ISSN 2658 – 7505
Выпуск №2, 2019
Электронный научный журнал «Вестник молодёжной науки России»

УДК 004.92:616-056.52

РАЗРАБОТКА ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ АНАЛИЗА РЕЗУЛЬТАТОВ ЛЕЧЕНИЯ ПАЦИЕНТОВ С РАЗЛИЧНЫМИ ФОРМАМИ ОЖИРЕНИЯ

И.А. Лызин, *e-mail: i-lyzin@mail.ru*
Томский политехнический университет (ТПУ), г. Томск

В данной статье рассматривается разработка информационной технологии анализа результатов лечения детей с различными формами ожирения. В настоящее время существует множество подходов к лечению ожирения, накоплены массивы данных клинико-лабораторных показателей пациентов. В связи с этим, актуальной становится задача разработки информационных технологий для оперативной обработки этих массивов и получения новых знаний. В процессе работы, проведено визуальное исследование структуры пропущенных данных клинико-лабораторных показателей, осуществлено восстановление пропущенных данных с использованием метода множественного восстановления, на основе полученных результатов построены визуальные образы с применением такого типа пиктографиков, как лица Чернова. Результатом работы стала информационная технология, осуществляющая функции: анализа выбросов, анализа пропусков, восстановления пропусков, построения визуальных образов.

Ключевые слова: ожирение, информационная технология, восстановление данных, анализ, многомерные данные, визуальные образы.

DEVELOPMENT OF INFORMATION TECHNOLOGY OF ANALYSIS OF RESULTS OF TREATMENT OF PATIENTS WITH DIFFERENT FORMS OF OBESITY

I.A. Lyzin
Tomsk Polytechnic University (TPU), Tomsk

This article discusses the development of the information technology analysis of the results of treatment of children with various forms of obesity. Currently, there are many approaches to the treatment of obesity, accumulated arrays of clinical-laboratory indicators of patients. In this connection the task of developing information technologies for the rapid processing of these arrays and obtaining new knowledge becomes urgent.

In the process, we carried out a visual study of the structure of the missing data of clinical and laboratory parameters, carried out the imputation of missing data using the method of multiple imputation missing data, based on the obtained results of the constructed visual images with the use of this type of pictographical as Chernoff faces.

The result of the work was the information technology that performs the functions of analysis of outliers, analysis of missing data, imputation-missing data, and construction visual images.

Key worlds: obesity, information technology, restoration of missed data, analysis, multidimensional data, visual images.

Проблему избыточного веса называют одной из болезней цивилизации, поскольку ее распространенность в современном обществе постоянно растает. В последнее время статистические наблюдения фиксируют рост частоты ожирения у детей и подростков. По данным всемирной организации здравоохранения, за последние четыре десятилетия в мире в десять раз увеличилась доля детей и подростков (от 6 до 17 лет), страдающих ожирением.

Значимость проблемы ожирения определяется угрозой инвалидизации пациентов молодого возраста и снижением общей продолжительности жизни. Ожирение снижает устойчивость к инфекционным и простудным заболеваниям, кроме того, резко увеличивает риск осложнений при оперативных вмешательствах [1].

В настоящее время существует множество подходов к лечению ожирения, накоплены массивы данных клинико-лабораторных показателей пациентов. В

связи с этим, актуальной становится задача разработки информационных технологий для оперативной обработки этих массивов и получения новых знаний.

Целью работы является создание информационной технологии оценки результатов лечения пациентов, страдающих ожирением разной степени, проводимого по пяти различным траекториям.

Научной новизной исследования является алгоритм оценки результатов лечения пациентов, заложенный в основу информационной технологии.

Предметом исследования является массив данных предоставленный НИИ курортологии и физиотерапии г. Томска, в котором содержатся группы клинико-лабораторных показателей пациентов. Состояние каждого пациента оценивалось при поступлении на лечение и по окончании курса реабилитации на основании набора лабораторных данных и химического анализа крови. Пациенты – дети и подростки в возрасте от 6 до 17 лет, с избыточной массой тела и ожирением. В состав базы включены данные 276 пациентов, поделенных на 5 групп. Деление пациентов на группы осуществлялось в зависимости от способа лечения.

Задача исследования заключается в определении эффективности проведенного лечения по средствам построения визуальных образов групп пациентов до и после лечения.

В качестве инструмента решения поставленной задачи был выбран язык R, так как R является популярным инструментом для визуализации данных. Язык R является одним из ведущих статистических инструментов в мире. Он активно применяется в генетике, молекулярной биологии и биоинформатике, науках об окружающей среде. Также R все больше используется в обработке медицинских данных. В качестве среды разработки выбрано – Rstudio. RStudio – свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R [2].

Анализ выбросов является первым этапом в подготовки данных. Под "выбросом" понимается наблюдение, которое "слишком" велико или "слишком" мало по сравнению с большинством других имеющихся наблюдений. Обычно для обнаружения выбросов используют диаграмму размахов. Диаграмма размаха, также известная как ящик с усами. Этот график упрощает данные, оставляя только несколько параметров: медиану, верхнюю и нижние квартили, существенный минимум и максимум, а также выбросы, которые существенно меньше или больше остальных чисел в ряду. Обычно этот способ применяется, если необходимо визуально сравнить несколько категорий данных. После того как выбросы были исключены можно переходить к следующему шагу.

Следующим этапом в подготовке данных является анализ пропущенных значений. Большое количество статистических методов предполагает, что в ходе наблюдений были получены полностью укомплектованные наборы данных. Но на практике пропуски в данных все же являются повсеместным явлением. Когда в исходном наборе присутствуют пропуски необходимо оценить причину их появления и их влияние на дальнейший анализ. Если пропущенные значения не обрабатываются должным образом в конечном итоге могут получиться результаты не соответствующие действительности. Поэтому анализ пропущенных значений является важной задачей. Обычно максимальный порог пропущенных данных составляет 15% от общего количества выборки. Если пропущенные данные превышают 15%, необходимо отказаться от этого показателя. Существует три типа недостающих данных:

1. Полностью случайные пропуски (MCAR). Данные являются MCAR, если вероятность пропуска данных в переменной X не связана с другими переменными и значениями самих X;

2. Случайные пропуски (MAR). Данные типа MAR встречаются, когда пропуски во всей совокупности данных случайно распределены не по всем переменным, а внутри определенных подгрупп этих переменных;

3. Отсутствуют не случайно (MNAR). Данные являются MNAR, если вероятность пропуска данных систематично связана с предполагаемыми значениями этих пропущенных данных.

Идентификация пропущенных значений обычно делается с использованием функций is.na() и complete.cases(). Анализ исходных данных показал, что процент пропусков не превышает 15% в некоторых случаях. Следовательно, можно переходить к процессу восстановления.

Метод множественного восстановления пропущенных данных – это способ заполнения пропусков при помощи повторного моделирования. Множественное восстановление часто применяется для работы с пропущенными данными в сложных ситуациях. При этом подходе из существующего набора данных с пропущенными значениями создается несколько полных наборов данных. Для замещения пропущенных значений в производных наборах данных используются методы Монте-Карло. К каждому из производных наборов данных применяются стандартные статистические методы, а на основании их результатов формируются оценки окончательных результатов и доверительные интервалы, которые учитывают неопределенность, созданную пропущенными значениями. Для устранения пропусков в данных был использован пакет mice языка R.

Процесс восстановления пропусков включает в себя несколько этапов:

1. Функция mice() обрабатывает исходную таблицу данных с пропущенными значениями, а возвращает объект, содержащий несколько полных наборов данных (пять по умолчанию). Каждый такой полный набор данных получается при восстановлении пропущенных данных исходной таблицы. Все полные наборы данных отличаются друг от друга значениями пропусков, полученных в ходе процедуры Гиббса. Этот процесс повторяется, пока значения для пропущенных данных не сойдутся: mice(data, m);

2. При помощи функции with() применяется статистическая модель (например, линейная или обобщенная линейная): with(imp, analysis);

3. Функция pool() объединяет результаты: pool(fit);

4. Функция complete() создает полный набор данных с заполненными пропусками: complete(imp, action=#).

После применения функции mice() анализируемые данные принимают следующую структуру рис. 1. К каждой переменной с пропусками был применен метод сочетания, предсказанного среднего (ppm). Восстановление данных применялось ко всем показателям исходной базы. Строка VisitSequence содержит информацию о последовательности замены данных в переменных – справа налево, начиная с ОТ дл и заканчивая окЛППНП пл. Матрица PredictorMatrix показывает, что при восстановлении пропущенных данных каждой переменной использовались все остальные переменные. В этой матрице строки соответствуют переменным, значения которых восстанавливались, столбцы – переменным, которые использо-

ISSN 2658 – 7505
Выпуск №2, 2019
Электронный научный журнал «Вестник молодёжной науки России»

вались для расчета замещаемых значений, а 1/0 указывает, использовалась ли переменная [3]. Результатом работы функции mice () стала база с восстановленными значениями.

```
Multiply imputed data set
Call:
mice(data = Base[3:12])
Number of multiple imputations:  5
Missing cells per column:
     ОТ дл      ОТ пл      ОБ дл      ОБ пл  Масса, кг дл  Масса, кг пл Избыток, % дл Избыток, % пл
       180        177        164        165            33            33            41            43
    ИМТ дл     ИМТ пл
        37         40
Imputation methods:
     ОТ дл      ОТ пл      ОБ дл      ОБ пл  Масса, кг дл  Масса, кг пл Избыток, % дл Избыток, % пл
     "pmm"      "pmm"      "pmm"      "pmm"         "pmm"         "pmm"         "pmm"         "pmm"
    ИМТ дл     ИМТ пл
     "pmm"      "pmm"
VisitSequence:
     ОТ дл      ОТ пл      ОБ дл      ОБ пл  Масса, кг дл  Масса, кг пл Избыток, % дл Избыток, % пл
         1          2          3          4             5             6             7             8
    ИМТ дл     ИМТ пл
         9         10
PredictorMatrix:
             ОТ дл ОТ пл ОБ дл ОБ пл Масса, кг дл Масса, кг пл Избыток, % дл Избыток, % пл ИМТ дл ИМТ пл
ОТ дл            0     1     1     1            1            1            1            1      1      1
ОТ пл            1     0     1     1            1            1            1            1      1      1
ОБ дл            1     1     0     1            1            1            1            1      1      1
ОБ пл            1     1     1     0            1            1            1            1      1      1
Масса, кг дл     1     1     1     1            0            1            1            1      1      1
Масса, кг пл     1     1     1     1            1            0            1            1      1      1
Избыток, % дл    1     1     1     1            1            1            0            1      1      1
Избыток, % пл    1     1     1     1            1            1            1            0      1      1
ИМТ дл           1     1     1     1            1            1            1            1      0      1
ИМТ пл           1     1     1     1            1            1            1            1      1      0
Random generator seed value:  NA
```

Рис. 1 Фрагмент из восстановления данных

После восстановления пропусков в исходных наборах были получены полностью укомплектованный набор данных, который можно использовать для моделирования визуальных образов используя лица Чернова. Чернова. Лица Чернова – это схема визуального представления мультивариативных данных в виде человеческого лица. [4]. Идея использования лиц заключается в том, что люди легко распознают лица и без труда замечают небольшие изменения, для человека очень естественно смотреть на лица, легко делать сравнения и выявлять отклонения. Но так как количество показателей намного больше чем можно поместить в лица (всего 18), необходимо подготовить восстановленные данные к построению визуальных образов. Подготовка данных предполагает определения наиболее информативных признаков из всего восстановленного набора данных. Информативность признаков была определена с использованием информативной меры Кульбака. Данный метод предлагает в качестве оценки информативности меру расхождения между двумя классами которая называется дивергенцией [5]. В данном случае в качестве классов выступают показатели до и после лечения. После расчета информативности по 5 группам были получены следующие результаты, которые представлены в таблице 1.

Следовательно, информативными признаками являются: ТМТ – тощая масса тела, NO – содержание оксида азота в крови, Каталаза – каталаза сыворотки крови, Лептин – содержание лептина в сыворотке крови, ФНО – содержание фактора некроза опухоли в сыворотке крови, АПФ – ангиотензинпревращающий фермент в сыворотке крови, ИЛ-4, окЛПНП – окисление липопротеидов низкой плотности в сыворотке крови, КК– уровень каллекриина в сыворотке крови.

Таблица 1

Информативность признаков по пяти группам

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5

ТМТ	33,3	ТМТ	30,4	ТМТ	26,6	Каталаза	19,4	NO	34
Гаркави	14,2	NO	19,1	NO	15,3	NO	18	Каталаза	21,7
NO	11,4	Гаркави	11	Каталаза	15,3	ТМТ	17,5	Гаркави	12,3
Каталаза	7,9	Лептин	7,9	Гаркави	14,6	Гаркави	8,9	ТМТ	9,6
ФНО	7,8	Каталаза	5,9	Лептин	12,5	АПФ	4,5	АПФ	8,1
Лептин	6,2	АПФ	4,4	ФНО	10,5	ФНО	4,4	ФНО	5,8
КК	2,9	ФНО	4	АПФ	5,8	ИЛ-4	2,9	ИЛ-4	5,2
АПФ	2,5	ИЛ-4	4	окЛППНП	5,5	Лептин	2,3	КК	4,7
окЛППНП	2,2	КК	1	КК	5,2	окЛППНП	1,9	Лептин	2,8
ИЛ-4	2,2	окЛППНП	0,7	ИЛ-4	3	КК	1	окЛППНП	0,8

Для построения лиц Чернова использована библиотека TeachingDemos функция faces. Набор конструктивных параметров лиц соответствуют информативным признакам. Полученные визуальные образы позволяют сравнить результаты до и после лечения рис. 2.

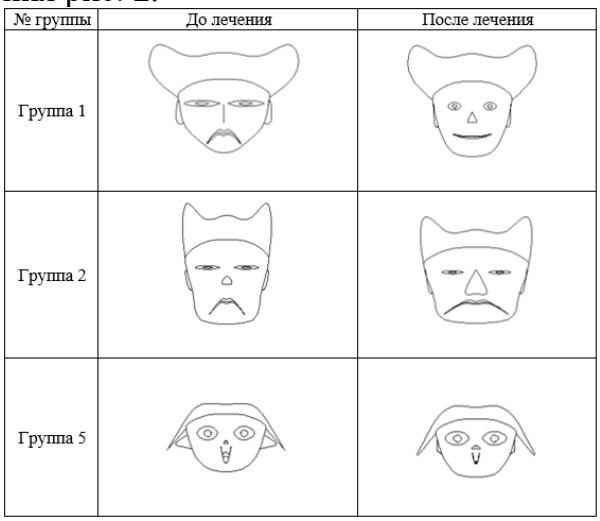

Рис. 2 Визуальные образы до и после лечения

Анализируя результаты можно заметить по каким группам лечение было наиболее эффективным. Можно выделить группу 1, по этой группе заметны значительные изменения. Визуальные образы группы 2 и группы 5 практически не изменились, следовательно, можно сделать вывод, что лечение не оказало положительного результата.

В результате выполнения работы, была разработана информационная технология анализа результатов лечения детей с эндокринопатиями.

С использованием разработанной информационной технологии было осуществлено восстановление пропусков в исходных наборах данных и проанализированы результаты лечения пациентов с использованием визуальных образов.

В рамках выполнения работы была построена структурная схема алгоритма информационной технологии и осуществлена её практическая реализация. Алгоритм информационной технологии представлен на рис.3.

ISSN 2658 – 7505
Выпуск №2, 2019
Электронный научный журнал «Вестник молодёжной науки России»

Рис. 3 – Алгоритм ИТ анализа результатов лечения детей

Разработанная информационная технология имеет широкую сферу применения и может использоваться для комплексного анализа данных в различных медицинских учреждениях.

Список цитируемой литературы

1. Всемирная организация здравоохранения / Ожирение и избыточный вес [электронный ресурс] – режим доступа: http://www.who.int/mediacentre/factsheets/fs311/ru/.
2. RStudio / Take control of your R code [Электрон. ресурс] - Режим доступа: https://www.rstudio.com/products/rstudio/.
3. Роберт И. Кабаков R в действии. Анализ и визуализация данных в R / Метод множественного восстановления данных / Москва – 2014 – С. 489.
4. Берестнева О.Г. Методы структурного анализа и визуализации экспериментальных данных в социальных и медицинских исследованиях / О.Г. Берестнева, И.А. Осадчая, А.Л. Бурцева. - Томск: Из-во Томского политехн. ун-та, 2014 – С. 17-19.
5. Гублер Е. В. Вычислительные методы анализа и распознавания патологических процессов – МЕДИЦИНА, 1978. – 198с.