УДК 004.67

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING ДЛЯ ВЫЯВЛЕНИЯ СКРЫТЫХ ЗАКОНОМЕРНОСТЕЙ В ЗАДАЧАХ АНАЛИЗА МЕДИЦИНСКИХ ДАННЫХ

Марухина О.В., Мокина Е.Е., Берестнева Е.В.

ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», Томск, e-mail: Marukhina@tpu.ru

В настоящее время ожирение является одним из самых распространенных хронических заболеваний, по данным Всемирной организации здравоохранения (World Health Organization) к началу XXI века избыточную массу имело около 30% населения планеты. В том числе, ожирение у детей и подростков является одной из актуальных проблем современного здравоохранения. Детское ожирение является фактором, обуславливающим более высокую вероятность ожирения, преждевременной смерти и инвалидности во взрослом возрасте. Накопленные в архивах медицинские данные актуальны для изучения механизмов ожирения среди детского населения, поскольку содержат огромный запас информации, сведений о различных случаях заболеваний, ходе и тяжести лечения, значения разнообразных клинических и лабораторных показателей. По результатам диспансеризации в Томской области каждый 15-й ребенок страдает ожирением. Решением проблемы заболевания ожирением может являться анализ клинико-лабораторных показателей, отражающих состояние и клиническую картину детей и нахождение закономерностей, помогающих корректировать лечение. В статье рассмотрены вопросы применения методов Data Mining в задачах медицинских исследований. Рассмотрено применение кластер-анализа в задаче выявления закономерностей (определение групп детей, имеющих схожий результат лечения) после проведения лечения. Полученный результат даст возможность правильно выбрать процедуру лечения для вновь поступивших пациентов.

Ключевые слова: многомерные данные, data mining, скрытые закономерности

USING DATA MINING FOR REVEALING HIDDEN REGULARITIES IN THE TASK OF ANALYZING MEDICAL DATA

Marukhina O.V., Mokina E.E., Berestneva E.V.

National Research Tomsk Polytechnic University, Tomsk, e-mail: Marukhina@tpu.ru

Currently, obesity is one of the most common chronic diseases, according to the World Health Organization (World Health Organization) to the beginning of the XXI century overweight had about 30% of the world population. In particular, obesity in children and adolescents is one of the urgent problems of modern health care. Childhood obesity is a factor that contributes to a higher likelihood of obesity, premature death and disability in adulthood. Accumulated in the archives of medical data relevant to the study of the mechanisms of obesity among children, because they contain a huge stock of information, information about various cases of diseases, the course and severity of the treatment, the values of a variety of clinical and laboratory parameters. According to the results of the clinical examination in the Tomsk region every 15 children is obese. Solution to the problem of the disease of obesity may be the analysis of clinical and laboratory parameters, reflecting the state and the clinical picture of children and finding regularities to help adjust the treatment. The article discusses the use of Data Mining methods in problems of medical research. The application of cluster analysis to identify patterns of problem (definition of groups of children with similar results of treatment) after treatment. This result will provide an opportunity to choose the right procedure for the treatment of newly admitted patients.

Keywords: multidimensional data, data mining, hidden patterns

Архивы медицинских данных, находящиеся в распоряжении различных медицинских учреждений, содержат огромный запас сведений о различных случаях каждого конкретного заболевания. Извлечение «скрытых» закономерностей из массива данных – одна из задач многих исследований медицинской тематики. Для решения таких задач применяются методы автоматического анализа, при помощи которых приходится практически добывать знания из «завалов» информации.

Термин Data Mining часто переводится как добыча данных, раскопка данных или «извлечение зерен знаний из гор данных». Технологию Data Mining достаточно точно определяет один из основателей этого направления Григорий Пиатецкий-Шапиро: «Data Mining – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1].

Data Mining – совокупность различных методов обнаружения знаний. Выбор метода часто зависит от типа имеющихся данных и от того, какую информацию необходимо получить. Вот некоторые из них: классификация, кластеризация, ассоциация (объединение), анализ временных рядов, нейронные сети, прогнозирование и анализ временных рядов и т.д.

Технология Data Mining применяется практически везде, где возникает задача

автоматического анализа данных. Основные сферы применения технологии Data Mining: наука, бизнес, исследования для правительства и Web-направление. Широкое распространение эти методы получили также в медицинских исследованиях.

Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания симптомов различных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как его нужно лечить [1]. Правила помогают выбирать средства медикаментозного воздействия, определять показания (противопоказания), ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения, изучать причины и механизмы возникновения различных патологий, исследовать эффективность хирургического вмешательства и т.п. Технология Data Mining позволяет обнаруживать в медицинских данных шаблоны, которые составляют основу указанных правил.

Метод Data Mining в медицинских исследованиях применяли разные авторы: Гудинова Ж.В. Применение Data Mining (Обнаружение полезных знаний в базах данных) как основа исследований и управления в сфере охраны здоровья населения и среды обитания [5]; А.В. Кузнецова, О.В. Сенько, Возможности использования методов Data Mining при медико-лабораторных исследованиях для выявления закономерностей в массивах данных [6]; Дзюра А.Е., Берестнева Е.В. Применение Data Mining в медико-психологических исследованиях [7].

Поиск скрытых закономерностей в медицинских данных. Исходный массив данных нашей задачи – матрица значений медицинских показателей, сформированная в НИИ курортологии г. Томска (рис. 1). Объекты исследования – дети, страдающие разной формой ожирения. Цель нашего исследования – выявление закономерностей (определение групп детей, имеющих схожий результат лечения) после проведения лечения.

В настоящее время на рынке программного обеспечения существует огромное разнообразие программных продуктов, реализующих самый разнообразный спектр методов Data Mining. Разработчики универсальных статистических пакетов, в дополнение к традиционным методам статистического анализа, включают в пакет определенный набор методов Data Mining. Это такие пакеты, как SPSS(SPSS, Clementine), Statistica (StatSoft), SAS Institute (SAS Enterprise Miner). Некоторые разработчики OLAP-решений также предлагают набор методов Data Mining, например семейство продуктов Cognos. Есть поставщики, включающие Data Mining решения в функциональность СУБД: это Microsoft (Microsoft SQL Server), Oracle, IBM (IBM Intelligent Miner for Data).

Для реализации поставленной задачи был использован достаточно популярный пакет Statistica, разработчиком которого является компания StatSoft. Компанией StatSoft была разработана целая система Statistica Data Miner, реализующая технологию Data Mining [3, 5]. Данная система спроектирована и реализована как универсальное всестороннее средство анализа данных (от взаимодействия с различными базами данных до создания готовых отчетов), реализующее так называемый графически-ориентированный подход. Система Statistica Data Miner подходит для выявления скрытых правил и закономерностей, проведения углубленных исследований, где «не работают» классические методы математической статистики. Авторами разработан алгоритм выявления скрытых закономерностей, схема которого представлена на рис. 2. Алгоритм был апробирован при решении подобных прикладных исследовательских задач [2]. Исходный массив данных задачи, представленной в данной статье, – матрица значений X_{nm} медицинских показателей, сформированная в НИИ курортологии г. Томска (n – количество пациентов (n = 269); m – число медицинских показателей (m = 111).

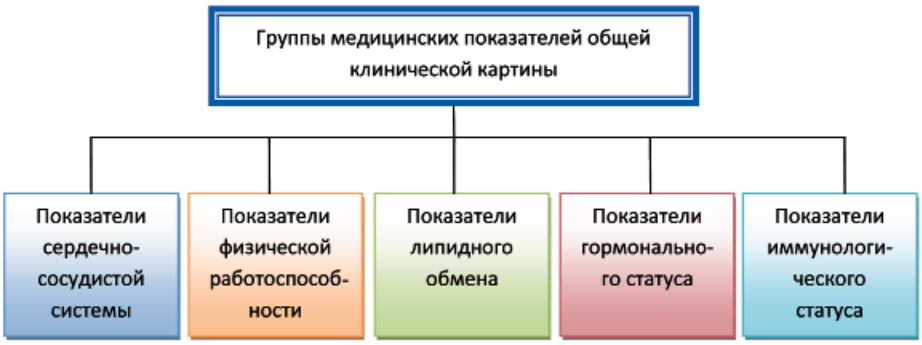

Рис. 1. Группы медицинских показателей

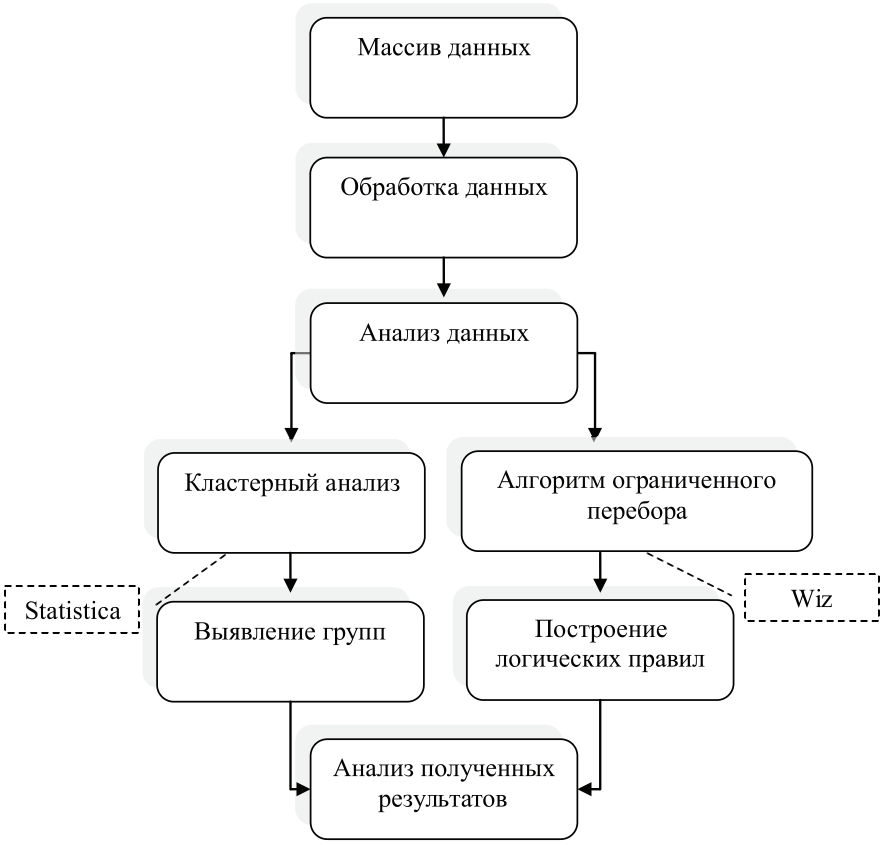

Рис. 2. Схема алгоритма выявления скрытых закономерностей

	ANOVA for continuous variables (Копия ожирение Number of clusters: 3 Total number of training cases: 104					
	Between SS	df	Within SS	df	F	p value
степень ожирения	355,681	2	34,5500	101	519,8806	0,000000
ОТ пл-дл	711,533	2	469,0973	101	76,5990	0,000000
ОБ пл-дл	657,269	2	319,9028	101	103,7568	0,000000
Масса пл-дл	764,623	2	148,1127	101	260,7031	0,000000
Избыток пл-дл	2870,852	2	368,3402	101	393,5981	0,000000
ИМТ пл-дл	107,774	2	127,4182	101	42,7145	0,000000

Рис. 3. Результат дисперсионного анализа

Метод кластерного анализа [1, 4] позволяет разбивать множества исследуемых объектов и признаков на однородные группы (кластеры). Для реализации процедуры кластерного анализа существует несколько методов, авторами был использован наиболее распространенный метод k-средних, целью которого является разбиение объектов на k кластеров.

Принципиальное отличие метода k-средних от иерархического кластер-анализа заключается в том, что исследователю необходимо изначально определить число кластеров, на которое требуется разбить изучаемую совокупность. Соответственно, желательно еще до начала анализа иметь гипотезу о структуре исследуемой совокупности. Авторы статьи проводили кластер-анализ для $k = 3, 4, 5$. Анализ результатов специалистами НИИ курортологии показал, что наилучший результат показывает кластер-анализ при $k = 3$.

В дисперсионном анализе межгрупповая дисперсия сравнивается с внутригрупповой дисперсией для принятия решения, являются ли средние для отдельных переменных в разных совокупностях значимо различными. Исходя из уровней значимости (p-value), все выбранные авторами переменные являются значимыми при решении задачи о распределении объектов по кластерам (рис. 3).

Одним из способов определения кластеров является проверка средних значений для каждого кластера и для каждого измерения (рис. 4).

Анализируя график, можно сделать вывод, что пациенты, входящие в Cluster 3, получили лучшие результаты при лечении. Это показывают переменные: масса тела, избыток массы тела (%) и индекс массы тела.

В таблице на рис. 5 представлены объекты, входящие в кластер 3. В основном это пациенты с 3 и более степенью ожирения. Столбцы со значениями указывают на то, каких они достигли результатов при лечении.

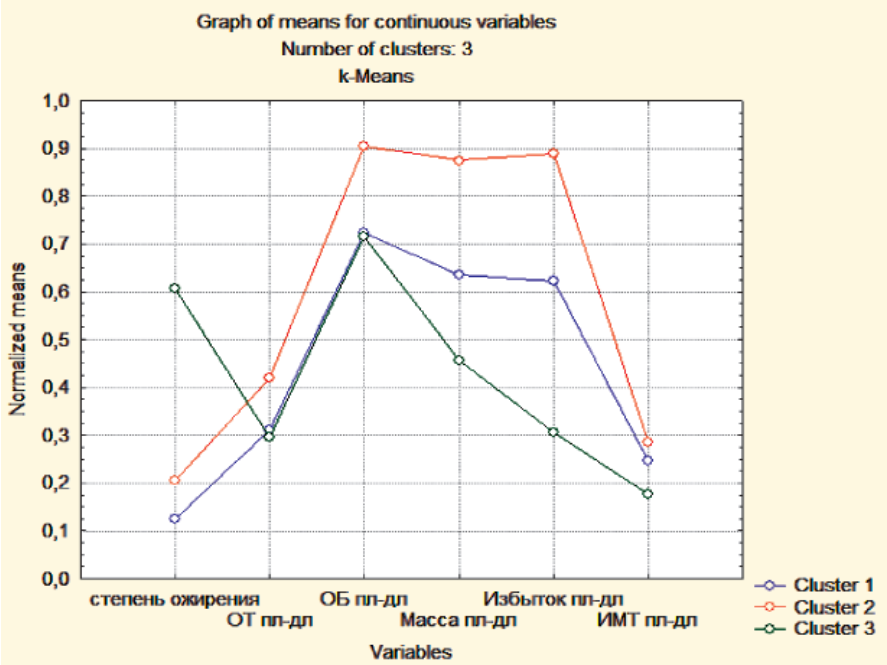

Рис. 4. График средних значений кластеров

Cluster members for cluster: 3 (Копия ожирение для матем с обозначениями)
Number of cases: 22

Case No.	степень ожирения	ОТ пл-дл	ОБ пл-дл	Масса пл-дл	Избыток пл-дл	ИМТ пл-дл	Distance to centroid
51	3,000000	-5,00000	-2,0000	-5,50000	-12,2000	-2,30000	0,324861
93	3,000000	-2,00000	-2,0000	-4,00000	-8,0000	-1,20000	0,169081
106	3,000000	-2,00000	-3,0000	-2,00000	-4,0000	-0,80000	0,486277
161	3,000000	-1,00000	-2,0000	-4,00000	-7,8000	-1,70000	0,199575
165	2,000000	-5,00000	-1,0000	-4,00000	-8,1000	-1,75000	0,353308
167	2,000000	-5,00000	-6,0000	-7,00000	-11,1000	-2,70000	0,539857
168	3,000000	0,00000	-11,0000	-8,00000	-12,1000	-2,80000	0,903876
177	2,000000	-7,00000	-4,0000	-6,50000	-9,5000	-1,50000	0,438983
178	2,000000	-6,00000	0,0000	-3,30000	-8,6500	-1,60000	0,434625
179	3,000000	-5,00000	-6,0000	-4,00000	-12,7000	-1,80000	0,399184
180	3,000000	-9,00000	-1,0000	-8,00000	-12,0000	-2,70000	0,627187
187	4,000000	0,00000	0,0000	-1,00000	-3,0000	-0,20000	0,806808
190	3,000000	-3,00000	0,0000	-4,00000	-13,0000	-2,20000	0,425255
198	3,000000	-6,00000	-5,0000	-1,15000	-9,2000	-0,50000	0,452387
203	3,000000	-3,00000	-5,0000	-4,10000	-7,1000	-2,10000	0,234156
208	3,000000	-2,00000	-5,0000	-3,10000	-6,0000	-1,80000	0,332121
209	3,000000	-1,00000	-2,0000	-3,00000	-6,0000	-1,00000	0,331586
220	3,000000	-7,00000	-7,0000	-6,10000	-13,0000	-4,10000	0,581273
224	2,000000	-2,00000	-2,0000	-4,00000	-9,5000	-1,60000	0,305314
242	2,000000	0,00000	-1,0000	-4,00000	-9,5000	-1,80000	0,383122
243	3,000000	-2,00000	-2,0000	-4,00000	-5,5000	-1,30000	0,305722
259	4,000000	-2,00000	-2,0000	-3,00000	-10,0000	-1,60000	0,448391

Рис. 5. Объекты кластера 3

Алгоритм ограниченного перебора. С целью обработки клинико-лабораторных показателей авторами был использован алгоритм ограниченного перебора (с применением пакета Wiz Why). Поиск логических правил осуществлялся для клинико-лабораторных показателей до лечения, так и для значения разности показателей до лечения и после лечения. В частности, были получены правила, характеризующие показатели пациентов до лечения к индексу массы тела после лечения. Эти правила позволяют понять, на каких пациентов и с какими признаками лечение подействовало эффективнее.

Для пояснения полученных правил, наиболее подробно рассмотрено правило № 5:

If IgM do is 0,80 ... 2,10 (average = 1,43)
and Lizocim do is 36,00
Then
IMT p-do is not more than -1,08
Rule's probability: 0,909
The rule exists in 10 records.
Significance Level: Error probability < 0,01

Это правило представляет собой конъюнкцию двух высказываний:

If IgM do is 0,80 ... 2,10 (average = 1,43) – если концентрация иммуноглубина М до лечения от 0,80 до 2,10, и Lizocim do is 36,00 активность лизоцима в сыворотке крови до лечения равняется 36, то IMT p-do индекс массы тела в результате проведенного лечения уменьшается более чем на 1,08.

Запись Rule's probability: 0,909 означает точность правила в данном случае равной 0,909. Следующая запись The rule exists in 10 records характеризует множество объектов, для котрых справедливо рассматриваемое правило, а запись Significance Level: Error probability < 0,01 касается статистической оценки уровня значимости.

В системе предусмотрена визуализация любого правила. Для этого надо позиционировать курсор на одном из его условий и щелчком правой кнопки мыши вызвать контексное меню, где выбрать «Rule Chart…». В окне отобразится диаграмма, иллюстрирующая отдельные компоненты правила (рис. 6).

Левая часть окна содержит правило в текстовом режиме. Правая часть – визуализация правила. Зеленая полоса показывает долю объектов, обладающих целевым значением, красная полоса иллюстрирует пропущенные объекты, белые разделы – другие объекты в этом поле. Длина полосы соответствует доле таких объектов.

Оценить информативность признака можно при помощи таблицы «Field Index», где перечисляются признаки, участвующие в полученных правилах, и приводится список их номеров. Анализируя табл. 1, можно сказать, что показатели: DV.PR. do, OXC do, TTG(0,23-3,4) do имеют наибольшее значение при выявлении закономерностей, а в табл. 2 показатели: DV.PR.p-do, CIK p-do, Kortizol p-do, OL p-do.

На основе анализа логических правил, представленных в приложении В, можно сделать вывод о том, что уменьшение или увеличение определенных значений показателей способствует лучшему результату при проведении лечения и наоборот, улучшение в лечении не наблюдается или наблюдается, но незначительное.

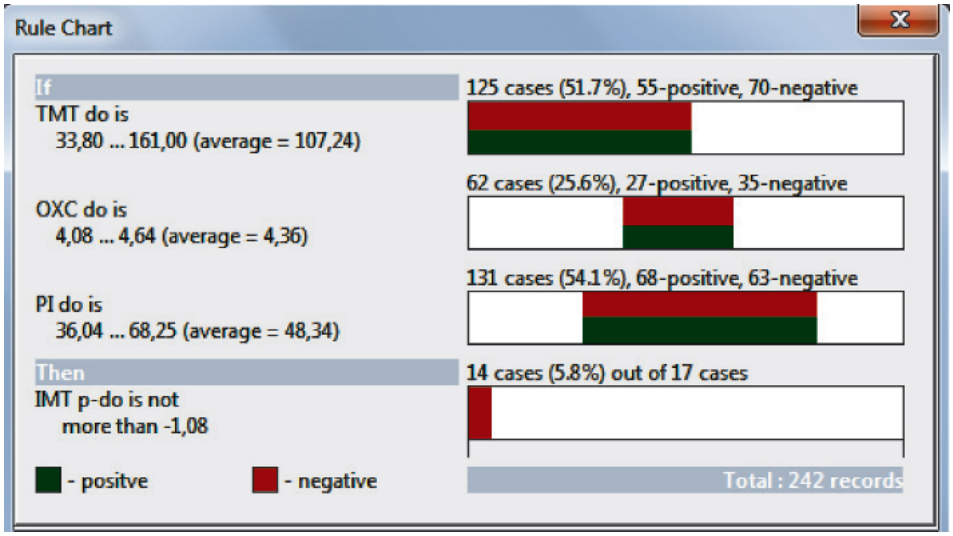

Рис. 6. Диаграмма правила

Таблица 1

Field Index (до лечения)

Field	Rule
DV.PR. do	2, 10, 12, 15, 21
FNO (ne > 2,5 pg/ml) do	13, 14
IA do	1, 6, 17, 20
IgM do	2, 5, 21
Kortizol do	1, 12, 14, 18
LPNP do	3
Lizocim do	5, 9, 13, 15
OXC do	3, 4, 11, 16, 19
PI do	4, 6
TAG do	7, 19, 20, 21
TMT do	1, 4, 6
TTG(0,23-3,4) do	2, 3, 7, 8, 10, 12

Таблица 2

Field Index (после лечения)

Field	Rule
APF p-do	2, 5, 28, 29, 33, 50, 65, 69, 70
CIK p-do	15, 20, 21, 31, 32, 34, 37, 41, 42, 47, 54, 57, 63, 64, 68, 69, 73
DAD p-do	1, 2, 6, 8, 13, 17, 19, 23, 26, 46,53, 59
DV.PR.p-do	9, 10, 12, 17, 22, 23, 24, 30, 31, 36, 39, 40, 48, 49, 54, 61, 62, 63, 69, 71, 72,
FNO p-do	43, 61, 62
IgG p-do	2, 5, 19, 20, 21, 30, 31, 32, 46, 52
IgM p-do	27
Insylin p-do	11, 13, 15, 16, 19, 28, 29, 33, 68
Kortizol p-do	4, 7, 8, 9, 18, 28, 29, 33, 39, 40, 47, 48, 49, 52, 54, 57, 60, 73, 74
LPONP p-do	21, 29, 40, 42, 45, 49, 56, 62, 67, 72
Lizocim p-do	37, 41, 42, 54, 63, 64, 73, 74
MG p-do	3, 10, 17, 24
NOMA p-do	14, 34, 37, 46, 53, 58, 59, 60, 70, 71, 72
OL p-do	4, 14, 18, 20, 21, 27, 30, 32, 35, 38, 41, 42, 48, 49, 50, 51, 59, 61, 62, 63, 71, 72, 74
OXC p-do	25, 26, 35
PI p-do	16, 36, 39, 40, 53, 58, 60, 64, 66, 67
SAD p-do	1, 10, 12, 25, 44, 45, 58, 70
T syp p-do	18, 19, 46, 51, 57, 64
TAG p-do	20, 28, 39, 41, 44, 48, 55, 61, 66, 71
TFN p-do	34, 37, 47, 52, 53, 58, 59, 60, 68, 70,73

Правило № 2:

If **DAD p-do** *is* **-12,00 ... 0,00** (average = **-4,58**)
and **IgG p-do** *is* **-21,00 ... -0,80** (average = **-5,20**)
and **APF p-do** *is* **-26,67 ... -9,40** (average = **-16,23**)
Then
IMT p-do *is not* ***more than* -1,08**
Rule's probability: **1,000**
The rule exists in **12** *records.*
Significance Level: Error probability < 0,0001

Из анализа правила № 2 следует, что снижение диастолического артериального давления, иммуноглобубина G и ангиотензинпревращающего фермента в сыворотке крови, способствует лучшему похудению пациентов, индекс массы тела уменьшается более чем на 1,08, достоверность 100 %.

Заключение

Результатом выявления скрытых закономерностей для данных поставленной задачи является в дальнейшем корректировка траектории лечения детей с различными степенями ожирения в НИИ курортологии г. Томска. В частности, полученные знания о результатах лечения в каждой такой группе дадут возможность правильно выбрать процедуру лечения для поступивших пациентов. Разработанный авторами алгоритм выявления скрытых закономерностей будет входить в состав информационной системы, разрабатываемой в рамках выполнения проекта РФФИ № 14-07-00675.

Работа выполнена при поддержке гранта РФФИ, проект № 14-07-00675.

Список литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 471 с.

2. Берестнева О.Г., Пеккер Я.С. Выявление скрытых закономерностей в сложных системах // Изв. Том. политехн. ун-та. – 2009. – Т. 315, № 5. – С. 138–143.

3. Бурева Н.Н. Многомерный статистический анализ с использованием ППП «Statistica»: учебно-методический материал по программе повышения классификации «Применение программных средств в научных исследованиях и преподавании математики и механики». – Нижний Новгород, 2007. – 112 с.

4. Дюк В., Самойленко А. Data Minning: учебный курс. – СПб.: Питер, 2001. – 386 с.

5. Чубукова И.А. Data Mining: учебное пособие. – Интернет – университет информационных технологий: БИНОМ: Лаборатория Знаний, 2006. – 382 с.

References

1. Ajvazjan S.A., Buhshtaber V.M., Enjukov I.S., Meshalkin L.D. Prikladnaja statistika. Klassifikacija i snizhenie razmernosti. M.: Finansy i statistika, 1989. 471 p.

2. Berestneva O.G., Pekker Ja.S. Vyjavlenie skrytyh zakonomernostej v slozhnyh sistemah // Izv. Tom. politehn. un-ta. 2009. T. 315, no. 5. pp. 138–143.

3. Bureva N.N. Mnogomernyj statisticheskij analiz s ispol'zovaniem PPP «Statistica»: uchebno-metodicheskij material po programme povyshenija klassifikacii «Primenenie programmnyh sredstv v nauchnyh issledovanijah i prepodavanii matematiki i mehaniki». – Nizhnij Novgorod, 2007. 112 p.

4. Djuk V., Samojlenko A. Data Minning: uchebnyj kurs. SPb.: Piter, 2001. 386 p.

5. Chubukova I.A. Data Mining: uchebnoe posobie. – Internet – universitet informacionnyh tehnologij: BINOM: Laboratorija Znanij, 2006. 382 p.

Рецензенты:

Уразаев А.М., д.б.н., профессор, Институт теории образования, ФГБОУ ВПО «Томский государственный педагогический университет», г. Томск;

Мещеряков Р.В., д.т.н., профессор, заведующий кафедрой безопасности информационных систем, Томский государственный университет систем управления и радиоэлектроники, г. Томск.

Работа поступила в редакцию 12.02.2015.