УДК 004.67

ОТБОР НАИЛУЧШИХ ФАКТОРОВ, ОПРЕДЕЛЯЮЩИХ КЛАССИФИКАЦИЮ ОБЪЕКТОВ

Ливинская В.А., к.т.н., доц., Иванова П.Р., студ., Дашко В.С., студ.

Белорусско-Российский университет, г.Могилёв, Республика Беларусь

Реферат. *В статье представлены результаты применения инструментов FeatureSelection, а также непараметрических методов сравнения нескольких и двух независимых групп для отбора факторов, наилучшим образом, определяющих классификацию объектов в клиническом исследовании.*

Ключевые слова: прикладная статистика, непараметрические критерии, критерий Краскела-Уоллиса, критерий Манна-Уитни, критерий ХИ-квадрат, реографическая кривая.

Анализ данных в различных сферах человеческой деятельности приобретает большую популярность, благодаря современным информационным технологиям, позволяющим регистрировать, накапливать и хранить большие массивы информации. Появилось особое направление в естественных науках – Datascience, в котором реализуются как классические методы прикладной статистики, так и создаются новые методы, позволяющие работать с большими массивами (BigData).

В данной работе представлен один результат одного из этапов применения классических методов прикладной статистики в клиническом медицине.Речь идет об отборе факторов, оказывающих наибольшее влияние на вариацию категориального признака. В качестве исходных данных для анализа использовались результаты оценки при помощи программного обеспечения Реоспектр (нейрософтизменений сопротивления живой ткани в переменном электрическом поле высокой частоты спомощью реогепатографии (РГГ) пациентов, имеющих различные патологии и проходящих лечение в областной больнице г. Могилева):

- время распространения пульсовой волны от сердца (Qx);
- время быстрого кровенаполнения (α1);
- время медленного кровенаполнения (α2);
- время восходящей части волны (α);
- время общей систолы (Тобщ), длительность катакроты (Ткат);
- реографический индекс (РИ) – отношение амплитуды артериальной части волны (Аарт) к стандартному калибровочному импульсу;
- диастолический индекс (ДИА) – процентное отношение амплитуды дикроты к Аарт,
- максимальную скорость быстрого наполнения (Vмакс) – отношение амплитуды систолического максимума реограммы (Асист) к α1;
- среднюю скорость медленного наполнения (Vср) – отношение (Аарт-Асист) / α2.

В работе [1] предлагался в качестве фактора, являющегося предиктором классификации патологий использовать один агрегированный (вместо нескольких факторов)-площадь под кривой реограммы, схематическое представление на рисунке 1.

По имеющимся параметрам с помощью макроса VBA, были рассчитаны площади под кривой для каждого пациента с помощью метода площадей.

Рисунок 1 – Схематическое представление реографической кривой

Одной из первых задач исследования являлся отбор факторов, подтверждающий статистически значимое различие у пациентов, относящихся к группам с различными патологиями.

Поскольку данные параметры были исчислены для выборки из 87 пациентов, возникла необходимость в автоматизации обработки исходной информации. Для этих целей использовался современный язык для анализа статистических данных R, который является свободно распространяемым программным обеспечением и не требует приобретения лицензии.С помощью языка R, из отдельных файлов в текстовом формате, содержащих информацию о каждом пациенте, был сформирован датасет из 9 столбцов, восемь из которых являлись физическими параметрами кривой, а девятый отвечал за принадлежность пациента к одной из 2 групп: пациенты после с наличием одного из 3 патологических состояний: синдрома полиорганной дисфункции (ОАО СПОД), пациенты с циррозом печени (ОАО печень), пациенты с наличием хронических заболеваний желудка и поджелудочной железы (ОАО гастро), а также контрольная группа – практически здоровые. Отнесение пациента к одной из двух групп описывается бинарной переменной, принимающей два значения 1-пациент болен, 0-пациент здоров.

Далее использовался метод, реализованный в пакете FSelector. Так как результирующая переменная категориальная, был проведен тест Chi-squarenанезависимость между входом и выходом с оценкой p-value и, как следствие, бинарным выводом о значимости или незначимости отобранного набора признаков.

Результат отбора факторов (ранжированы в списке по убыванию значимости) представлены в таблице 1.

Как видно, все факторы оказывают статистически значимое влияние на принадлежность к группе (p-value<0,05), что подтверждает эффективность использования метода реографии для диагностикию

На следующем этапе ставилась задача отобрать факторы, наиболее влияющие на разделение пациентов на 4 группы (каждая из патологий рассматривалась как отдельная группа). В качестве группирующей переменной выбиралась категориальная переменная, принимающая 4 значения: норма, ОАО(печень), ОАО(гастро), ОАО(СПОД).

Таблица 1 – Результат ранжирования факторов по критерию принадлежности кгруппе здоровых пациентов

	Chi-square	p-value
Vср, Ом/с	101,5841	0,000000
Альфа2, сек	85,4666	0,000000
Vмакс, Ом/с	79,3226	0,000000
Альфа, сек	73,7892	0,000003
Площадь под кривой	66,3733	0,000001
РИ, у.е.	65,4901	0,000002
ДИА, %	58,5670	0,000021
Ткат, сек	56,6677	0,000709
Тобщ, сек	49,5584	0,005116
Альфа1, сек	42,7220	0,027900

Перед выбором метода, подтверждающего различие в двух независимых группах, в каждой из выборок проверялась гипотеза о принадлежности к нормальному распределению с помощью теста Шапиро-Уилка. Во всех случаях гипотеза о принадлежности была отвергнута, поэтому дальше применялись методы непараметрической статистики. Для проверки гипотезы об отсутствии различий показателей во всех группах одновременно, использовался непараметрический аналог дисперсионного анализа критерий Краскела – Уоллиса. Гипотеза была отвергнута во всех случаях, кроме показателя время быстрого кровенаполнения (α1). Этот фактор в дальнейшем был исключен из анализа. Для выявления различий между двумя группами – норма и отдельная патология применялся критерий Манна – Уитни. Результат проверки гипотезы об отсутствии различий (ошибка первого рода) представлен в таблице 2.

Таблица 2 – Результат проверки гипотезы об отсутствии различий (p-level)

группа	Фактор									
	Qx, сек	α2, сек	α, сек	Тобщ, сек	Ткат, сек	РИ, у.е.	ДИА, %	Vмакс, Ом/с	Vср, Ом/с	Площадь под кривой
ОАО(гастро)	0,16	0,00	0,00	0,12	0,03	0,00	0,00	0,00	0,00	0,00
ОАО(печень)	0,02	0,09	0,04	0,91	0,54	0,76	0,30	0,31	0,00	0,33
ОАО(СПОД)	0,12	0,00	0,00	0,00	0,02	0,00	0,74	0,00	0,36	0,00

Таким образом, по результатам наблюдений 86 пациентов, сопоставляя значения ошибки первого рода с принятым уровнем значимости 0,05 в качестве предиктора (т. е. фактора, наилучшим образом прогнозирующим отнесение к определенной группе больных) можно считать:

– для пациентов групп ОАО(гастро) и ОАО(СПОД) – площадь под реографической кривой как интегральный показатель остальных факторов;

– для пациентов группы ОАО(печень) таким фактором можно считать показатель средней скорости медленного наполнения (Vср, Ом/с) как фактор, имеющий самую малую ошибку первого рода.

Список использованных источников

1. Точило, С. А. Интегративный показатель состояния артериального печеночного кровотока у пациентов при критических состояниях / С. А.Точило и [др.]. // Вестник Витебского государственного медицинского университета. – 2019. – Т. 18. – № 3. – С. 52–60.

УДК 004.67

ROC-АНАЛИЗ КАК ИНСТРУМЕНТ БИНАРНОЙ КЛАССИФИКАЦИИ

Ливинская В.А., к.т.н., доц., Иванова П.Р., студ., Дашко В.С., студ.

Белорусско-Российский университет, г.Могилёв, Республика Беларусь

Реферат. *В статье представлено исследование, которое демонстрирует возможность применения метода площадей в анализе результатов клинических исследований.*

Ключевые слова: метод площадей, ROC-анализ, реографическая кривая.

Основной целью данной работы являлось выявление статистически значимых различий в показателях оценки гемодинамики печени с помощью реогепатографии (РГГ) пациентов, имеющих различные патологии и проходящих лечение в областной больнице г. Могилева [1]. Данная методика основана на фиксации изменений сопротивления живой ткани в переменном электрическом поле высокой частоты. С использованием программного обеспечения Реоспектр (Нейрософт) были получены определенные физические характеристики реографической кривой (рис. 1), описывающей динамику сопротивления живой ткани за определенный временной промежуток (табл. 1).

Таблица 1 – Параметры реографической кривой

Параметр РРГ	Обозначение
Амплитуды артериальной части волны	Аарт
Время распространения пульсовой волны от сердца	Q_x
Систолический максимум. Реограммы	Асист
Время восходящей части волны	α
Время быстрого кровенаполнения	$α_1$
Время медленного кровенаполнения	$α_2$
Время общей систолы	Тобщ
Длительность катакроты	Ткат