МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра «Биомедицинская инженерия»

УТВЕРЖДАЮ
Проректор по учебной работе
О.Г. Локтионова
2017г.

«МЕТОДЫ СБОРА И АНАЛИЗА МЕДИКО-БИОЛОГИЧЕСКОЙ ИНФОРМАЦИИ»

Методические указания по выполнению курсовой работы для студентов направления подготовки 12.03.04 “Биотехнические системы и технологии” (бакалавр)

Курск 2017

УДК 004.93:61

Составители: С.А. Филист, К.Д.А. Кассим

Рецензент:
Доктор технических наук, профессор *А.Ф. Рыбочкин*

Методы сбора и анализа медико-биологической информации: Методические указания по выполнению курсовой работы / Юго-Зап. гос. ун-т; сост.: С.А. Филист, К.Д.А. Кассим - Курск, 2017. 63 с.

Предназначены для студентов направления подготовки 12.03.04 “Биотехнические системы и технологии” (бакалавр) дневной и заочной форм обучения

Текст печатается в авторской редакции

Подписано в печать 10.11.17. Формат 60х84 1/16.
Усл.печ.л.1,0 Уч.-изд.л.0,9 Тираж 100 экз. Заказ 1799 Бесплатно.
Юго-Западный государственный университет.
305040, г.Курск, ул. 50 лет Октября, 94.

Введение

В последние годы значительно возросли требования к выпускникам высших технических учебных заведений. Непрерывный рост объема знаний, накопленных каждой отраслью науки, приводит к увеличению количества информации, которую нужно усвоить студенту за время обучения.

Объем профессиональных знаний, полученный за время обучения, составляют лишь исходный баланс, от которого начинается рост и формирование специалиста. Современный инженер должен уметь учиться самостоятельно, быть в курсе достижении современной науки, развивать творческое мышление, уметь обобщать литературные данные и личный опыт работы, уметь выступать перед аудиторией. Настоящий специалист должен обладать навыками исследователя, применять в своей работе элементы научного поиска.

Профессиональные знания инженера, вместе с другими техническими дисциплинами дисциплинами, формирует и дисциплина «Методы сбора и анализа медико-биологической информации», поскольку в настоящее время роль вычислительных средств в биомедицинских исследованиях особенно актуальна.

Одним из эффективных путей повышения качества подготовки молодых специалистов является всемирное развитие исследовательской работы студентов. Она проводится в виде научно-исследовательской работы, выполняемой во внеучебное время, и учебно-исследовательской работы, включённой в учебный процесс. Эти формы научной работы студентов являются обязательным компонентом подготовки молодых специалистов и широко используются на кафедре Биомедицинской инженерии Юго-Западного государственного университета.

Курсовая работа, при охвате всего контингента обучающихся на кафедре, является одной из важных форм приобретения студентами умений и навыков по самостоятельному сбору, оформлению и представлению экспериментального материала.

Курсовая работа дисциплины «Методы сбора и анализа медико-биологической информации» имеет целью закрепление теоретических знаний, полученных в ходе прослушивания лекционного курса, а также приобретение новых навыков при

работе с реальными экспериментальными данными, полученными самостоятельно. Курсовая работа выполняется в соответствии с определенными правилами и ГОСТами.

1.1. Состав проекта

Курсовая работа должна состоять из пояснительной записки объемом не менее 20 листов машинописного текста, выполненного на листах белой бумаги формата А4 без нанесения рамки и основной надписи. При необходимости работа может иллюстрироваться плакатами или чертежами формата А1. Оформление чертежей и пояснительной записки должно соответствовать стандартам ГОСТ, ЕСКД, а также другим отраслевым стандартам.

1.2. Структура пояснительной записки

Пояснительная записка должна содержать в себе следующие пункты:

1. Титульный лист
2. Задание на курсовую работу
3. Техническое задание
4. Содержание
5. Введение
6. Материалы «Состояние вопроса»
7. Основные разделы проекта (в зависимости от темы)
8. Заключение
9. Список используемых источников
10. Приложения (в случае необходимости)

Во введении дается укрупненное технико-экономическое обоснование разработки, ее актуальность, формулируется цель работы, вытекающая из актуальности решаемой задачи.

Раздел «Состояние вопроса» содержит современное состояние в данной области с позиций литературных источников. Кроме того, в данном разделе следует перечислить методы и способы анализа исходных данных. Данный раздел следует заканчивать формулировками задач работы, т.е. формулировками последовательности действий, которые должны привести к достижению поставленной цели.

Основные разделы работы определяются решаемыми проблемами и задачами. Здесь может быть множество разделов, например аналитический, расчетный, экспериментальный и другие. Каждый из разделов должен иметь заголовок, отражающий конкретное содержание раздела, например «предварительная обработка таблиц экспериментальных данных». Записи «Аналитический раздел» и прочие не допускаются.

В заключении приводится развернутая последовательность действий, которая была проделана для достижения поставленной цели. Иными словами, следует привести последовательности расчета и синтеза, начиная от предварительного анализа данных и заканчивая конечной математической моделью.

УРОВЕНЬ СЛОЖНОСТИ: «ВЫСОКИЙ»

2. Теоретические основы синтеза признакового пространства для оценки адаптационных свойств организма человека на основе данных, получаемых из пальцевой фотоплетизмограммы

2.1. Исследование носителей информативных параметров в сигнале фотоплетизмограммы

2.1.1. Кодовые точки фотоплетизмограммы

Информативные параметры фотоплетизмограммы группируются по двум признакам:

1. По вертикальной оси исследуются амплитудные характеристики пульсовой волны, соответствующие анакротическому и дикротическому периоду. Несмотря на то, что эти параметры являются относительными, их изучение в динамике предоставляет ценную информацию о силе сосудистой реакции. В этой группе признаков изучаются амплитуда анакротической и дикротической волны, индекс дикротической волны. Последний показатель имеет абсолютное значение и имеет собственные нормативные показатели.
2. По горизонтальной оси исследуются временные характеристики пульсовой волны, предоставляющие информацию о длительности сердечного цикла, соотношении и длительности систолы и диастолы. Эти параметры имеют абсолютные значения и могут сравниваться с существующими нормативными показателями. В этой группе параметров изучаются длительность анакротической фазы пульсовой волны, длительность дикротической фазы пульсовой волны, длительность фазы изгнания, длительность пульсовой волны, индекс восходящей волны, время наполнения, продолжительность систолической фазы сердечного цикла, продолжительность диастолической фазы сердечного цикла, время отражения пульсовой волны, частота сердечных сокращений.

На рисунке 2.1. представлены основные кодирующие точки объемного пульса.

Точка В1 соответствует началу периода изгнания систолического периода, точка В2 соответствует моменту максимального расширения сосуда в фазу форсированного изгнания, точка В3 соответствует протодиастолическому периоду, точка В4 соответствует началу диастолы, точка В5 соответствует наступлению конца диастолы и указывает на завершение сердечного цикла.

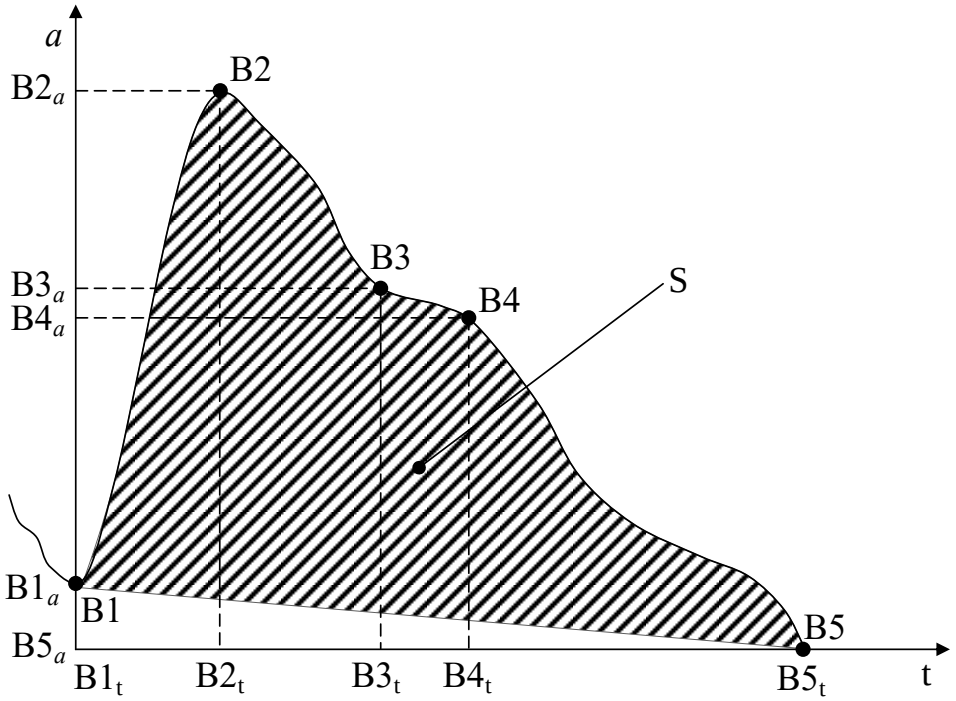

Рисунок 2.1 – Основные кодирующие точки объемного пульса

На рисунке 2.1 можно выделить амплитудные параметры фотоплетизмограммы (ось ординат *a*) и временные параметры фотоплетизмограммы (ось абсцисс t).

2.1.2. Амплитудные параметры фотоплетизмограммы

Амплитуда пульсовой волны (более точное название – «амплитуда анакротической фазы»).

- Измеряется в относительных единицах;
- Значение вертикальной оси вычисляется по формуле:

$$АПВ = B2_a - B1_a, \qquad (2.1)$$

- Нормативных значений не имеет, оценивается в динамике.

Амплитуда дикротической волны.

- Измеряется в относительных единицах;
- Значение вертикальной оси вычисляется по формуле:

$$АДВ = В4_а - В5_а, \quad (2.2)$$

- В норме составляет 1/2 от величины амплитуды пульсовой волны.

Высота инцизуры.

- Измеряется в относительных единицах;
- Значение вертикальной оси вычисляется по формуле:

$$ВИ = В3_а - В5_а, \quad (2.3)$$

- В норме составляет 2/3 от величины амплитуды пульсовой волны.

Индекс дикротической волны.

- Измеряется в процентах;
- Значение вертикальной оси вычисляется по формуле:

$$ИДВ = (В3_а - В5_а)/(В2_а - В1_а)\cdot 100, \quad (2.4)$$

- Нормативное значение составляет 63 - 73%;

2.1.2. Временные параметры фотоплетизмограммы

Длительность анакротической фазы пульсовой волны.

- Измеряется в секундах;
- Значение горизонтальной оси вычисляется по формуле:

$$ДАФ = В3_t - В1_t. \quad (2.5)$$

- Нормативное значение не установлено.

Длительность дикротической фазы пульсовой волны.

- Измеряется в секундах;
- Значение горизонтальной оси вычисляется по формуле:

$$ДДФ = В5_t - В3_t, \quad (2.6)$$

- Нормативное значение не установлено.

Длительность фазы изгнания.

- Параметр, отражающий диастолическую активность;
- Измеряется в секундах;
- Значение горизонтальной оси вычисляется по формуле:

$$ДФИ = В5_t - В3_t, \quad (2.7)$$

- Нормативное значение не установлено.

Длительность пульсовой волны.

- Измеряется в секундах;
- Значение горизонтальной оси вычисляется по формуле:

$$ДПВ = В5_t - В1_t. \quad (2.8)$$

Нормативные значения этого параметра представлены в таблице 2.1.

Таблица 2.1 – Нормативные значения по возрастным группам

Возраст, лет	Длительность пульсовой волны, с
0 - 1	0,43 - 0,50
1 - 3	0,50 - 0,57
3 - 5	0,57 - 0,60
5 - 8	0,60 - 0,67
8 - 10	0,67 - 0,70
10 - 20	0,70 - 1,00
20 - 30	1,00 - 0,92
30 - 40	0,92 - 0,88
40 - 50	0,88 - 0,83
50 - 60	0,83 - 0,75
60 - 70	0,75 - 0,71
80 - 90	0,73 - 0,70

Индекс восходящей волны.

- Отражает фазу наполнения в систолический период сердечного цикла, соответствует отношению длительности восходящего сегмента анакротической волны к общей длительности пульсовой волны;

- Значение горизонтальной оси;
- На пульсовой волне вычисляется по формуле:

$$ИВВ = (В2_t - В1_t)/(В5_t - В1_t)\cdot 100, \quad (2.9)$$

- Нормативное значение соответствует 15 - 24%.

Время наполнения.

- Измеряется в секундах;
- Значение горизонтальной оси;
- Соответствует промежутку от начала пульсовой волны до вершины анакротической волны вычисляется по формуле:

$$ВН = В2_t - В1_t, \quad (2.10)$$

- Нормативное значение находится в пределах 0.06 - 0.12 с.

Продолжительность систолической фазы сердечного цикла.

- Измеряется в секундах;
- Значение горизонтальной оси вычисляется по формуле:

$$ДС = В4_t - В1_t, \quad (2.11)$$

- Нормативный параметр вычисляют как произведение длительности ДПВ и 0,324.

Продолжительность диастолической фазы сердечного цикла.

- Измеряется в секундах вычисляется по формуле:

$$ДД = В5_t - В4_t, \quad (2.12)$$

- В норме равна остатку вычитания длительности систолы от общей продолжительности пульсовой волны.

Время отражения пульсовой волны.

- Измеряется в секундах;
- Соответствует времени расслабления миокарда в протодиастолическую фазу;
- Значение горизонтальной оси вычисляется по формуле:

$$ВОВ = В4_t - В2_t, \quad (2.13)$$

- Нормативное значение лежит в диапазоне 0,03 – 0,04 с.

Частота сердечных сокращений.

- Измеряется в ударах в минуту.
- Значение горизонтальной оси вычисляется по формуле:

$$ЧСС = 60 / ДПВ . \quad (2.14)$$

Нормативные значения приведены в таблице 2.2.

Таблица 2.2 – Нормативные значения частоты сердечных сокращений по Кассирскому:

Возраст, лет	ЧСС в мин $^{-1}$
0 - 1	140-120
1 - 3	120-105
3 - 5	105-100
5 - 8	100-90
8 - 10	90-85
10 - 20	85-60
20 - 30	60-65
30 - 40	65-68

Несмотря на большое количество представленных параметров, большинство из них требует вычисление кодовых точек B3 и B4, которые определяют координаты инцизуры. Исследования показали, что эти кодовые точки локализовать очень сложно, поэтому необходим поиск информативных признаков, вычисляемых без использования координат этих точек.

2.1.3. Амплитудно-временные параметры фотоплетизмограммы

В качестве амплитудно-временных параметров фотоплетизмограммы нами был предложен параметр, характеризующий площадь фотоплетизмограммы (площадь фигуры, заштрихованной на рисунок 2.1).

- Измеряется в относительных единицах;
- Размерность: (амплитуда) х (время) вычисляется по формуле:

$$S = \sum_{i=1}^{N} (a_i - (\frac{B5_a - B1_a}{B5_t - 1} i + B5_a - B5_t \frac{B1_a - B5_a}{1 - B5_t})), \qquad (2.15)$$

где a_i – величина i - го отсчета фотоплетизмограммы, N - число отсчетов в анализируемой фотоплетизмограмме в интервале $[B1_t, B5_t]$;

- Нормативных значений не имеет, оценивается в динамике.

2.2. Разработка способов выделения информативных параметров фотоплетизмосигнала

Как показал анализ информативных параметров фотоплетизмограммы, выполненный в разделе 2.1, в качестве носителей информации могут быть использованы как амплитудные, так и временные параметры фотоплетизмограммы, поэтому для исследования фотоплетизмограммы необходимо использовать комбинацию амплитудных и временных методов анализа.

2.2.1. Техника измерения фотоплетизмограммы, используемой для классификации адаптационного резерва организма

При определении адаптационного резерва в большинстве случаев используют динамические показатели информативных признаков, получаемые в процессе проведения нагрузочных проб, которые характеризуют изменение информативного признака после нагрузочной пробы или время его восстановления после нагрузочной пробы до исходного значения или до некоторого уровня относительно исходного значения. При этом основное требование к нагрузочной пробе состоит в том, чтобы она не переводила функциональные системы исследуемого из одного класса функционального состояния в другой. Кроме того, необходимо исследовать толерантность и чувствительность каждого информативного признака к такой нагрузочной пробе.

Таким образом, в процессе исследования необходим использовать две нагрузочные пробы, одна из которых позволили бы управлять адаптационным резервом организма, а другая являлась бы индикатором уровня адаптационного резерва.

В процессе исследования адаптационного резерва необходимо перевести испытуемого из одного класса функционального состояния регуляторных систем в другой. Для этого использовался велоэргометр с функциями контроля частоты сердечных сокращений.

В качестве индикатора функционального состояния было предложено использовать окклюзионную фотоплетизмографию. Известная методика окклюзионной фотоплетизмографии заключается в следующем: на уровне верхней трети плеча накладывается тонометрическая манжета и в нее нагнетается воздух до давления, на 30 мм рт.ст. превышающее артериальное давление. Давление в манжете сохраняется в течение 5 минут, затем воздух быстро стравливается. В течение первых 30 секунд в норме возникает пиковая объемная и линейная скорости кровотока, постепенно снижающиеся к 3-й минуте.

На основании этой методики нами предложена следующая техника получения фотоплетизмограммы. Для этого используется плечевой автоматический тонометр, например, МТ-40. На большой палец испытуемого накладывается датчик фотоплетизмограммы, а на левое плечо окклюзионная манжета. После этого записывается фотоплетизамосигнал в течение одной минуты. Затем включается автоматический измеритель артериального давления. Процесс измерения давления длится приблизительно одну минуту. После чего в течение одной минуты записывается фотоплетизмосигнал после окклюзии.

На рисунке 2.2 показана запись фотоплетизмограммы на трех вышеперечисленных стадиях эксперимента.

В качестве информативных дополнительных признаков, которые получаются в процессе проведения эксперимента, могут быть использованы ЧСС, САД и ДАД.

Прежде чем приступить к анализу фотоплетизмограммы, выберем апертуру ее анализа. Так как волны третьего порядка наблюдать весьма проблематично, то ограничимся анализом волн первого и второго порядка. Учитывая, что средняя частота колебаний, соответствующих волнам второго порядка, составляет 0,2 Гц, ограничимся апертурой наблюдения фотоплетизмограммы 30 с, на которой могут разместиться, в среднем, шесть дыхательных циклов.

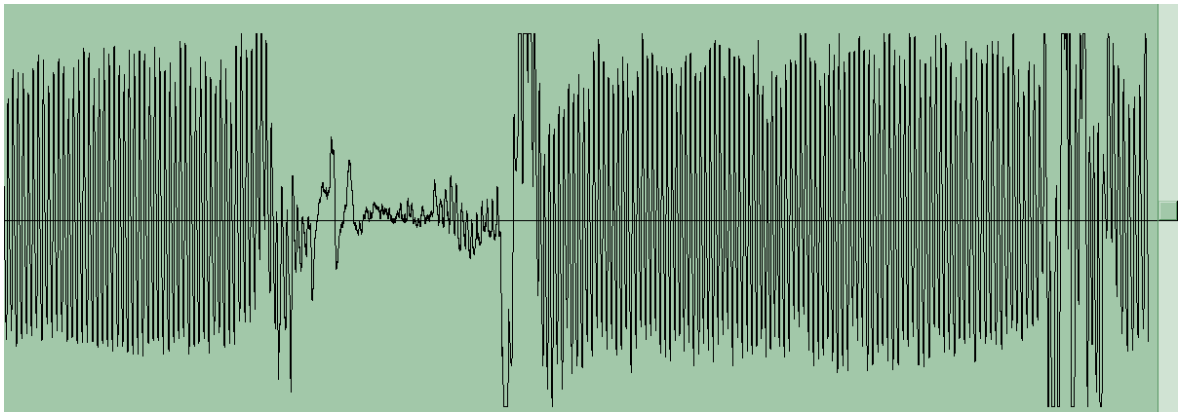

Рисунок 2.2 – Запись фотоплетизмограммы в процессе эксперимента по предложенной методике

В качестве иллюстрации, подтверждающий верность принятого решения, на рисунке 2.3 показан амплитудный спектр Фурье в окне длиной 30000 отсчетов, а на рисунке 2.4 показан амплитудный спектр Фурье того же самого сигнала, но в окне длиной 3000 отсчетов.

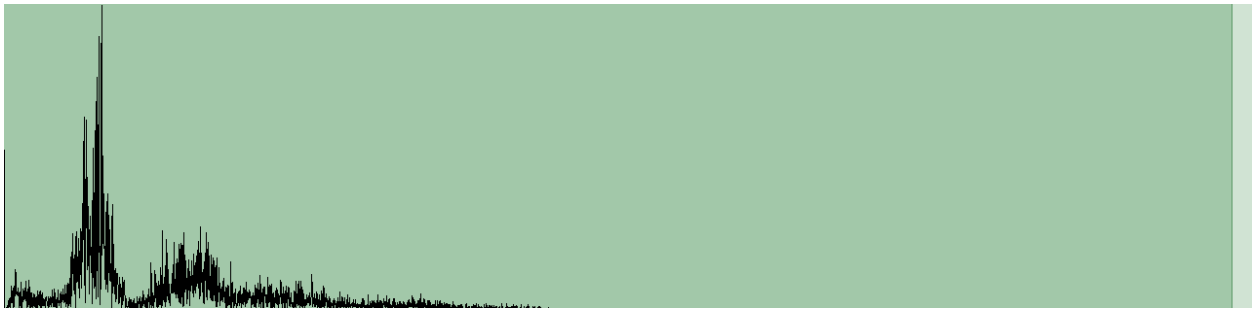

Рисунок 2.3 – Амплитудный спектр Фурье фрагмента фотоплетизмограммы(30000 отсчетов, частота дискретизации 100 Гц)

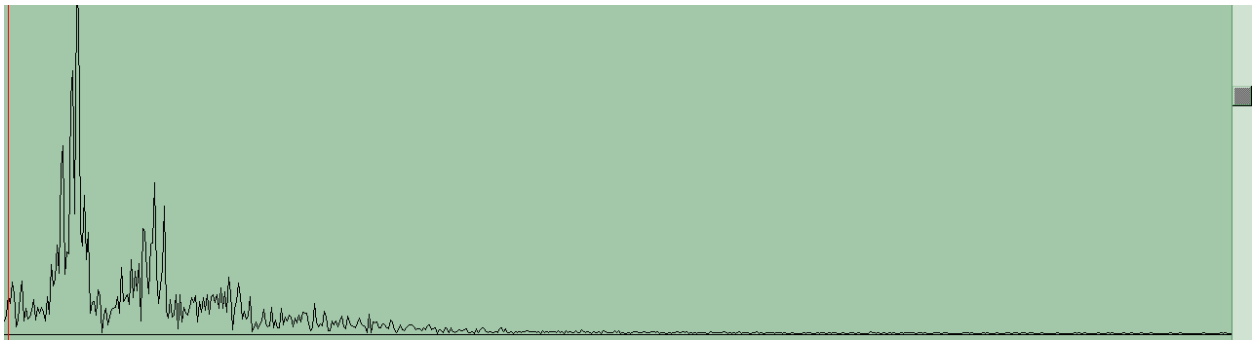

Рисунок 2.4 – Амплитудный спектр Фурье фрагмента фотоплетизмограммы (3000 отсчетов, частота дискретизации 100 Гц)

Представленные рисунки показывают, что увеличение длины окна не оказывает существенного влияния на структуру сигнала, более того, статистические исследования аналогичных спектров фотоплетизмограмм различных пациентов показали, что с ростом ширины окна третья гармоника кардиосигнала становится менее выраженной.

2.2.2. Выделение информативных параметров в амплитудно-временном пространстве

В качестве информативного параметра фотоплетизмограммы в амплитудно-временной области используем параметр S, определяемый согласно выражению (2.15), способ получения которого иллюстрирует рисунок 2.1. Для вычисления этого параметра необходимо определить кодовые точки фотоплетизмограммы, показанные на рисунке 2.1. Так как выражение (2.15) использует всего две кодовые точки, то этот процесс сводится к тривиальной сегментации фотоплетизмограммы.

На рисунке 2.5 приведена иллюстрация работы одного из предложенных и исследованных в работе алгоритма сегментации на фрагменте фотоплетизмограммы длительностью 30 с, а на рисунке 2.6 показан график параметров S, полученный на том же самом фрагменте фотоплетизмограммы.

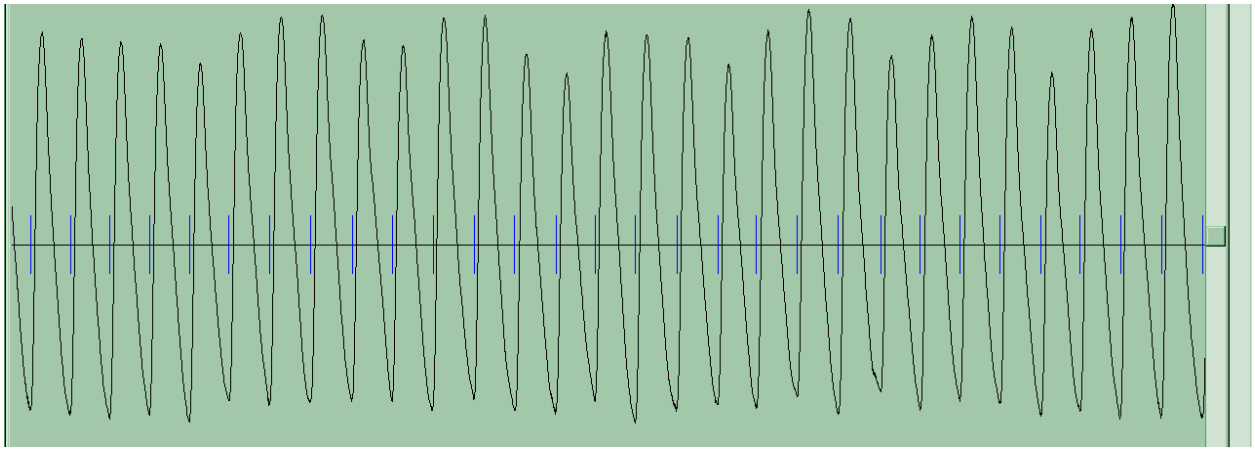

Рисунок 2.5 – Фрагмент сегментированного фотоплетизмосигнала (сегментация по кодовым точкам B1 и B5)

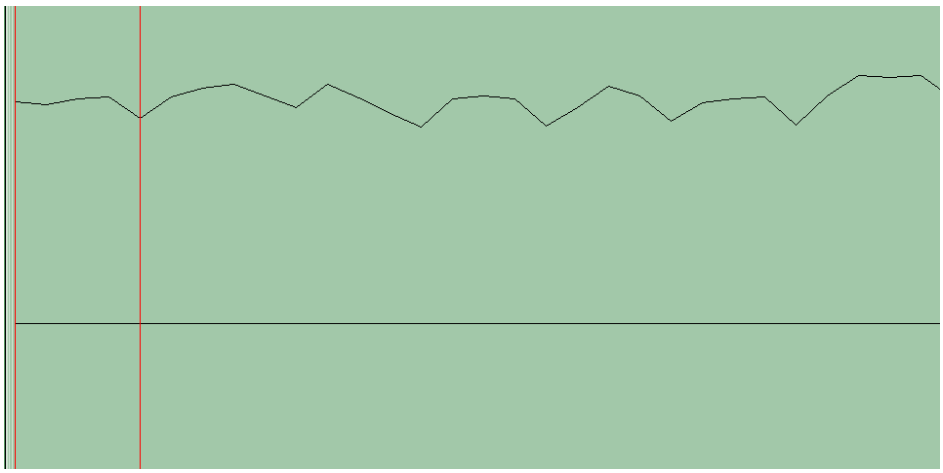

Рисунок 2.6 - График изменения параметра S на фрагменте фотоплетизмограммы рисунка 2.5: j – номер кардиоцикла

В качестве информативного параметра используем среднее значение параметра S на 30-секундном фрагменте, определяемого как:

$$X1 = \overline{S} = \frac{1}{N}\sum_{j=1}^{N} S(j), \qquad (2.16)$$

где N – число целых кардиоциклов на интервале 30 с.

2.2.3. Аплитудно-частотные информативные параметры

Одной из важнейших характеристик фотоплетизмограммы является ее спектр. Морфологию спектра фотоплетизмограммы иллюстрирует рисунок 2.7.

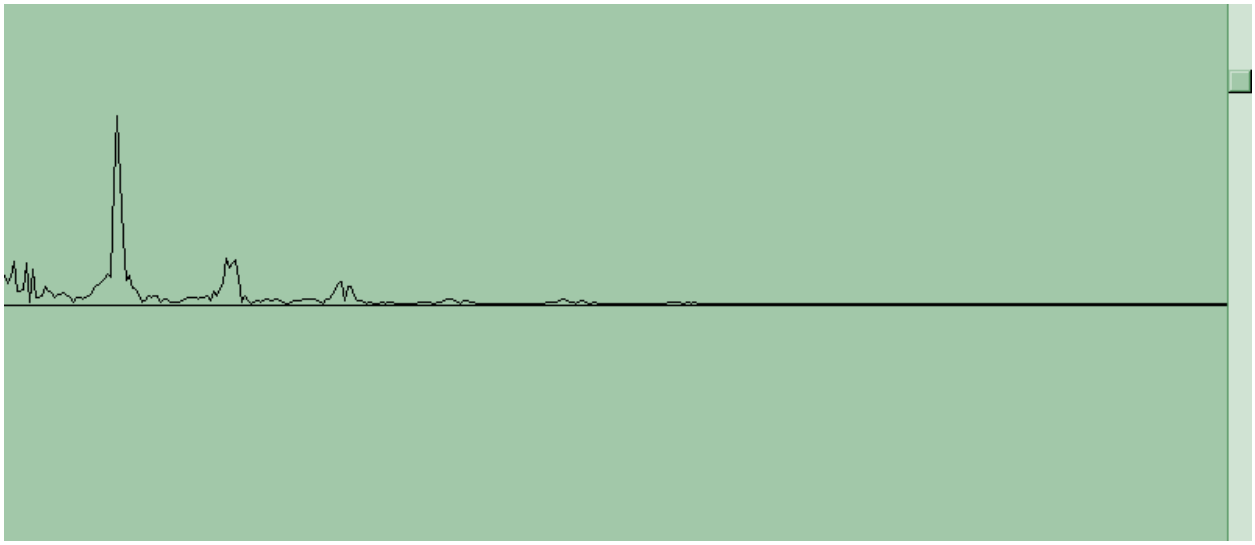

Рисунок 2.7 – Амплитудный спектр Фурье фотоплетизмограммы

Спектр Фурье фотоплетизмограммы определен в окне шириной 30 с. По энергетическому составу в нем преобладают волны первого порядка и их высшие гармоники. Поэтому, исходя из рисунка 2.7, в оконном спектре фотоплетизмограммы можем выделить сегмент дыхательной волны (I), сегмент первой гармоники кардиоцикла (II) и сегменты высших гармоник кардиоцикла (III). В третьем сегменте могут быть от одной до пяти гармоник.

В качестве информативных параметров выбираем амплитудные и частотные параметры спектра.

Так как амплитуда первой гармоники кардиоцикла максимальна, то все остальные амплитудные параметры целесообразно нормировать относительно этой гармоники, то есть амплитуда первой гармоники не рассматривается как информативный параметр. Это объясняется тем, что амплитуды гармоник зависят от ряда мешающих факторов, которые будут рассмотрены в следующем разделе. Следовательно, амплитудные параметры спектра фотоплетизмограммы дадут столько информативных признаков, сколько гармоник в третьем сегменте фотоплетизмограммы. В то же время, первая гармоника весьма полезна тем, что ее координата локализуются с высокой точностью за счет высокой амплитуды и, локализовав ее координату, можно определить координаты кратных ей гармоник третьего сегмента с достаточной точностью, несмотря на возможность высокого зашумления.

Следовательно, принимая во внимание только три первых гармоники кардиоцикла, получаем два информативных параметра, определяемых по следующим формулам:

$$X2 = \frac{|F(\omega)|_{2\max}}{|F(\omega)|_{1\max}}, \qquad (2.17)$$

$$X3 = \frac{|F(\omega)|_{3\max}}{|F(\omega)|_{1\max}}, \qquad (2.18)$$

где $|F(\omega)|_{i\max}$ – модуль максимальной амплитуды в полосе i-й гармоники кардиоцикла спектра фотоплетизмограммы.

Такое нормирование еще полезно и тем, что если в спектре отсутствует одна или несколько кратных гармоник, то информативные признаки, связанные с этими параметрами спектра фотоплетизмограммы, принимают значения нуля.

Частотные параметры спектра фотоплетизмограммы привязаны также к ЧСС. На каждой гармонике фотоплетизмограммы выделяем два информативных параметра:

1) ширина гармоники (ширина частотного диапазона, который занимает гармоника);

2) частота гармоники.

Рисунок 2.8 иллюстрирует методику определения амплитудных и частотных параметров спектра фотоплетизмограммы, которая может быть отнесена к любой гармоники фотоплетизмограммы.

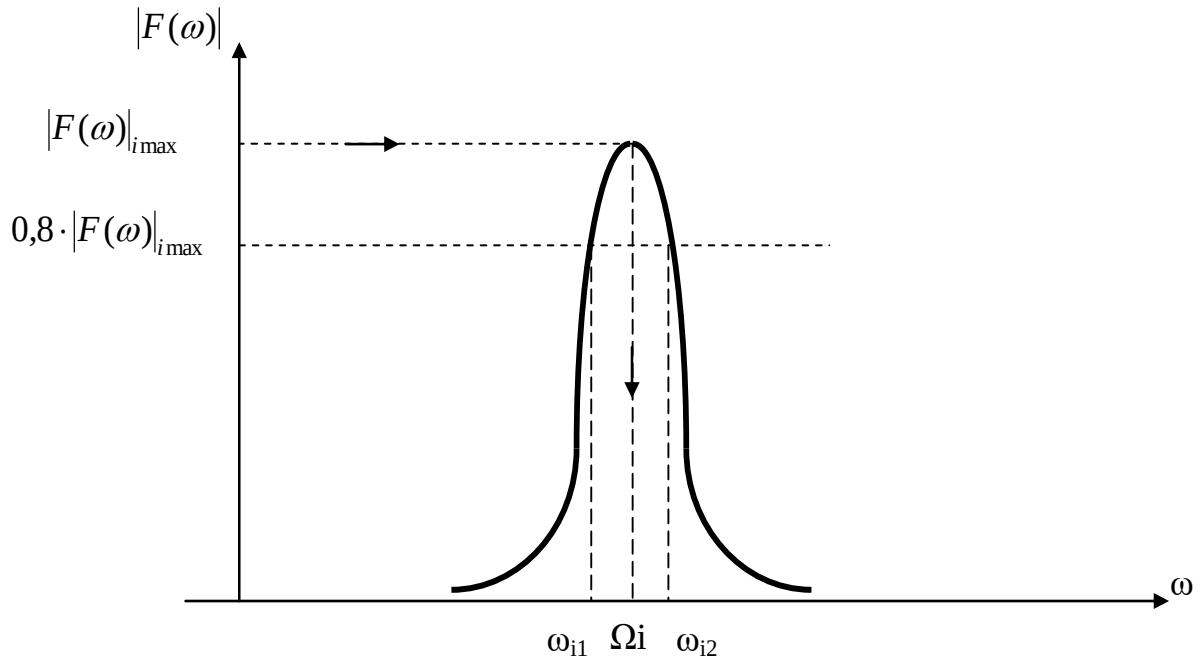

Рисунок 2.8 – Способ определения абсцисс полосы частот, занимаемой гармоникой фотоплетизмограммы

Все гармоники кардиоцикла на фотоплетизмограмме размыты, то есть занимают некоторую частотную полосу (это объясняется нестационарностью сигнала в окне). На рисунке эта полоса ограничена координатами ω_{i1} и ω_{i2}.

Информативные параметры, соответствующие частотным координатам гармоник кардиоцикла фотоплетизмограммы, определяются по следующим формулам:

$$X4 = \omega_{12} - \omega_{11}, \quad (2.19)$$

$$X5 = \omega_{22} - \omega_{21}, \quad (2.20)$$

$$X6 = \omega_{32} - \omega_{31}, \quad (2.21)$$

$$X7 = \Omega_1 = (\omega_{12} + \omega_{11})/2, \quad (2.22)$$

$$X8 = \Omega_2 = (\omega_{22} + \omega_{21})/2, \quad (2.23)$$

$$X9 = \Omega_3 = (\omega_{32} + \omega_{31})/2. \quad (2.24)$$

Кроме спектральных полос гармоник кардиоцикла на спектрограмме фотоплетизмограммы имеются спектральные полосы, вызванные дыхательным циклом (0,33…0,4 Гц приблизительно) и спектральная полоса 0,1 Гц. Координаты этих спектральных полос определяются эмпирически, так как не у всех пациентов они ярко выражены. Поэтому в качестве информативного параметра может быть использована только спектральная плотность в соответствующем диапазоне частот.

В связи с вышеперечисленными сложностями их локализации, в данной работе эти информативные параметры в признаковое пространство не включены.

2.2.4. Помехи при измерении фотоплетизмосигнала по предлагаемой методики

Помехи, оказывающие влияние на фотоплетизмосигнал, могут быть разделены на три категории.

К первой категории относятся помехи аппаратурного плана: шумы датчика и усилителя, помехи по питанию (от сети 220 В 50 Гц) и т.п.

Ко второй категории отнесем помехи, связанные с внешними источниками электромагнитного излучения, которые фиксирует фотоэлектрический датчик фотоплетизмограммы. Наиболее существенными помехами здесь могут быть электромагнитные излучения люминесцентных ламп и экранов мониторов.

К третьей категории отнесем помехи, связанные с биообъектом. К ним относятся механические помехи, вызванные

нестабильностью контакта фотоэлектрического датчика с поверхностью биообъекта (они особенно проявляются после физической нагрузки, когда человек устал и ему трудно фиксировать конечности в стабильном состоянии), и температура конечностей, которая существенно влияет на амплитуду пульсовой волны.

Помехи первой категории слабо влияют на фотоплетизмосигнал, так как их спектр лежит значительно выше спектра полезного сигнала. Они могут только привести к перегрузке измерительного тракта, поэтому должны подавляться в самом измерительном тракте специально предусмотренными схемотехническими решениями.

Помехи второй категории могут оказать существенное влияние на форму фотоплетизмограммы, так их частота хотя и выше спектра полезного сигнала, но соизмерима с ним. На рисунке 2.9 вверху показан фотоплетизмограмма и ее спектр при помехах такого рода. Внизу приведены фотоплетизмограмма и ее спектр того же самого пациента, полученные через несколько дней при отсутствии помех.

На верхнем рисунке справа хорошо видно, что спектр помехи значительно сдвинут вправо относительно спектра полезного сигнала. Спектр полезного сигнала и на верхнем и на нижнем рисунке занимает полосу приблизительно 8 Гц, а спектр помехи на верхнем рисунке сосредоточен вблизи 15 Гц.

Чтобы избавиться от этих помех, желательно в процессе эксперимента не включать лампы дневного света и проводить эксперимент на значительном расстоянии от экрана монитора.

Помехи третьей категории оказываю существенное влияние на амплитуду и форму пульсовой волны. На рисунке 2.10 показано три фрагмента фотоплетизмограмм с помехами такого рода.

Такие помехи затрудняют, а порой и делают невозможной, сегментацию фотоплетизмограмм и требуют усложнения алгоритмов сегментации, результаты работы которых представлены на рисунке 2.5. Радикальный способ борьбы с ними – отказ от использования параметров, связанных с интенсивностью сигнала. Однако интенсивность сигнала фотоплетизмограммы на определенных сегментах является важным информативным параметром.

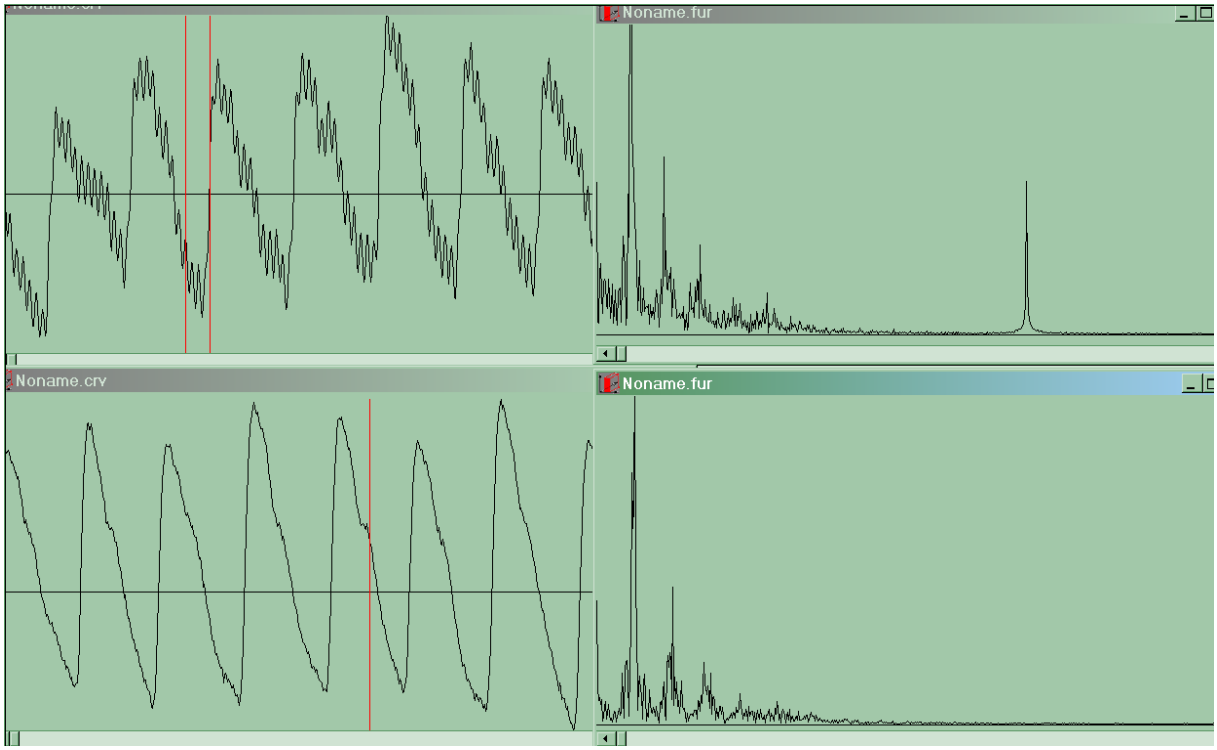

Рисунок 2.9 – Фотоплетизмосигнал и его спектр при наличии помех кадровой развертки монитора (вверху) и фотоплетизмосигнал и его спектр при отсутствии помех от кадровой развертки монитора (внизу)

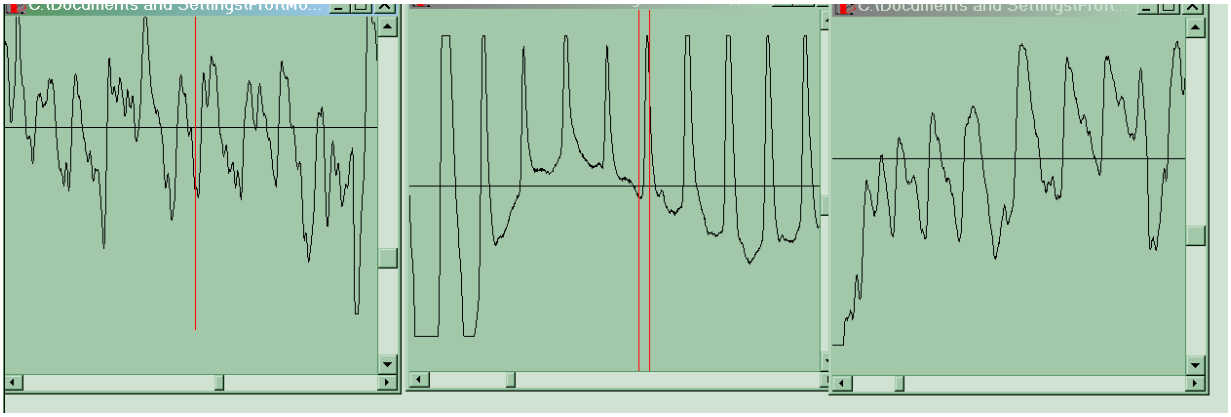

Рисунок 2.10 – Фрагменты фотоплетизмограмм с помехами третьей категории

2.3. Синтез пространства информативных признаков

Прежде чем приступать к синтезу признакового пространства, определим его размерность. Размерность признаковых пространств, используемых при оценке адаптационного резерва человека, варьируется в широких пределах. Преобладают одномерные признаковые пространства, в частных случаях с агрегированным

признаком. Целесообразно ограничиться трехмерным признаковым пространством. Это объясняется тем, что это наименьшая размерность, позволяющая наглядное представление классов и при возникшей необходимости размерность признакового пространства всегда можно увеличить за счет апробированных признаков.

Но прежде чем привести признаковое пространство к такому удобному виду, синтезируем признаковое пространство максимальной размерности путем включения в него всевозможных информативных параметров фотоплетизмограммы, полученных в разделе 2.2, а затем, при возможности, сократим его, до требуемого размера путем удаления из него малоинформативных и линейно зависимых информативных признаков.

Поэтому на первом этапе синтеза признакового пространства выявим все признаки, которыми может быть описана фотоплетизмограмма. Как было установлено в разделе 2.2, для описания фотоплетизмограммы до третьей гармоники включительно могут быть использованы девять информативных признаков. Важно, что восемь из них (X2…X9) не привязаны к амплитуде сигнала, то есть на них практически не оказывают влияния помехи, рассмотренные в разделе 2.2.4. Что же касается информативного признака X1, то он связан с амплитудой пульсовой волны, следовательно, значительно подвержен помехам третьей категории, особенно он чувствителен к температуре конечностей. Но отказываться от него не представляется возможным, так как это единственный признак, который несет информацию об интенсивности пульсовой волны в синтезируемом признаковом пространстве во временной области.

Исследования показали, что этот параметр очень чувствителен к окклюзии, что иллюстрирует рисунок 2.11. Следовательно, его целесообразно измерять дважды: до окклюзии и после окклюзии. Так как эти измерения проходят практически одновременно, то это частично позволяет отстроиться от помех, связанных с температурой конечностей. Затем из двух измеренных параметров получим один параметр, независящий от помех третьей категории. Один из возможных вариантов предлагаемого решения описывает следующая формула:

$$Y1 = (\sum_{i=1}^{10} S_i(t1) - \sum_{i=1}^{10} S_i(t2)) / \sum_{i=1}^{10} S_i(t1), \qquad (2.25)$$

где t1- интервал времени до окклюзии, t2- интервал времени после окклюзии.

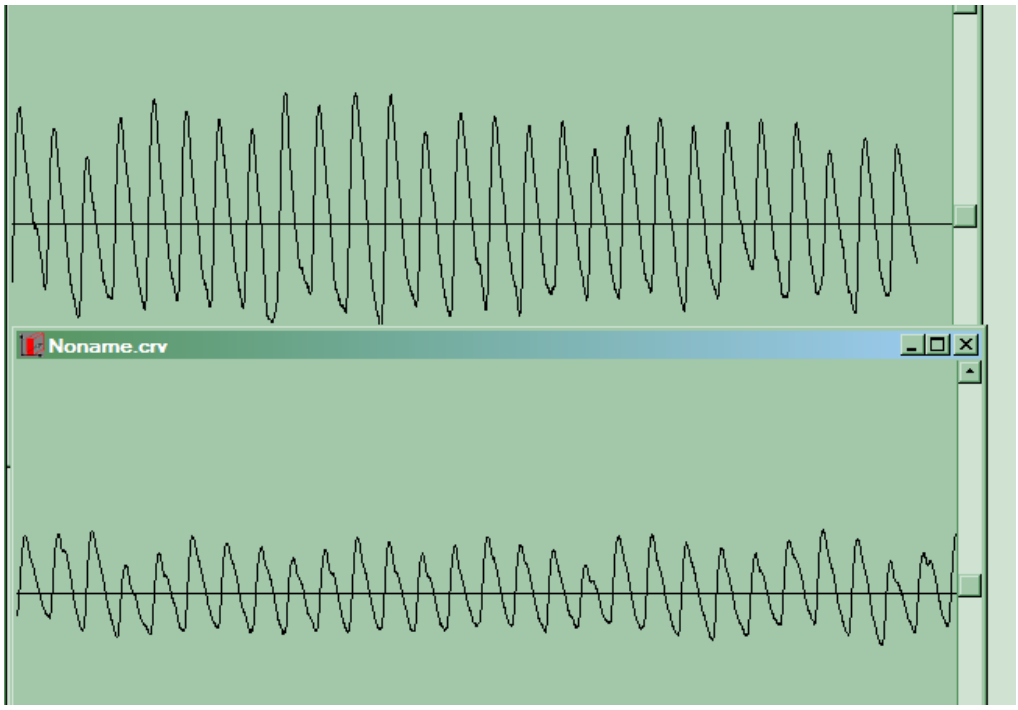

Рисунок 2.11 – Пальцевая фотоплетизмограмма до окклюзии плечевой артерии (фрагмент вверху) и после окклюзии (фрагмент внизу)

Информативные признаки X2 и X3 могут быть включены в пространство информативных признаков без модификации, то есть Y2=X2 и Y3=X3. Эти параметры характеризуют степень выраженности второй и третьей гармоник в сигнале. Они могут измеряться до окклюзии и после нее.

Спектральные характеристики фотоплетизмограмммы также можно измерять до окклюзии и после окклюзии. Но, в отличие от параметров интенсивности, для спектральных параметров окклюзия служит индикатором состояния функциональных систем, а не способом отстройки от помех. Поэтому информативные параметры, определяемые спектральными характеристиками сигнала, необходимо измерить до окклюзии и после окклюзии. Изменение этих параметров можно наблюдать на примере спектрограмм, показанных на рисунке 2.11.

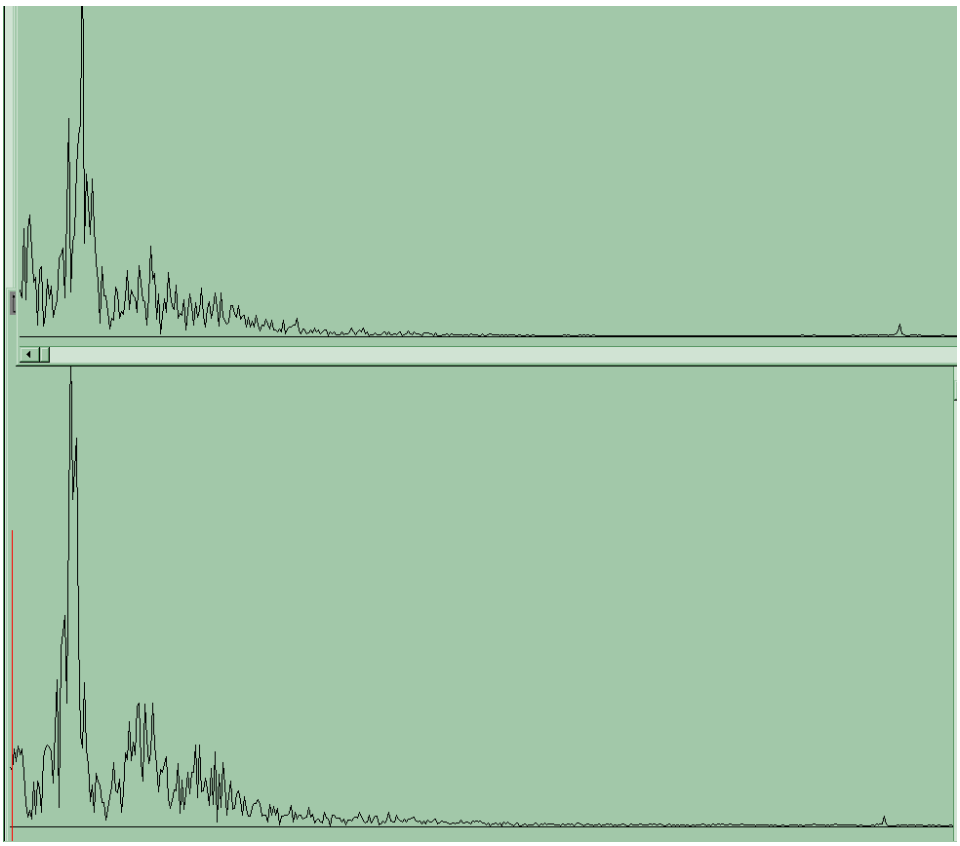

Рисунок 2.11 – Спектр Фурье фотоплетизмограммы до окклюзии (вверху) и после окклюзии (внизу)

Рисунок 2.12 показывает, что реальная картина спектра значительно сложнее, чем это представлено на рисунке 2.8. В связи с этим зафиксировать однозначно координаты ω_{i1} и ω_{i2} на уровне $0{,}8\left|F(\omega)\right|_{i\max}$ весьма проблематично. Можно изменить уровень на 0,5 или другое число, но это не приведет к однозначности этих координат.

Статистические исследования спектров фотоплетизмограмм пациентов с различными адаптационными резервами покали, что при частоте дискретизации фотоплетизмограммы 100 Гц ширина гармоник кардиоцикла не превышает величины десяти отсчетов или 0,333 Гц. Поэтому величины ω_{i1} для всех гармоник кардиоцикла могут быть привязаны к частотам Ω_i и определяться по следующим формулам:

$$\omega_{i1} = \Omega_i - 0{,}333/2, \qquad (2.27)$$

$$\omega_{i2} = \Omega_i - 0{,}333/2, \qquad (2.28)$$

Координаты амплитуды первой гармоники кардиоцикла Ω_1 определяем по величине $|F(\omega)|_{1max}$, как это показано стрелками на рисунке 2.8. Координаты же остальных гармоник вычисляем как $\Omega_2 = 2 \cdot \Omega_1$, $\Omega_3 = 3 \cdot \Omega_1$. Это сразу делает информативные признаки X7, X8 и X9 линейно зависимыми, поэтому вместо них вводим один признак $Y7 = \Omega_1$.

Четвертый, пятый и шестой информативные признаки определяют нормированную площадь под кривой спектрограммы в интервале ω_{i1} и ω_{i2} определяются по следующим формулам:

$$Y4 = \frac{1}{|F(\omega)|_{1max}} \sum_{i=\omega 11}^{\omega 12} |F(\omega_i)|, \qquad (2.29)$$

$$Y5 = \frac{1}{|F(\omega)|_{1max}} \sum_{i=\omega 21}^{\omega 22} |F(\omega_i)|, \qquad (2.30)$$

$$Y6 = \frac{1}{|F(\omega)|_{1max}} \sum_{i=\omega 31}^{\omega 32} |F(\omega)|, \qquad (2.31)$$

где ω_{i1} – нижняя граница полосы i-й гармоники кардиоцикла, выраженная в отсчетах; ω_{i2} – верхняя граница полосы i-й гармоники кардиоцикла, выраженная в отсчетах; $|F(\omega_i)|$ - i-й отсчет амплитудного спектра фотоплетизмограммы.

Полученные информативные признаки сведены в таблице 2.3.

Таблица 2.3 – Структура признакового пространства, полученного на основе анализа пальцевой фотоплетизмограммы

Комплексные информативные признаки	Информативные признаки, полученные до окклюзии						Информативные признаки, полученные после окклюзии					
Y1	2	3	4	5	6	7	8	9	10	11	12	13

Как видно из таблицы 2.3, информативные признаки делятся на три категории. К первой категории относим признаки, которые

получены по фрагментам фотоплетизмограммы, записанным перед и после окклюзии. В эту категорию попал всего лишь один признак, но их число при необходимости может быть увеличено как за счет признаков второй и третьей категорий, так и за счет ввода новых признаков. Этот признак определяется по формуле (2.25).

Ко второй категории относятся информативные признаки Y2…Y7, вычисляемые по фотоплетизмограмме, полученной перед окклюзией. Информативные признаки Y2…Y6 вычисляются по формулам (2.17), (2.18), (2.29). (2.30) и (2.31), соответственно, а информативный признак Y7 вычисляется согласно выражению:

$$Y7 = \Omega_1; \qquad (2.32)$$

где $|F(\Omega_1)| = \sup|F(\omega_i)|$, $\omega_i \in \{0...\Omega\}$; $\Omega = F\partial/2$; $F\partial$ - частота дискретизации фотоплетизмосигнала.

К третьей категории относятся информативные признаки Y8…Y13, вычисляемые по фотоплетизмограмме, полученной после окклюзии. Они вычисляются по тем же самым формулам, что и информативные признаки Y2…Y7.

Также отметим, что информативные признаки Y2, Y3 и Y8, Y9 характеризуют распределение энергии фотоплетизмосигнала между гармониками кардиоцикла, а информативные признаки Y4…Y6 и Y10…Y12 характеризуют энергию конкретной гармоники кардиоцикла. Информативные признаки Y7 и Y13 соответствуют ЧСС до и после окклюзии.

2.4. Задание на курсовую работу

Курсовая работа включает получение таблицы экспериментальных данных (ТЭД), строки которой идентичны таблице 2.3. Разница состоит только в том, что в реальной таблице добавляется еще один столбец, характеризующий класс функционального состояния, например, после нагрузки-до нагрузки. Число классов и методика ввода биообъект в соответствующее функциональное состояние является индивидуальной для каждого варианта задания. Функциональное состояние может быть изменено с помощью велоэргометра или прибора низкоинтенсивной лазерной терапии (по согласованию с

преподавателем). В эксперимент могут привлекаться студенты младших курсов.

В таблице должно быть не менее тридцати строк.

1. Заполнить пропуски в ТЭД
2. Удалить артефакты.
3. Найти основные статистические показатели данных
4. Используя корреляционный анализ определить статистически зависимые признаки, на основание чего сократить пространство признаков
5. Используя пакета дискриминантного анализа из Statistica 6 получить решающее правило для разделения классов
6. Оценить эффективность решающего правила. Если классификация неудовлетворительная, введите дополнительные информативные признаки, например, пол, гипотоник/гипертоник, САД, ДАД и т.п.

Примечание: при обработке данных можно пользоваться программой SPWIN

УРОВЕНЬ СЛОЖНОСТИ: «ПРОДВИНУТЫЙ»

3. Задание на курсовую работу

Вариант 1

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6 определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 2

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).

4. Используя пакет STATISTICA6 определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 3

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6 определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 4

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6 определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 5

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных ряда X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.

6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 6

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 7

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.

2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 8

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 9

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 10

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.

6.Оценить информативность признаков, при необходимости удалить неинформативные.

7. Выбрать тип решающего правила.

8. Найти численные значения параметров решающего правила.

9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 11

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.

2. По одномерным гистограммам оценить структуру классов.

3. Провести статистический анализ данных (найти основные статистические параметры рядов).

4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.

5. Провести визуализацию данных при помощи построения двумерных проекций.

6. Оценить информативность признаков, при необходимости удалить неинформативные.

7. Выбрать тип решающего правила.

8. Найти численные значения параметров решающего правила.

9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 12

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.

2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 13

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 14

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по пять вариационных рядов X1…X5 и X11…55. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с равномерым законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 15

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит пошесть вариационных рядов X1…X6 и X11…66. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с нормальным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.

6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

Вариант 16

1. Используя пакет MathCad сформировать две обучающих выборки для двух диагностируемых классов. В каждую выборку входит по четыре вариационных ряда X1…X4 и X11…44. Формирование вариационных рядов осуществляется посредством генератора случайных чисел с равномерным законом распределения.
2. По одномерным гистограммам оценить структуру классов.
3. Провести статистический анализ данных (найти основные статистические параметры рядов).
4. Используя пакет STATISTICA6, определить функции дискриминации и расстояния Махаланобиса.
5. Провести визуализацию данных при помощи построения двумерных проекций.
6. Оценить информативность признаков, при необходимости удалить неинформативные.
7. Выбрать тип решающего правила.
8. Найти численные значения параметров решающего правила.
9. Используя данные таблиц в качестве контрольной выборки найти значение вероятности правильной классификации полученной модели.

3.1. Пример выполнения курсовой работы

Задание:

1. Заполнить пропуски в ТЭД
2. Удалить артефакты.

3. Найти основные статистические показатели данных

3. По одномерным гистограммам оценить структуру классов.

4. Выбрать тип решающего правила.

5. Найти численные значения параметров решающего правила.

6. Используя данные ТЭД в качестве контрольной найти значение вероятности правильной классификации полученной модели.

3.1.1. Математическая методология.

Обычно в распоряжении исследователя имеются лишь данные выборки, например, значения количественного признака $x_1, x_2, \ldots x_n$, полученные в результате n наблюдений. Через эти данные и выражают оцениваемый параметр. При n > 50 для оценки математического ожидания и дисперсии следует пользоваться формулами 3.1 и 3.2 соответственно:

$$M[x] = \sum x_i / n, \qquad (3.1)$$

$$D[x] = \sum (x_i - M[x])^2 n, \qquad (3.2)$$

где n – длина выборки;

$$СКО = D[x]^{\frac{1}{2}} \qquad (3.3)$$

Медианой называется то значение, которое удовлетворяет условию:

$$P(x > M) = P(x < M), \qquad (3.4)$$

где M – медиана.

Модой называется то возможное значение, при которой плотность распределения максимальна.

Коэффициент вариации:

$$V = D[x]^{\frac{1}{2}} \cdot 100\% / M \qquad (3.5)$$

Коэффициент асимметрии:

$$As = (\sum (x_i - M)^3) / D[x]^{\frac{3}{2}}. \qquad (3.6)$$

Для статистического описания материала используются коэффициент корреляции, выборочный коэффициент отношения, которые вычисляются по формулам (3.3), (3.4) соответственно:

$$R_{xy} = \frac{\sum (x_i - M[x]) \cdot (y_i - M[y])}{(\sum (x_i - M[x])^2 \cdot \sum (y_i - M[y])^2)^{\frac{1}{2}}} \qquad (3.7)$$

$$\eta_{xy} = \delta_{yx} / \delta, \qquad (3.8)$$

где $\delta_{yx} = \sqrt{\sum n_x \cdot (y_{x\,ср} - y_{ср})^2 / n}$, $\delta_y = \sqrt{\sum n_y \cdot (y - y_{ср})^2 / n}$

Коэффициент корреляции принимает значения от –1 до+1.

С помощью получившихся коэффициентов корреляции можно составить корреляционную матрицу и построить графы связности признаков с учетом отброса статистически незначимых данных. Количество и толщина линий определяется рангом. Если ранг равен нулю, то связи между признаками нет, если единице, то средняя, если ранг равен двум, то связь сильная, если трем, то очень сильная.

Чувствительность:

$$Se = PS / S \cdot 100\%, \qquad (3.9)$$

где PS – количество больных с идентифицированным значением признака, S – общее число больных.

Специфичность:

$$Sp = NH / H \cdot 100\%, \qquad (3.10)$$

где NH – число здоровых, у которых отсутствует рассматриваемый признак, H – общее число здоровых людей.

$$Эфф = \frac{PS + R}{S + H} \cdot 100\%, \qquad (3.11)$$

где R – число здоровых, попавших в доверительный интервал для здоровых, H - общее число здоровых людей, S – общее число больных.

Для построения решающих правил находим доверительный интервал по формуле: $P_0 \pm \Delta P$, где P_0 - среднее, $\Delta P = \delta_p / \sqrt{n-1}$.

3.1.2. Результаты исследования

Исходные данные представлены в таблицах 3.1 и 3.2.

Таблица 3.1 – Исходные данные

№	X1	X2	X3	X4	X5
1	2	3	4	5	6
1	211	-112	876	6981	1297
2	205	-100	873	6997	1297
3	208	-97	874	6998	1306
4	198	-85	889	6993	1293
5	199	-111	868	6995	
6	212	-95	876	6992	1301
7	199	-107	882	6993	1297
8	204	-102	875	6990	1295
9	198	-104	882	7001	1315
10		-103	877	7009	1293
11	207	-100	881	7010	1287
12	204	-104	867	7005	1307
13	221	-94	885	6984	1292
14	208	-101	856	7000	1307
15	189	-92	891	6994	1292

Продолжение таблицы 3.1.

1	2	3	4	5	6
16	210	-105	898	7002	1293
17	25	-96	852	6998	1305
18	206	-92	877	6995	1315
19	193	-92	867	6995	1290
20	186	-97	883	6990	1309

Таблица 3.2 – Исходные данные

№	X11	X22	X33	X44	X55
1	168	-99	832	7017	1278
2	184	-103	910	7019	1335
3	169	-57	826	6996	1262
4	206	-68	857	6965	1296
5	261	-128	847	7025	1345
6	188	-69	889	7029	1308
7	156	-28	850	7046	1352
8	192	-107	925	7060	1266
9	241	-140	900	6963	1278
10	185	-90	857	7022	1336
11	139	-148	894	7023	1310
12	222	-63	837	6973	1255
13		-125	900	7034	1226
14	201	-118	921	7022	1355
15	222	-118	878	6975	1341
16	191	-99	836	6974	1336
17	239	-102	851	6975	1300
18	198	-139	847	7066	1239
19	179	-52	898	6945	1274
20	238	-115	849	7013	1320

3.1.3 Расчетная часть

Заполним пропуски. После заполнения получились таблицы 3.3 и 3.4.

Таблица 3.3 – Исходные данные с заполненными пропусками

№	X1	X2	X3	X4	X5
1	211	-112	876	6981	1297
2	205	-100	873	6997	1297
3	208	-97	874	6998	1306
4	198	-85	889	6993	1293
5	199	-111	868	6995	**1315**
6	212	-95	876	6992	1301
7	199	-107	882	6993	1297
8	204	-102	875	6990	1295
9	198	-104	882	7001	1315
10	**206**	-103	877	7009	1293
11	207	-100	881	7010	1287
12	204	-104	867	7005	1307
13	221	-94	885	6984	1292
14	208	-101	856	7000	1307
15	189	-92	891	6994	1292
16	210	-105	898	7002	1293
17	25	-96	852	6998	1305
18	206	-92	877	6995	1315
19	193	-92	867	6995	1290
20	186	-97	883	6990	1309

Таблица 3.4 – Исходные данные с заполненными пропусками

№	X11	X22	X33	X44	X55
1	2	3	4	5	6
1	168	-99	832	7017	1278
2	184	-103	910	7019	1335
3	169	-57	826	6996	1262
4	206	-68	857	6965	1296
5	261	-128	847	7025	1345
6	188	-69	889	7029	1308
7	156	-28	850	7046	1352
8	192	-107	925	7060	1266
9	241	-140	900	6963	1278
10	185	-90	857	7022	1336

Продолжение таблицы 4.3.

1	2	3	4	5	6
11	139	-148	894	7023	1310
12	222	-63	837	6973	1255
13	**222**	-125	900	7034	1226
14	201	-118	921	7022	1355
15	222	-118	878	6975	1341
16	191	-99	836	6974	1336
17	239	-102	851	6975	1300
18	198	-139	847	7066	1239
19	179	-52	898	6945	1274
20	238	-115	849	7013	1320

Удалим артефакты. Все значения каждого параметра должны находиться в диапазоне M±2σ. Если это не так, то заменяем это значение медианой соответствующего показателя.

Найдем статистические показатели, необходимые для этого (таблицы 3.5-3.6).

Таблица 3.5 – Статистические показатели

Показатель	X1	X2	X3	X4	X5
Мат. ожидания	198,5	-104,5	879,5	6985,5	1303
Дисперсия	1591,25	70,05	125,25	163,05	82,8
СКО	39,89048	8,369588	11,19151	12,7691	9,099451

Таблица 3.6 – Статистические показатели

Показатель	X11	X22	X33	X44	X55
Мат. ожидания	203	-107	840,5	7015	1299
Дисперсия	1591,25	70,05	125,25	163,05	82,8
СКО	39,89048	8,369588	11,19151	12,7691	9,099451

В данном случае все значения попадают в данный диапазон. Таким образом, артефактов нет.

Найдем основные статистические характеристики данных показателей (таблицы 3.7-3.8).

Таблица 3.7 – Основные статистические характеристики данных показателей

Показатель	X1	X2	X3	X4	X5
Медиана	206,5	-101,5	879	7009,5	1290
Мода					
Эксцесс	18,15675	-0,18185	0,574618	0,26748	-1,10268
Коэффициент асимметрии	1,56607	982744	13,83187	4575619	2936,744
Коэффициент вариации	20,09596	-8,00917	1,272486	0,182794	0,698346
Минимум	25	-112	852	6981	1287
Максимум	221	-85	889	7010	1315

Таблица 3.8 – Основные статистические характеристики данных показателей

Показатель	X11	X22	X33	X44	X55
Медиана	162	-119	875,5	7022,5	1323
Мода					
Эксцесс	-0,43571	-0,51893	-1,34469	-0,93136	-1,06646
Коэффициент асимметрии	1,56607	982744	13,83187	4575619	2936,744
Коэффициент вариации	19,65048	-7,82204	1,331531	0,182026	0,700497
Минимум	139	-148	826	6945	1226
Максимум	261	-28	925	7066	1352

Если коэффициент As<0, то это левосторонняя асимметрия, если As>0, то правосторонняя. Коэффициент вариации служит для сравнения величин рассеяния по отношению к выборочной средней двух вариационных рядов.

Чем больше коэффициент, тем больше рассеяние.

Построим одномерные гистограммы для таблиц 3.1 и 3.2 соответственно, шаг в которых рассчитаем по формуле:

$$\text{Шаг} = (\text{максимальное значение} - \text{минимальное значение})/(1+\log_2 \text{N})$$

где в данном случае N-20.

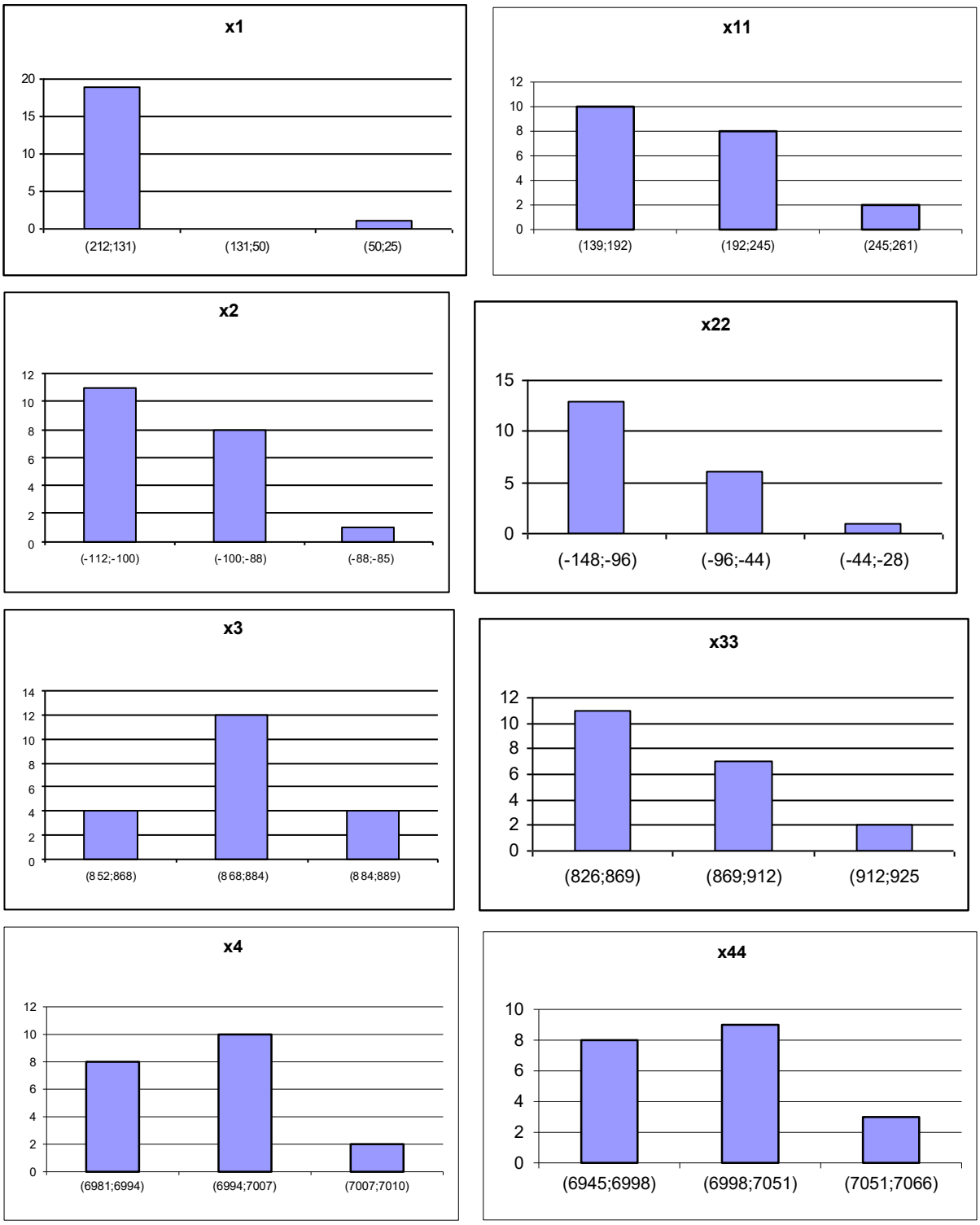

Рисунок 3.1 – Одномерные гистограммы для таблиц 3.1 и 3.2

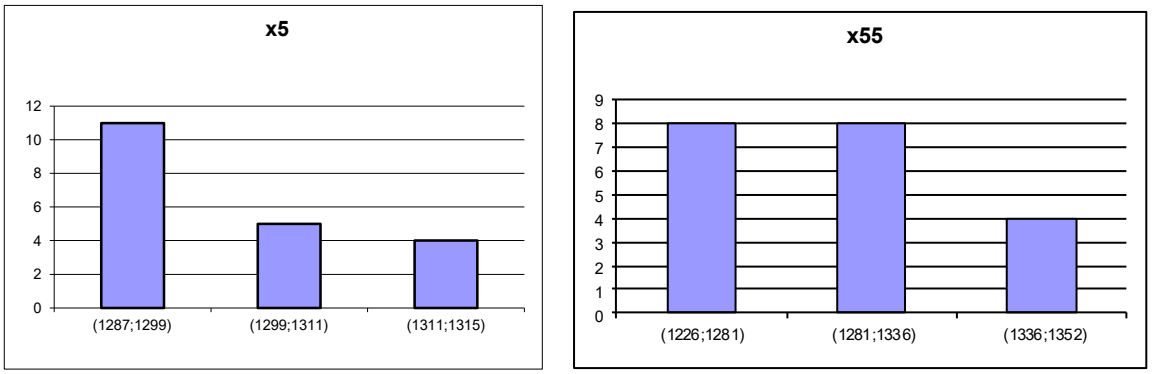

Рисунок 3.2 – Одномерные гистограммы для таблиц 3.1 и 3.2

Составим корреляционную матрицу. Для этого воспользуемся формулой 3.7.

Корреляционная матрица для таблицы 3.1 представлена в виде таблице 3.9.

Таблица 3.9 – Корреляционная матрица для таблицы 3.1

№	1	2	3	4	5
1	1	-0,1	0,5	-0,1	-0,1
2	-0,1	1	-0,01	0,46	-0,3
3	0,5	-0,01	1	-0,29	-0,26
4	-0,1	0,46	-0,29	1	-0,24
5	-0,1	-0,3	-0,26	-0,24	1

Корреляционная матрица для таблицы 3.3 представлена в виде таблицы 3.10.

№	11	22	33	44	55
11	1	-0,36	-0,1	-0,2	-0,03
22	-0,36	1	-0,03	-0,28	0,05
33	-0,1	-0,03	1	-0,05	0,05
44	-0,2	-0,28	-0,05	1	-0,02
55	-0,032	0,05	0,05	-0,02	1

Определим значимость каждого коэффициента корреляции и для каждого значимого коэффициента определим соответствующие регрессионные модели.

Если $\text{arctgR} > t(p)/\sqrt{n-3}$, где $t(p) = 1.67$, n – длина выборки, то коэффициент значим.

Итак, в нашем случае значимыми являются коэффициенты:

R11 = R22 = R33 = R44 = R55 = 1

R1111 = R2222 = R3333 = R4444 = R5555 = 1

R14 = R41 = 0.5

R25 = R52 = -0.3

Зависимость показателей для таблицы 3.1 и 3.2 показана на рисунках 3.3-3.4.

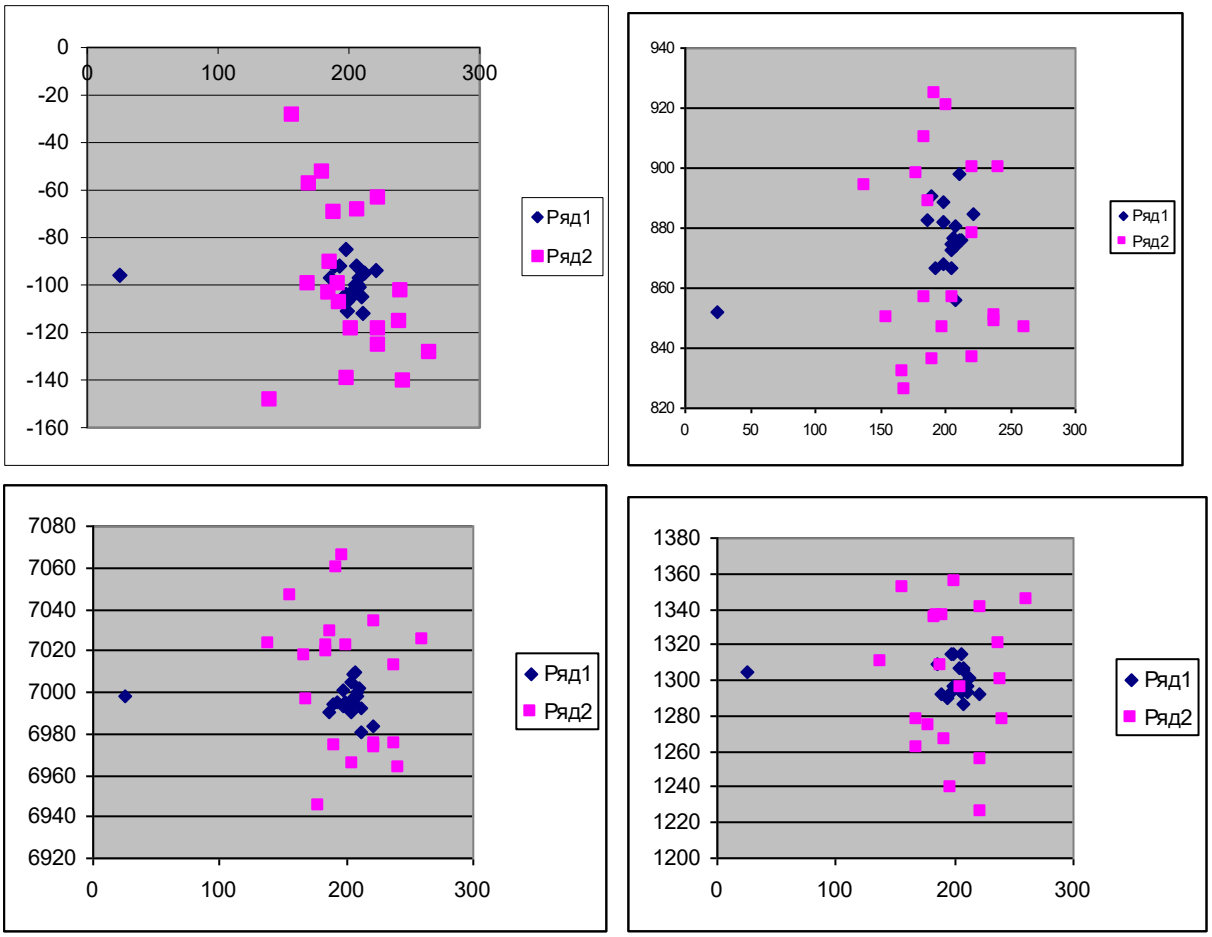

Рисунок 3.3 – Зависимость показателей для таблиц 3.1 и 3.2

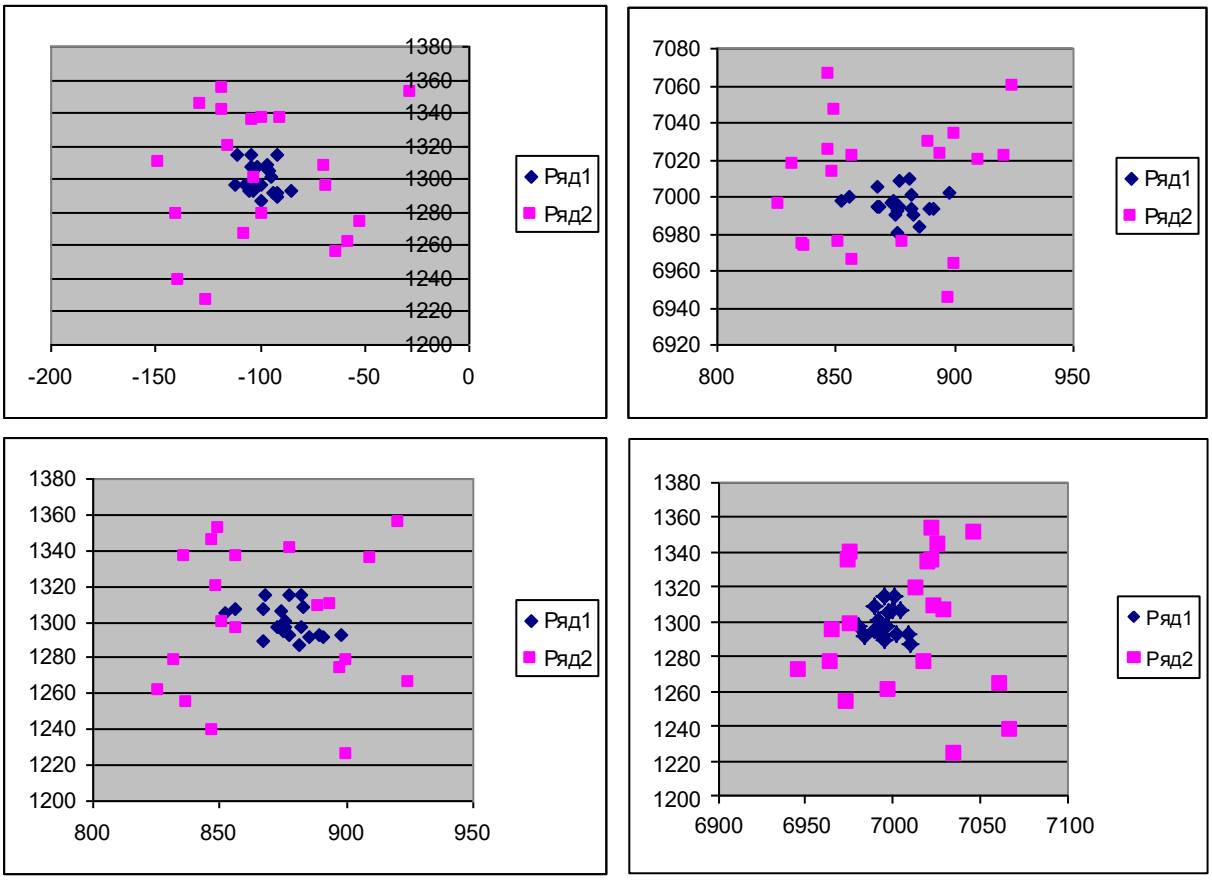

Рисунок 3.4 – Зависимость показателей для таблиц 3.1 и 3.2

По полученным данным найдем тип решающего правила.

Построим линейную разделяющую поверхность для двух классов.

Для этого необходимо найти координаты точки C, которая характеризует среднюю составляющую по формуле:

$$C((M[x]1 + M[y]1)/2;\ (M[x]2 + M[y]2)/2;\ \ldots)$$

$$C(-25{,}25;\ 538{,}75;\ 2075;\ 3660{,}75;\ 175{,}75)$$

Теперь необходимо найти координаты вектора AB, которые соединяют точки X и Y.

$$AB(-11{,}5;\ -1242{,}5;\ 51{,}5;\ 4201{,}5;\ -30{,}5)$$

Найдем уравнение плоскости, которая проходит через точку C и имеет нормаль AB. Если принять, что AB(A, B,C,D) и C(Xo,Yo, Zo, Eo), тогда плоскость будет иметь уравнение:

$$-11{,}5\cdot(X1+25{,}5)-1242{,}5\cdot(X2-538{,}75)+51{,}5\cdot(X3-20{,}75)+$$
$$+4201{,}5\cdot(X4-3660{,}75)-30{,}5\cdot(X5-175{,}75)=0$$
$$-11{,}5\cdot X1-1242{,}5\cdot X2+51{,}5\cdot X3+4201{,}5\cdot X4-30{,}5\cdot X5-18367996=Y$$

Найдем значения Y для строк каждого класса и найдем минимум и максимум для них (таблицы 3.11-3.12)

Таблица 3.11 – Нахождение значения Y для строк каждого класса, минимума и максимума.

№	X1	X2	X3	X4	X5	y
1	211	-112	876	6981	1297	
2	205	-100	873	6997	1297	
3	208	-97	874	6998	1306	
4	198	-85	889	6993	1293	
5	199	-111	868	6995	**1315**	
6	212	-95	876	6992	1301	
7	199	-107	882	6993	1297	
8	204	-102	875	6990	1295	
9	198	-104	882	7001	1315	
10	**206**	-103	877	7009	1293	
11	207	-100	881	7010	1287	
12	204	-104	867	7005	1307	
13	221	-94	885	6984	1292	
14	208	-101	856	7000	1307	
15	189	-92	891	6994	1292	
16	210	-105	898	7002	1293	
17	25	-96	852	6998	1305	
18	206	-92	877	6995	1315	
19	193	-92	867	6995	1290	
20	186	-97	883	6990	1309	
max						
min						

Таблица 3.12 – Нахождение значения Y для строк каждого класса, минимума и максимума.

№	X11	X22	X33	X44	X55	y
1	168	-99	832	7017	1278	
2	184	-103	910	7019	1335	
3	169	-57	826	6996	1262	
4	206	-68	857	6965	1296	
5	261	-128	847	7025	1345	
6	188	-69	889	7029	1308	
7	156	-28	850	7046	1352	
8	192	-107	925	7060	1266	
9	241	-140	900	6963	1278	
10	185	-90	857	7022	1336	
11	139	-148	894	7023	1310	
12	222	-63	837	6973	1255	
13	**222**	-125	900	7034	1226	
14	201	-118	921	7022	1355	
15	222	-118	878	6975	1341	
16	191	-99	836	6974	1336	
17	239	-102	851	6975	1300	
18	198	-139	847	7066	1239	
19	179	-52	898	6945	1274	
20	238	-115	849	7013	1320	
max						
min						

УРОВЕНЬ СЛОЖНОСТИ: «ПОРОГОВЫЙ»

4. Исследование эффективности классификации двухальтернативной выборки геометрическими методами распознавания

4.1. Этапы выполнения курсовой работы

В процессе выполнения курсовой работы необходимо сформировать две обучающих выборки. Каждая выборка представляет определенный класс индивидуумов, характеризующихся набором информативных признаков. Число объектов в обучающих выборках и число информативных признаком определяется индивидуальным заданием.

Затем проводится разведочный анализ, который заключается в исследовании гистограмм информативных признаков в многомерных пространствах и их сравнения (по классам).

Одномерные гистограммы по каждому признаку строятся для двух обучающих выборок, при этом шаг рассчитаем по формуле:

$$\text{Шаг} = (\text{максимальное значение} - \text{минимальное значение})/(1 + \log_2 \text{N})$$

где N – объем обучающей выборки.

В результате этого исследования исключаются те признаки, гистограммы которого имеют площадь перекрытия выше 80%.

Затем необходимо провести корреляционный анализ, устанавливающий линейную зависимость признаков, который должен подтвердить или не подтвердить верность исключения выбранного признака.

После этого исследуется три способа классификации данных, основанных на геометрических методах распознавания;

1 – построение разделяющей гиперплоскости;

2 – определение расстояний Махаланобиса;

3 – использование пакета дискриминантного анализа из Statistica 6.

В итоге формируются три решающих правила разделение совокупностей.

На следующем этапе формируются две контрольных выборки и проверяется эффективность решающих правил.

Обучающие и контрольные выборки формируются студентом самостоятельно на основании их статистических характеристик, приведенных в индивидуальном задании.

В заключении сравнивается эффективность решающих правил, разработанных на основе трех методов. Для наглядности сравнения решающих правил необходимо построить гистограммы, отражающие диагностическую чувствительность, диагностическую специфичность и диагностическую эффективность каждого решающего правила.

4.2. Индивидуальные задания

Исходные данные для первой совокупности представлены в таблице 4.1.

Таблица 4.1 – Исходные данные для первой совокупности

№	Nобуч/ Nконтр ольной	Математические ожидания информативных признаков						Дисперсии информативных признаков					
1	30/20	-18	112	101	495	10	51	100	500	100	1500	10	80
2	40/35	-21	109	140	491	17	49	81	145	39	1600	9	81
3	35/40	-21	100	132	513	20	-49	144	98	45	1460	9	144
4	15/30	20	95	120	499	16	-40	64	60	78	1000	18	64
5	18/25	26	73	66	509	17	-46	49	100	87	1000	20	49
6	25/19	-19	91	107	122	17	66	100	98	90	140	25	100
7	34/20	-15	68	93	122	23	-43	36	45	65	90	25	36
8	41/35	-19	72	78	127	13	-49	49	40	66	60	16	49
9	36/42	16	148	72	123	87	42	9	36	78	100	100	9
10	18/30	17	70	118	128	86	-60	16	26	76	98	100	16
11	19/25	-18	83	97	115	82	52	25	28	78	45	64	25
12	28/19	-21	107	141	124	80	-58	25	56	98	40	64	25
13	30/23	-22	112	97	127	70	55	49	80	100	36	49	49

Исходные данные для второй совокупности представлены в таблице 4.2.

Таблица 4.2. – Исходные данные для первой совокупности

№	Nобуч/ Nконтрольной	Математические ожидания информативных признаков						Дисперсии информативных признаков					
1	20/40	-8	120	90	490	15	61	10	100	100	400	10	36
2	40/35	-25	209	150	441	27	69	8	145	39	500	9	49
3	42/40	-28	110	80	413	27	-29	14	98	45	300	9	36
4	19/30	30	85	128	409	19	-30	6	60	78	100	18	9
5	18/25	26	93	76	559	27	-56	4	100	87	100	20	49
6	35/19	-9	71	88	122	27	86	10	98	90	100	25	49
7	24/20	-5	70	89	132	33	-33	8	45	65	80	25	36
8	31/35	-9	92	60	147	23	-69	4	40	66	9	16	49
9	36/42	20	120	87	133	97	52	9	36	78	10	10	36
10	48/30	20	80	81	158	116	-50	16	26	76	9	10	16
11	19/25	-19	78	100	125	102	82	25	28	78	4	4	25
12	28/19	-29	120	131	134	89	-78	25	56	98	4	4	25
13	35/25	-28	140	87	147	80	59	4	80	100	6	9	49

4.3. Получение контрольных и обучающих выборок

Используя генератор случайных чисел пакета MathCad можно создать как обучающие, так и контрольные выборки с любым объемом и любой размерности пространства информативных признаков. Пример листа MathCad, в котором создается выборка из 28 объектов с нормально распределенным законом распределения признаков (признаковое пространство четырехмерное) показан на рисунок 4.1.

$n := 28 \qquad m := 4 \qquad i := 0..m-1 \qquad j := 0..n-1$

$s_0 := 225 \qquad sd_0 := 16 \qquad s_1 := 228 \qquad sd_1 := 18 \qquad s_2 := 231$

$sd_2 := 15 \qquad s_3 := 235 \qquad sd_3 := 16$

$V\lambda_i := \overrightarrow{ceil\left(rnorm\left(n, s_i, sd_i\right)\right)}$

Рисунок 4.1 – Лист MathCad с генератором случайных чисел

$V\lambda_0$ =

	0
0	218
1	215
2	218
3	210
4	199
5	226
6	224
7	234
8	261
9	238
10	241
11	239
12	240
13	236
14	209
15	227
16	213
17	237
18	223
19	215
20	214
21	217
22	234
23	222
24	227
25	246
26	214
27	226

$V\lambda_1$ =

	0
0	248
1	245
2	176
3	190
4	232
5	217
6	207
7	230
8	242
9	234
10	229
11	215
12	221
13	216
14	226
15	247
16	232
17	238
18	242
19	228
20	234
21	231
22	236
23	220
24	216
25	283
26	223
27	205

$V\lambda_2$ =

	0
0	243
1	226
2	213
3	216
4	231
5	232
6	235
7	239
8	235
9	221
10	207
11	253
12	246
13	214
14	234
15	257
16	227
17	201
18	210
19	212
20	224
21	232
22	223
23	227
24	240
25	235
26	238
27	214

$V\lambda_3$ =

	0
0	218
1	218
2	229
3	214
4	190
5	256
6	248
7	238
8	219
9	221
10	221
11	240
12	229
13	256
14	213
15	242
16	243
17	264
18	225
19	244
20	260
21	206
22	256
23	219
24	243
25	219
26	216
27	242

Рисунок 4.2 – Пример выборки из 28 элементов в четырехмерном признаковом пространстве.

Каждый информативный признак в выборке представлен вектором-столбцом Vλ из 28 элементов. Случайные числа, распределенные по нормальному закону распределения, генерируются функцией rnorm(n, s, sd), где n-число элементов в выборке, s – математическое ожидание информативного признака, sd -дисперсия информативного признака.

Варьируя эти параметры (s и sd) вы можете менять структуру распределения классов в признаковом пространстве и исследовать

эффективность классификации при различных классовых структурах.

Функция ceil округляет случайное число до ближайшего целого.

Значок → обозначает операцию векторизации, то есть одновременное выполнение скалярной операции над всеми элементами вектора.

Пример полученной выборки в четырехмерном признаковом пространстве показан на рисунке 4.2.

4.4. Порядок построения линейной разделяющей гиперплоскости

По полученным обучающим выборкам найдем разделяющую гиперплоскость, которая проходит через середину отрезка, соединяющего центроиды двух обучающих выборок и перпендикулярна к нему.

Построим линейную разделяющую поверхность для двух классов.

Если вектор X_i характеризует i-й объект первого класса, а вектор Y_j характеризует j–й объект второго класса, то координаты центроид A и B , а, следовательно, и отрезка AB, их соединяющего, определяются как M[X] и M[Y].

Координаты точки C, которая лежит на середине отрезка AB, определяются по формуле:

$$C((M[x_1]+M[y_1])/2;(M[x_2]+M[y_2])/2;...;(M[x_N]+M[y_N])/2),$$

где N –число информативных признаков или размерность признакового пространства.

Если плоскость перпендикулярна вектору n(a; b; c), например, в трехмерном пространстве, то уравнение плоскости в этом пространстве записывается как:

$$ax+by+cz+d=0.$$

Уравнение плоскости, перпендикулярной вектору n(a; b; c) и проходящей через точку (x0; y0; z0) записывается как:

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

Чтобы перейти от вектора AB к вектору **n**, необходимо из координат $M[x_i]$ вычесть координаты $M[y_i]$, то есть из координат точки A вычесть координаты точки B.

4.5. Дискриминантный анализ

Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы). Например, некий исследователь в области образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы. После выпуска большинство учащихся, естественно, должно попасть в одну из названных категорий. Затем можно использовать дискриминантный анализ для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути.

Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, выздоровел полностью (группа 1), частично (группа 2) или совсем не выздоровел (группа 3). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

Классификация:

Другой главной целью применения дискриминантного анализа является проведение классификации. Как только модель установлена и получены дискриминирующие функции, возникает

вопрос о том, как хорошо они могут предсказывать, к какой совокупности принадлежит конкретный образец?

1)Априорная и апостериорная классификация. Прежде чем приступить к изучению деталей различных процедур оценивания, важно уяснить, что эта разница ясна. Обычно, если вы оцениваете на основании некоторого множества данных дискриминирующую функцию, наилучшим образом разделяющую совокупности, и затем используете те же самые данные для оценивания того, какова точность вашей процедуры, то вы во многом полагаетесь на волю случая. В общем случае, получают, конечно, худшую классификацию для образцов, не использованных для оценки дискриминантной функции. Другими словами, классификация действует лучшим образом для выборки, по которой была проведена оценка дискриминирующей функции (апостериорная классификация), чем для свежей выборки (априорная классификация). Трудности с (априорной) классификацией будущих образцов заключается в том, что никто не знает, что может случиться. Намного легче классифицировать уже имеющиеся образцы. Поэтому оценивание качества процедуры классификации никогда не производят по той же самой выборке, по которой была оценена дискриминирующая функция. Если желают использовать процедуру для классификации будущих образцов, то ее следует "испытать" (произвести кросс-проверку) на новых объектах.

2)Функции классификации. Функции классификации не следует путать с дискриминирующими функциями. Функции классификации предназначены для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект. Имеется столько же функций классификации, сколько групп. Каждая функция позволяет вам для каждого образца и для каждой совокупности вычислить веса классификации по формуле:

$$S_i = c_i + w_{i1} \cdot x_1 + w_{i2} \cdot x_2 + ... + w_{im} \cdot x_m$$

В этой формуле индекс i обозначает соответствующую совокупность, а индексы $1, 2, \ldots, m$ обозначают m переменных; c_i

являются константами для i-ой совокупности, w_{ij} - веса для j-ой переменной при вычислении показателя классификации для i-ой совокупности; x_j - наблюдаемое значение для соответствующего образца j-ой переменной. Величина S_i является результатом показателя классификации.

Поэтому вы можете использовать функции классификации для прямого вычисления показателя классификации для некоторых новых значений.

3)Классификация наблюдений. Как только вы вычислили показатели классификации для наблюдений, легко решить, как производить классификацию наблюдений. В общем случае наблюдение считается принадлежащим той совокупности, для которой получен наивысший показатель классификации (кроме случая, когда вероятности априорной классификации становятся слишком малыми). Поэтому, если вы изучаете выбор карьеры или образования учащимися средней школы после выпуска (поступление в колледж, в профессиональную школу или получение работы) на основе нескольких переменных, полученных за год до выпуска, то можете использовать функции классификации, чтобы предсказать, что наиболее вероятно будет делать каждый учащийся после выпуска. Однако вы хотели бы определить вероятность, с которой учащийся сделает предсказанный выбор. Эти вероятности называются апостериорными, и их также можно вычислить. Однако для понимания, как эти вероятности вычисляются, вначале рассмотрим так называемое расстояние Махаланобиса.

4.5.1. Расстояние Махаланобиса.

В общем, расстояние Махаланобиса является мерой расстояния между двумя точками в пространстве, определяемым двумя или более коррелированными переменными. Например, если имеются всего две некоррелированных переменные, то вы можете нанести точки (образцы) на стандартную 2М диаграмму рассеяния. Расстояние Махаланобиса между точками будет в этом случае равно расстоянию Евклида, т.е. расстоянию, измеренному,

например, рулеткой. Если имеются три некоррелированные переменные, то для определения расстояния вы можете по-прежнему использовать рулетку (на 3М диаграмме). При наличии более трех переменных вы не можете более представить расстояние на диаграмме. Также и в случае, когда переменные коррелированы, то оси на графике могут рассматриваться как неортогональные (они уже не направлены под прямыми углами друг к другу). В этом случае простое определение расстояния Евклида не подходит, в то время как расстояние Махаланобиса является адекватно определенным в случае наличия корреляций.

4.5.2. Расстояние Махаланобиса и классификация.

Для каждой совокупности в выборке вы можете определить положение точки, представляющей средние для всех переменных в многомерном пространстве, определенном переменными рассматриваемой модели. Эти точки называются *центроидами* группы. Для каждого наблюдения вы можете затем вычислить его расстояние Махаланобиса от каждого центроида группы. Снова, вы признаете наблюдение принадлежащим к той группе, к которой он ближе, т.е. когда расстояние Махаланобиса до нее минимально.

1) Апостериорные вероятности классификации. Используя для классификации расстояние Махаланобиса, вы можете теперь получить вероятность того, что образец принадлежит к конкретной совокупности. Это значение будет не вполне точным, так как распределение вокруг среднего для каждой совокупности будет не в точности нормальным. Так как принадлежность каждого образца вычисляется по априорному знанию модельных переменных, эти вероятности называются апостериорными вероятностями. Короче, апостериорные вероятности - это вероятности, вычисленные с использованием знания значений других переменных для образцов из частной совокупности. Некоторые пакеты автоматически вычисляют эти вероятности для всех наблюдений (или для выбранных наблюдений при проведении кросс-проверки).

2) Априорные вероятности классификации. Имеется одно дополнительное обстоятельство, которое следует рассмотреть при классификации образцов. Иногда вы знаете заранее, что в одной из

групп имеется больше наблюдений, чем в другой. Поэтому априорные вероятности того, что образец принадлежит такой группе, выше. Например, если вы знаете заранее, что 60% выпускников вашей средней школы обычно идут в колледж, (20% идут в профессиональные школы и остальные 20% идут работать), то вы можете уточнить предсказание таким образом: при всех других равных условиях более вероятно, что учащийся поступит в колледж, чем сделает два других выбора. Вы можете установить различные априорные вероятности, которые будут затем использоваться для уточнения результатов классификации наблюдений (и для вычисления апостериорных вероятностей).

На практике, исследователю необходимо задать себе вопрос, является ли неодинаковое число наблюдений в различных совокупностях в первоначальной выборке отражением истинного распределения в популяции, или это только (случайный) результат процедуры выбора. В первом случае вы должны положить априорные вероятности пропорциональными объемам совокупностей в выборке; во втором - положить априорные вероятности одинаковыми для каждой совокупности. Спецификация различных априорных вероятностей может сильно влиять на точность классификации.

Итог классификации. Общим результатом, на который следует обратить внимание при оценке качества текущей функции классификации, является матрица классификации. Матрица классификации содержит число образцов, корректно классифицированных (на диагонали матрицы) и тех, которые попали не в свои совокупности (группы).

Другие предостережения. При повторной итерации апостериорная классификация того, что случилось в прошлом, не очень трудна. Нетрудно получить очень хорошую классификацию тех образцов, по которым была оценена функция классификации. Для получения сведений, насколько хорошо работает процедура классификации на самом деле, следует классифицировать (априорно) различные наблюдения, то есть, наблюдения, которые не использовались при оценке функции классификации. Вы можете гибко использовать условия отбора для включения или исключения из вычисления наблюдений, поэтому матрица классификации

может быть вычислена по "старым" образцам столь же успешно, как и по "новым". Только классификация новых наблюдений позволяет определить качество функции классификации (см. также кросс-проверку); классификация старых наблюдений позволяет лишь провести успешную диагностику наличия выбросов или области, где функция классификации кажется менее адекватной.

В общем, Дискриминантный анализ - это очень полезный инструмент для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп, (2) - для классификации наблюдений в различные группы.

4.6. Оценка эффективности методов распознавания

В качестве расчетных показателей качества диагностических решающих правил используется: диагностическая чувствительность (ДЧ), диагностическая специфичность (ДС), прогностическая значимость положительных результатов ($ПЗ^{+}$), прогностическая значимость отрицательных результатов ($ПЗ^{-}$), диагностическая эффективность решающего правила (ДЭ).

Эти показатели вычислялись по данным таблицы распределений результатов контрольных испытаний (таблица 4.3).

Таблица 4.3 – Таблица контрольных испытаний

Обследуемые	Результаты срабатывания правил		Всего
	положительные	отрицательные	
n_{ω_r}	ИП	ЛО	ИП+ЛО
n_{ω_0}	ЛП	ИО	ЛП+ИО
Всего	ИП+ЛП	ЛО+ИО	ИП+ЛП+ЛО+ИО

где r – номер класса исследуемого заболевания; n_{ω_r} - количество людей в контрольной выборке в исследуемом классе заболеваний; n_{ω_0} - количество здоровых людей в контрольной выборке; ИП – истинно положительный результат равный количеству людей класса ω_r правильно классифицируемых рассматриваемым

правилом; ЛП – ложно положительный результат равный количеству людей класса ω_0 ошибочно отнесенных решающим правилом к классу $\omega_г$; ЛО – ложно отрицательный результат: количество людей класса $\omega_г$ ошибочно отнесенных решающим правилом к классу ω_0; ИО – истинно отрицательный результат: количество людей класса ω_0 правильно классифицируемых решающим правилом.

Для приведенных в таблице 4.3 обозначений расчет показателей качества осуществляется в соответствии с выражениями:

$$\begin{cases} ДЧ = ИП / n_{\omega_г} \\ ДС = ИО / n_{\omega_0} \\ ПЗ^{+} = ИП /(ИП + ЛП) \\ ПЗ^{-} = ИО /(ЛО + ИО) \\ ДЭ = (ИП + ИО) /(ИП + ЛП + ЛО + ИО) \end{cases}$$

Список литературы

1. Боровиков, В. STISTICA. Искусство анализа данных на компьютере: Для профессионалов [Текст] / В. Боровиков. 2-е изд. (+CD). СПб.: Питер, 2003. 688 с.
2. Горелик, А.Л. Методы распознавания: Учеб. пособие для вузов [Текст] / А.Л. Горелик, В.А. Скрипкин. М.: Высшая школа, 2004. 261 с.
3. Омельченко, В.П. Практикум по медицинской информатике [Текст] : серия «Учебники, учебные пособия» / В.П. Омельченко, А.Л. Демидова. Ростов-на-Дону: Феникс, 2001. 304 с.