ПРИМЕНЕНИЕ МАТЕМАТИКО – СТАТИСТИЧЕСКИХ МЕТОДОВ В ЗАДАЧЕ ИССЛЕДОВАНИЯ ЗНАЧЕНИЙ КЛИНИКО – ЛАБОРАТОРНЫХ ПОКАЗАТЕЛЕЙ ПАЦИЕНТОВ ДО И ПОСЛЕ ЛЕЧЕНИЯ

Ю.А. Емельянова, научный руководитель: О.В. Марухина
(г. Томск, Национальный исследовательский Томский политехнический университет)
yuliyaemelianova@yandex.ru, marukhina@tpu.ru

THE APPLICATION OF MATHEMATICAL AND STATISTICAL METHODS IN THE STUDY OF VALUE OF PATIENTS' CLINICAL AND LABORATORY INDICATORS BEFORE AND AFTER TREATMENT

Y.A. Emelianova, O.V. Marukhina
(Tomsk, Tomsk Polytechnic University)

Annotation: multivariate data analysis has been actively developing and applying practically in all fields of study latly. An important task in medicine during the study of diseases, the treatment of the patient is search and selection of informative features for reliable diagnosis setting.

Keywords: children and adolescent obesity, informative value, the Kulbak method, system of support of making medical decisions, Wilcoxon test.

Введение. В настоящее время вся информация хранится в электронном виде в базах данных и занимает большие объемы. Возможно, хранится информация, которая уже не будет использоваться и она является не актуальной. За счет того, что объем данных увеличивается в медицине и время на принятия решений сокращается, возникают различные диагностические ошибки и неверные назначения врачей. Решением данных проблем может быть внедрение в область медицины различных ИТ-решений.

В ходе анализа научных публикаций по данной теме выявлено, что разработки и исследования в этой области ведутся во всем мире и в различных направлениях. Существуют некоторые экспертные системы в области медицины: MYCIN - диагностирование бактерий и диагностика заболеваний свертываемости крови; DENDRAL - определение молекулярной структуры; CASNET - диагностика глаукомы; DXplain - поддержка клинических решений; Germwatcher - диагностика инфекций; Puff анализ нарушения дыхания. Среди российских разработок наиболее известны экспертные системы: ДИН (распознавание неотложного состояния) и ДИАГЕН (диагностика наследственных болезней у детей).

Новизна данного исследования заключается в создании нового алгоритма диагностики состояния пациента по степени ожирения на основе анализа клинико-лабораторных показателей. Так же в формировании уникальной базы знаний об исходах лечения на основе данных, собранных в НИИ Курортологии и физиотерапии.

Оценка информативности по Кульбаку. В медицине решать задачи диагностики, определения диагноза, распознавания заболевания или его отсутствия, можно только тогда, когда получены и проанализированы информативные признаки, присущие пациенту. Информативные признаки - полезная для данной цели информация, полученная из исходной информации. Поэтому из множества признаков необходимо определить наиболее информативные, которые характеризуют психофизическое состояние объекта. Ведь от сбора данных по этим признакам будет зависеть постановка диагноза. Существуют разные методы оценки информативности, например, метод Шеннона, метод накопленных частот, остановимся на методе оценки информативности по Кульбаку.

Метод Кульбака предлагает в качестве оценки информативности – меру расхождения между двумя классами, которая называется дивергенцией. Согласно этому методу информативность вычисляется по формуле 1:

$$J(x_i) = \sum_i J(x_{ij}) = \sum_i 10lg\frac{P(x_{ij}/A_1)}{P(x_{ij}/A_2)}0{,}5[P(x_{ij}/A_1) - P(x_{ij}/A_2)] \quad (1)$$

$J(x_i)$ –информативность признака,

P_1- вероятность попадания признака в первом классе A_1,

P_2- вероятность попадания признака во втором классе A_2,
j – номер диапазона признака x_i [1].

Метод Кульбака служит для определения информативности признака, который участвует в распознавании только двух классов и данный метод оперирует вероятностями, поэтому объемы выборки наблюдений признака по двум распознаваемым классам могут быть различны [2].

Томским НИИ Курортологии и физиотерапии были предоставлены данные для оценки информативности следующих групп: клиника, сердечно-сосудистая система, физическая работоспособность, липидный обмен, биохимия крови, гормональный статус, иммунологический статус, состояние калликреин-кининовой системы, окислительная способность плазмы крови. Данные они получили в ходе проведения клинического исследования в период с 2006 по 2013 год, в котором приняли участие 464 ребенка, страдающих ожирением в возрасте от 10 до 15 лет.

Информативность групп были получены в результате выполнения программы, созданной командой, занимающаяся данной темой в ТПУ. В результате была посчитана информативность для каждой группы признаков. Например, группа «Физическая работоспособность» включает в себя следующие признаки: толерантность к физической нагрузке, общая работоспособность, индекс инсулинорезистентности и двойное произведение (насыщение миокарда кислородом). Информативность данной группы представлена в таблице 1.

Таблица 1. Информативность группы

Признак	Информативность
ТФН	1,4
Общая раб-ть	0,99
HOMA	0,44
Дв.Пр.	0,01

Из таблицы видно, что информативным признаком в данной группе является ТФН (рис.1), менее информативным – Дв.Пр.(рис.2).

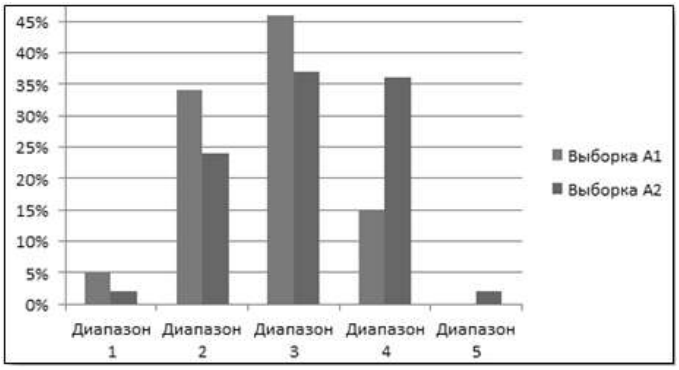

Рис.1. Признак «ТФН»

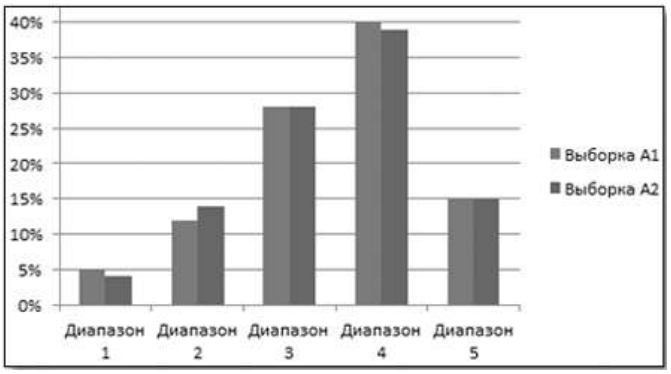

Рис.2. Признак «Дв.Пр»

Для каждой группы были построены графики наиболее информативного и менее информативного признака. Выборка А1 – значения до лечения, выборка А2 – значения после лечения. На рисунке 1 видим, что выборка А1 смещена в левую сторону, выборка А2 – в правую и заметна разница между выборками. Чем больше разница между выборками, тем информативнее признак. На рисунке 2 видим обратное, разница между выборками незначительная, следовательно, признак менее информативный.

Критерий Вилкоксона. Другой задачей исследования является применение метода оценки достоверности сдвига до и после воздействия лечения. Критерий Вилкоксона – статистический критерий, используемый для оценки различий между двумя независимыми выборками.

Проверяемая с помощью критерия Вилкоксона нулевая гипотеза состоит в том, что центры распределений, из которых происходят сравниваемые выборки, смещены относительно друг друга на некоторую величину. Алгоритм метода состоит в следующем:

- берется список данных из первой и второй выборки;
- находится разность по каждому замеру, вычитаются значения из второй выборки значения первой выборки;
- подсчитывается количество отрицательных и положительных значений, большее количество значений будет представлять из себя типичный сдвиг – наибольшее количество сдвигов;
- находятся значения разностей по модулю;
- все имеющиеся значения ранжируют, игнорируя их знак;
- ранги значений, принадлежащих к первой группе, суммируют, получая величину W;
- полученное W сравнивают со значением, которое можно было бы ожидать при верной нулевой гипотезе.

Расчеты выполнены с помощью скриптового языка R, предназначенному для статистической обработки данных, в среде разработки программного обеспечения RStudio. Для выполнения теста Вилкоксона в системе R используется функция wilcox.test() [3]. Переменная z- значение признака до лечения, переменная m – значение признака после лечения. Этим переменных присваиваем столбцы соответственно s (ТФН до лечения) и t (ТФН после лечения) из файла «data», содержащего данные по ожирению детей. На рис.3 представлен расчёт критерия Уилкоксона для признака ТФН из группы «физическая работоспособность».

Рис.3. Реализация в RStudio

В таблице 2 представлены результаты расчёта критерия Уилкоксона для группы «физическая работоспособность». В таблице используются следующие обозначения: W – это тестовая величина Уилкоксона, является суммой рангов в одной из двух выборок, p - величина, используемая при тестировании статистических гипотез, это вероятность ошибки при отклонении нулевой гипотезы.

Таблица 2. Критерий Уилкоксона

Признак	W	p	Сравнение с p=0,05
HOMA	17070	0,2835	p>0,05, следовательно, различия статистически не значимы, различия не достоверны.
ТФН	5055	0,0001	p<0,05 различия статистически значимы
Дв. Пр.	7042	0,9617	p>0,05 различия статистически не значимы
Общая р-ть	1175	0,1046	p>0,05 различия статистически не значимы

Из таблицы видим, что согласно полученному значению p (p-value=0,00001) признака ТФН до и после лечение, различия являются статистически значимы, т.е. различия достоверны, у остальных признаков различия между выборками не значимы.

Заключение. В результате в каждой группе определили наиболее информативные признаки:

- группа клиника – тощая масса тела (ТМТ);
- физическая работоспособность - толерантность к физической нагрузке;
- сердечно-сосудистая система - систолическое артериальное давление (САД);
- липидный обмен - липопротеиды низкой плотности;
- биохимия крови -щелочная фосфатаза в сыворотке крови;
- гормональный статус - ИНФ не >45 пг/мл;
- иммунологический статус - циркулирующие иммунные комплексы;
- состояние калликреин-кининовой системы - уровень каллекриина;
- окислительная способность плазмы крови - содержание оксида азота в крови.

При расчёте критерия Уилкоксона получили следующие результаты: в группе клиника – различия выборок избыток, индекс массы тела и тощая масса тела являются статистиче-

ски значимыми. Физическая работоспособность – значимое различие имеет признак толерантность к физической нагрузке, сердечно-сосудистая система – диастолическое и систолическое артериальное давление, липидный обмен - липопротеиды низкой плотности, в группах биохимия крови и гормональный статус получили результаты, что для всех показателей $p>0,05$, следовательно различия между выборками статистически не значимы. Иммунологический статус – значимые циркулирующие иммунные комплексы в сыворотке крови и концентрация иммуноглобулина А в сыворотке крови. В группах состояние калликреин-кининовой системы и окислительная способность плазмы крови для всех показателей $p<0,05$, следовательно различия статистически значимы.

Данные по результатам исследования будут использоваться при формировании базы знаний и создании интеллектуальной системе поддержки принятия врачебных решений, а так же учитываться при сборе необходимых признаков для вновь прибывших пациентов.

ЛИТЕРАТУРА

1. Гублер Е.В. Вычислительные методы анализа и распознавания патологических процессов, 1978, 269 с.

2. Голованова И.С. Выбор информативных признаков. Оценка информативности. [Электронный ресурс]. URL: http://ime.tpu.ru/study/discypliny/INF-PR.pdf (дата обращения: 10.10.2017).

3. Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R. [Электронный ресурс]. URL: http://r-analytics.blogspot.ru/ (дата обращения: 21.10.2017).

ЦИФРОВАЯ ЭКОНОМИКА: ОБЛАЧНЫЕ ТЕХНОЛОГИИ В ЗАДАЧАХ МАТЕМАТИЧЕСКОГО АНАЛИЗА В СФЕРЕ МЕДИЦИНЫ

Зимина Е.Ю.
НИУ ВШЭ, ezimina@hse.ru

DIGITAL ECONOMY: CLOUD TECHNOLOGIES IN PROBLEMS OF MATHEMATICAL ANALYSIS OF MEDICAL DATA

Zimina E. U.
Higher School of Economics, ezimina@hse.ru

Abstract. The article includes the observation of the cloud services and technologies usage in telemedicine. The article contains a review of mathematical analysis of medical data using cloud technology, which produces storage, analysis and forecasting on the basis of obtained data. In addition, the possibility of integrating cloud technologies with external systems is considered.

Key words: Telemedicine, digital economy, cardiology, big data

Введение

По прогнозу института McKinsey на 2025 год в сфере цифровой экономики наиболее значимыми областями применения технологий будут Mobile Internet, Automation of knowledge work, Internet of Things и Cloud. [1]

На рис 1 приводятся прогнозные оценки мирового рынка по каждой из названных технологий, суммарный объем рынка которых составит около 30 триллионов долларов. При том, что доля медицинского использования этих технологий оценивается ориентировочно в 30%.

Доля нефтегазового сектора мировой экономики оценивается при этом только в 1.5 триллиона долл.