим, что каждый объект (студент) характеризуется двумя параметрами x1 и x2. Теперь можно отметить точки с координатами (x1i, x2i) на координатной плоскости в Excel, откуда видно, что номера 1, 5 и 7 составляют отдельный кластер (рис.4).

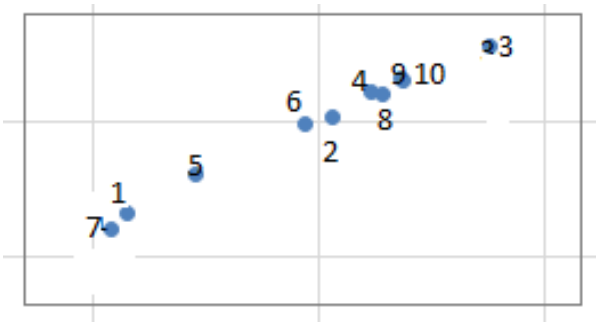

Рис. 4. Визуализация исходных данных

Таким образом, переход к главным компонентам, равным двум, в каждом кластере позволяет усилить визуальные возможности метода обработки.

Также можно использовать сценарий *Нейросеть* в Deductor, который позволяет визуализировать процесс снижения размерности данных.

4. *Формирование области кластера с помощью систем линейных неравенств (осуществляется в программе MS Excel).* Полученная многомерная область однородных групп имеет обычно сложную форму, поэтому ее аппроксимируют вписанным или описывающим прямоугольным гиперпараллелепипедом. Достоинством данной формы задания областей является простота описания области. Результаты анализа погрешностей аппроксимации наиболее распространенных областей гиперпараллелепипедами показывают, что погрешность резко возрастает с увеличением размерности пространства контролируемых параметров (количества контролируемых параметров). Уменьшения погрешности аппроксимации можно достигнуть более точным заданием границ областей, например, аппроксимировать системой гиперпараллелепипедов. В этом случае наиболее распространенным является задание области системой линейных неравенств:

$$\sum_{j=1}^{n} a_{ij} x_j + m_i \le 0, \; i = 1,2,...m. \qquad (1)$$

Исходными данными для определения количества линейных неравенств и значений их коэффициентов являются граничные точки области. Грани многогранника представляют собой гиперплоскости,

которые соответствуют уравнениям неравенств (1) при замене знака "≤" на знак "=".

Для наших данных получим

область ограничений первого кластера (точки 1, 5 и 7)

$$\begin{cases} -7x_1 + 3x_2 \le 2 \\ 5x_1 - x_2 \le 2 \\ -19x_1 + 7x_2 \le -4 \end{cases}$$

область ограничений второго кластера (точки 2, 4, 3 и 6)

$$\begin{cases} -2x_1 + x_2 \le 0 \\ -8x_1 + 5x_2 \le 20 \\ 13x_1 - 7x_2 \le -1 \\ -x_1 + x_2 \le 1 \end{cases}$$

5. *Нахождение точки внутри области примерно на равном расстоянии от границы области – центра кластера*. Необходимо найти точку внутри области примерно на равном расстоянии от границы области. В большинстве случаев применяется квадратичный критерий ошибки приближения.

Исходная система ограничений:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \ge b_1; \\ .....; \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \le b_m. \end{cases} \qquad (2)$$

Левые части обозначим:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n; \\ .....; \\ y_m = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n; \end{cases}$$

Берём y_1 как целевую функцию и на множестве (2) находим максимальное и минимальное значения y_1. Затем берем y_2 и опять находим максимальное и минимальное значения y_2 и т.д. ($y_{i\ min}$, $y_{i\ max}$).

$\bar{y}_i = \dfrac{y_{i\min} + y_{i\max}}{2}$ – среднее значение.

Тогда система ограничений примет вид:

$$\begin{cases} \bar{y}_1 = a_{11}\bar{x}_1 + ... + a_{1n}\bar{x}_n; \\ .....; \\ \bar{y}_m = a_{m1}\bar{x}_1 + ... + a_{mn}\bar{x}_n. \end{cases}$$

Это система линейных алгебраических уравнений.

$$\vec{y}_{cp} = \bar{A} \cdot \vec{x}_{cp};$$

$$\vec{x}_{cp} = \left(\overline{A}^T \cdot \overline{A}\right)^{-1} \cdot \overline{A}^T \cdot \vec{y}_{cp}.$$

Этот алгоритм осуществляется также в программе Excel. В большинстве случаев применяется квадратичный критерий ошибки приближения [9]. Использование центров кластеров в методе k-means позволяет повысить устойчивость к сбойным результатам.

Центр первого кластера имеет координаты (0, 85; 1,72), а второго (1,05; 2,12). Возвращаясь к исходным многомерным данным по формуле $X = W^T*Y$, можно получить усредненный результат теста эталонного представителя группы.

«Эталонный» представитель для 1-ого кластера имеет результаты (0,44; 0,50; 1,14; 1,04; 1,70; 1,78), а для 2-го кластера (0,55; 0,63; 1,44; 1,31; 2,16; 2,26). Значит, наиболее типичный представитель первой группы студент № 1, а второй группы – студент № 6.

Для того, чтобы студенты научились решать прикладные задачи, нужны очень подробные методические указания, как пользователям пакета прикладных программ, так и руководства к решению задач. Кроме того, изучение пакетов прикладных программ должно предварять изучение общего курса, чтобы студенты знали интерфейс программы и привыкли им пользоваться [2].

Предложенный подход и рассмотренные алгоритмы позволяют проводить обработку экспериментальных данных в многомерном пространстве и без сложного математического инструментария и получать очень важные результаты исследований в условиях неопределенности. Переход после этапа многомерной кластеризации к главным компонентам в каждом кластере позволяет усилить визуальные возможности последовательного метода обработки, что является важным моментом на этапе обучения. Вместо перечисления всех объектов можно дать список “эталонных” представителей групп (центров кластеризации), указать номера (имена) объектов, входящих в состав каждой группы. При небольшом числе групп описание данных становится обозримым и легко интерпретируемым.

Таким образом, применение адаптированных технологий интеллектуальной обработки данных с использованием аналитической платформы Deductor имеет много преимуществ: интересно студентам; визуализация результатов; реализация серьезных алгоритмов и задач без знания математического аппарата; имеет реальное практическое применение [10].

🕮 БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Методы интеллектуальной обработки данных: учебное пособие / Т.Г. Авачева и др.; ФГБОУ ВО РязГМУ Минздрава России. – Рязань: РИО УМУ, 2016. – 104 с.
2. Котюргина А.С., Никитин Ю. Б. О возможности обучения студентов основному курсу математики с применением пакетов прикладных программ // Научно-методический электронный журнал «Концепт». – 2017. – Т. 2. – С. 147–153. – URL: http://e-koncept.ru/2017/570032.htm.
3. Авачёва Т.Г., Дмитриева М.Н., Ельцов А.В., Кривушин А.А. Информационные технологии в обучении физике и математике студентов фармацевтических специальностей [Текст] // Психолого-педагогический поиск. – 2017. - № 1 (34). – С. 114 – 127.
4. Дюк В. А., Флегонтов А. В., Фомина И. К. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // Известия РГПУ им. А.И. Герцена. 2011. №138. С.77-83. URL: http://cyberleninka.ru/article/n/primenenie-tehnologiy-intellektualnogo-analiza-dannyh-v-estestvennonauchnyh-tehnicheskih-i-gumanitarnyh-oblastyah (дата обращения: 17.06.2017).
5. Марухина О.В., Мокина Е.Е., Берестнева Е.В. Применение методов Data Mining для выявления скрытых закономерностей задачах анализа медицинских данных //Фундаментальные исследования. – 2015. – № 4. – С. 107-113; URL: https://www.fundamental-research.ru/ru/article/view?id=37131 (дата обращения: 04.06.2017).
6. Ельцов А.В., Махмудов М.Н. Интеграция процессов познания и моделирования при обучении физике [Текст] // Психолого-педагогический поиск. 2015. № 2 (34). С. 145-151.
7. Кривушин А.А. Возможности виртуального физического эксперимента на занятиях по астрономии и физике [Текст] // Учебная физика. 2015. № 5. С. 57-61.
8. Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений: сборник материалов межвуз. научн.-практ. конф. – Рязань: Лаборатория баз данных, 2010. – 155 с.
9. Авачева Т.Г., Дорошина Н.В., Кабанов А.Н. Разбиение дискретного конечного множества элементов на основе метода оптимума номинала выпуклой области. Современные технологии в науке и образовании – СТНО-2017 [текст]: сб. тр. между- нар. науч.-техн. и науч.-метод. конф.: в 4 т. Т.3. / под общ. ред. О.В. Миловзорова. – Рязань: Рязан. гос. радиотехн. ун-т, 2017; Рязань. – 300 с.

10. Дорошина Н. В., Кабанов А.Н. Использование аналитических пакетов при обучении студентов в задачах интеллектуальной обработки данных. Материалы ежегодной научной конференции Ряз. гос. мед. университета имени академика И.П. Павлова/ редкол: Р.Е. Калинин, В.А. Кирюшин, И.А. Сучков; ФГБОУ ВО РязГМУ Минздрава России – Рязань: РИО РязГМУ, 2016. С. 158-161.