**НИТУ «МИСиС»**

Квалификация (степень): **магистр**

Курс: **1**

Семестр: **1**

Дисциплина: Нейронные сети

**ОТЧЕТ**

**ПО ЛАБОРАТОРНОЙ РАБОТЕ**

на тему:

**«Метод главных компонент»**

**Выполнил:**

Группа: МПИ-20-4-2

Студент: Новицкий Д. А.

**Проверил:** доц., к.т.н. Курочкин И. И.

Оценка:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Москва 2020**

**Оглавление**

[Постановка задачи 3](#_Toc61029538)

[Ход выполнения лабораторной работы 3](#_Toc61029539)

[Результаты 3](#_Toc61029540)

[Визуализация 3](#_Toc61029541)

[Демонстрация работоспособности лабораторной работы 4](#_Toc61029542)

[Входные/выходные данные 4](#_Toc61029543)

[Загрузка файлов 4](#_Toc61029544)

[Описание работы программы 5](#_Toc61029545)

[Используемые датасеты 5](#_Toc61029546)

[Anuran Calls Dataset 5](#_Toc61029547)

[Diabetes Dataset 6](#_Toc61029548)

[Detecting Malware Using A Hybrid Approach Datasets 7](#_Toc61029549)

[Этапы работы программы 7](#_Toc61029550)

[Результаты работы программы 9](#_Toc61029551)

[Список используемой литературы 14](#_Toc61029552)

# Постановка задачи

## Ход выполнения лабораторной работы

1. Предобработка данных.

2. Построение матрицы корреляций и составление списка признаков со слабой корреляцией.

3. Подготовка данных для метода главных компонент на основе составленного списка.

4. Реализация метода главных компонент.

5. Реализация критериев выбора числа главных компонент (Кайзера, сломанной трости, каменистой осыпи).

6. Определение числа главных компонент.

7. Построение матрицы факторной нагрузки для всех значимых главных компонент.

8. Определение суммарной нагрузки для признаков. Составление ранжированного списка признаков (по убыванию нагрузки).

9. Анализ остатков.

## Результаты

1. Матрица корреляций.

2. Список признаков.

3. Количество отобранных главных компонент по 3 критериям.

4. Матрица факторной нагрузки для всех значимых главных компонент.

5. Ранжированный список признаков с суммарной нагрузкой.

6. Результаты анализа остатков.

## Визуализация

1. Матрица корреляций.

2. График % описываемой дисперсии в зависимости от количества главных компонент. (пометить на графике числа ГК по разным критериям).

3. Матрица факторной нагрузки.

4. Ранжированный список признаков с суммарной нагрузкой.

## Демонстрация работоспособности лабораторной работы

1. Протестировать методику обработки данных (см. ход выполнения ЛР) на 3 датасетах с размерностью ~20, ~60, ~1000 признаков (хотя бы один датасет должен быть уникальным).

2. Прокомментировать сокращение числа признаков после отбора списка признаков со слабой корреляцией.

3. Продемонстрировать различное количество главных компонент по различным критериям.

4. Продемонстрировать "хороший" и "плохой" случаи применения МГК. Пояснить, что Вы считаете "плохим", а что "хорошим" случаем работы МГК.

5. Прокомментировать результаты анализа остатков.

## Входные/выходные данные

1. Входные данные (датасеты) в виде текстового файла. (к примеру, экспорт таблицы из Excel в формате TXT или CSV).

2. Выходные данные (результаты) сохраняются в отчете (текст, таблицы и графики).

## Загрузка файлов

1. Работающее приложение.

2. Датасеты.

3. Текстовый файл со сценариями запуска (к примеру: myapp.exe dataset1 results).

4. Отчет по ЛР с приведенными результатами и подробными пояснениями (к примеру, если Вы получили какой-то график, приведите ниже сценарий и параметры запуска вычислительного приложения).

# Описание работы программы

## Используемые датасеты

В данной лабораторной работе используются различные наборы датасетов:

* Anuran Calls Dataset [1];
* Diabetes Dataset [2];
* Detecting Malware Using A Hybrid Approach Datasets [3].

Рассмотрим подробнее каждый из представленных датасетов.

### Anuran Calls Dataset

Всего в данном датасете присутствует:

* 7195 строк
* 22 столбца

Этот набор данных использовался в нескольких классификационных задачах, связанных с проблемой распознавания бесхвостых видов по их вызовам. Это набор данных с несколькими метками с тремя столбцами меток. Этот набор данных был создан с сегментацией 60 аудиозаписей, принадлежащих 4 различным семействам, 8 родам и 10 видам. Каждый звук соответствует одному экземпляру (отдельной лягушке), идентификатор записи также включен в качестве дополнительного столбца. Мы использовали спектральную энтропию и метод двоичного кластера для обнаружения аудиокадров, принадлежащих каждому слогу. Сегментация и извлечение признаков проводились в Matlab. После сегментации мы получили 7195 слогов, которые стали экземплярами для обучения и тестирования классификатора. Эти записи были получены на месте в реальных шумовых условиях (фоновый звук). Некоторые виды происходят из кампуса Федерального университета Амазонаса в Манаусе, другие - из Мата-Атлантика, Бразилия, а один из них - из Кордовы, Аргентина. Записи были сохранены в формате wav с частотой дискретизации 44,1 кГц и разрешением 32 бита, что позволяет нам анализировать сигналы до 22 кГц. Для каждого извлеченного слога было вычислено 22 MFCC с использованием 44 треугольных фильтров. Эти коэффициенты были нормализованы между [-1; 1]. Количество экземпляров в классе:

Семьи:

* Bufonidae: 68
* Dendrobatidae: 542
* Hylidae: 2165
* Leptodactylidae: 4420

Род:

* Adenomera: 4150
* Ameerega: 542
* Dendropsophus: 310
* Hypsiboas: 1593
* Leptodactylus: 270
* Osteocephalus: 114
* Rhinella: 68
* Scinax: 148

Виды:

* AdenomeraAndre: 672
* AdenomeraHylaedactâ: 3478
* Ameeregatrivittata: 542
* HylaMinuta: 310
* HypsiboasCinerascens: 472
* HypsiboasCordobae: 1121
* LeptodactylusFuscus: 270
* OsteocephalusOophaâ: 114
* Rhinellagranulosa: 68
* ScinaxRuber: 148

### Diabetes Dataset

Всего в данном датасете присутствует:

* 100 000 строк
* 55 столбцов

Набор данных представляет 10 лет (1999-2008 гг.) Клинической помощи в 130 больницах США и интегрированных сетях оказания помощи. Он включает более 50 функций, представляющих результаты для пациентов и больниц. Информация была извлечена из базы данных для встреч, которые удовлетворяли следующим критериям.

* Это стационарный прием (госпитализация).
* Это встреча с диабетом, то есть такая, во время которой в систему в качестве диагноза был внесен любой вид диабета.
* Продолжительность пребывания составляла от 1 до 14 дней.
* Во время встречи были проведены лабораторные испытания.
* Во время встречи вводили лекарства.

Данные содержат такие атрибуты, как номер пациента, раса, пол, возраст, тип госпитализации, время в больнице, медицинская специальность принимающего врача, количество выполненных лабораторных тестов, результат теста HbA1c, диагноз, количество лекарств, диабетических лекарств, количество амбулаторных пациентов, стационарные и неотложные посещения за год до госпитализации и т. д.

### Detecting Malware Using A Hybrid Approach Datasets

Всего в данном датасете присутствует:

* 2955 строк
* 1087 столбцов

Информация по файлам:

* staDynBenignLab.csv: 1086 функций, извлеченных из 595 файлов в MS Windows 7 и 8, полученный каталог Program Files.
* staDynVxHeaven2698Lab.csv: 1087 функций, извлеченных из 2698 файлов набора данных VxHeaven.
* staDynVt2955Lab.csv: 1087 функций, извлеченных из 2955, предоставленных Virus Total в 2018 году.

Информация об атрибутах:

* Статические функции: ASM, Hex-дамп и PE-заголовок (дискретный, непрерывный).
* Динамические функции: извлечены из песочницы с кукушкой.

## Этапы работы программы

Работа программы состоит из следующих последовательных этапов:

1. Загрузка и предобработка датасетов.
   1. Перемешивание данных.
   2. Удаление строк, в которых присутствуют пустые значения.
   3. Замена строковых значений на категориальные.
2. Построение матрицы корреляций и составление списка со слабой корреляцией.

2.1. Получение матрицы корреляций.

2.2. Получение списка столбцов со слабой корреляцией.

2.3. Отрисовка матрицы корреляции.

2.4. Сохранение матрицы корреляции в jpeg файл.

1. Подготовка данных для метода главных компонент.

3.1. Масштабирование (стандартизация). Предобработка данных, после которой каждый признак имеет среднее 0 и дисперсию 1.

1. Реализация метода главных компонент.

4.1. Создание и «обучение» метода главных компонент (с количеством главных компонент, равных количеству столбцов).

4.2. Получение суммы нагрузок (суммы дисперсий).

1. Реализация критериев выбора числа главных компонент (Кайзера, сломанной трости, каменистой осыпи).

5.1. Реализация критерия Кайзера (выбор только тех факторов, у которых собственные значения больше 1).

5.2. Реализация критерия сломанной трости.

5.3. Реализация критерия каменистой осыпи.

1. Получение результирующего значения числа главных компонент (рассчитывается как среднее арифметическое по предыдущим критериям).

6.1. Сохранение полученных значений главных компонент по каждому критерию в txt файл.

6.2. Вывод полученных значений главных компонент по каждому критерию на графике с последующим сохранением в jpeg файл.

1. Построение матрицы факторной нагрузки для всех значимых главных компонент.

7.1. Получение результирующей матрицы после применения метода главных компонент.

7.2. Расчёт матрицы факторной нагрузки.

7.3. Сохранение матрицы факторной нагрузки в xlsx файл.

1. Определение суммарной нагрузки для признаков. Составление ранжированного списка признаков (по убыванию нагрузки).

8.1. Получение списка с суммарной нагрузкой для признаков.

8.2. Сохранение полученного списка в xlsx файл.

1. Анализ остатков.

9.1. Получение списка остатков.

9.2. Сохранение полученного списка в файл.

## Результаты работы программы

Описанный в данной работе метод обработке данных был протестирован на трёх описанных датасетах. Получилось, что метод обработки данных хорошо работает на всех рассматриваемых датасетах.

После составления списка со слабой корреляцией получились следующие данные:

* 7 столбцов со слабой корреляцией было определено в датасете с 22 признаками.
* 6 столбцов со слабой корреляцией было определено в датасете с 55 признаками.
* Более 100 столбцов со слабой корреляцией было определено в датасете с 1087 признаками.

Списки столбцов хранятся в файлах «…\_low\_corr\_list.txt».

Количество различных главных компонент по различным критериям и различным датасетам представлены на рисунках 1-3, матрица корреляции для датасета «Anuran\_Calls\_Dataset\_22» представлена на рисунке 4.

Исходя из рисунков 1-3 можно сделать вывод о том, что для датасета «DetectingMalwareUsingAHybridApproach\_Datasets\_1087» метод главных компонент является «хорошо» применимым, поскольку при оставлении уже 20 главных компонент возможно сохранить большую часть (>90%) информации, представленной в датасете. Для датасета «Anuran\_Calls\_Dataset\_22» метод главных компонент является в «средней степени» применимым, поскольку применение метода главных компонент позволит сохранить достаточно большую часть информации, однако, исходя из небольшого числа признаков и с сохранением достаточно большого количества остаточной дисперсии (~0,05) для «последних» главных компонент, каждая из представленных главных компонент содержит значимое количество информации. Для датасета «Diabetic\_Dataset\_55» метод главных компонент является «плохо» применимым, поскольку каждая из главных компонент несёт большое количество значимой информации, а также потому, что значение остаточной дисперсии уменьшается не равномерно, а ближе к функции .

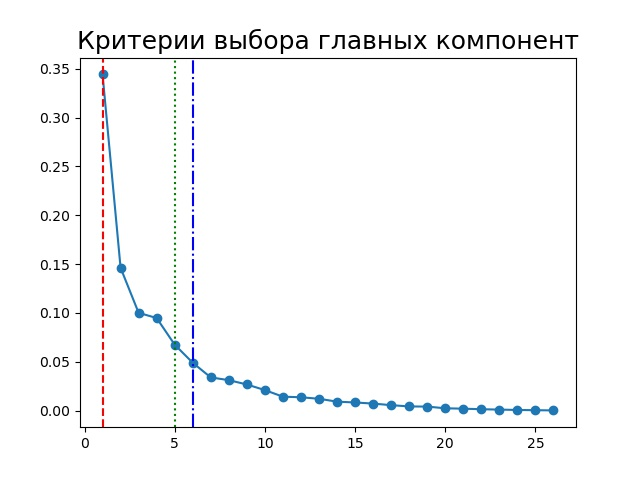


Рис. 1. Реализация критериев выбора числа главных компонент для датасета «Anuran\_Calls\_Dataset\_22»\*

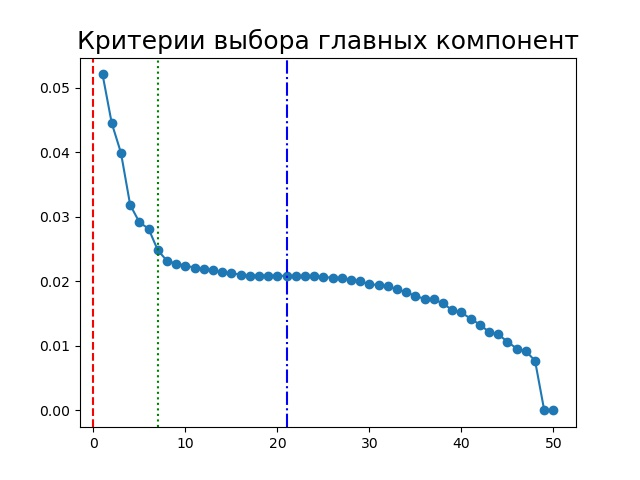


Рис. 2. Реализация критериев выбора числа главных компонент для датасета «Diabetic\_Dataset\_55»

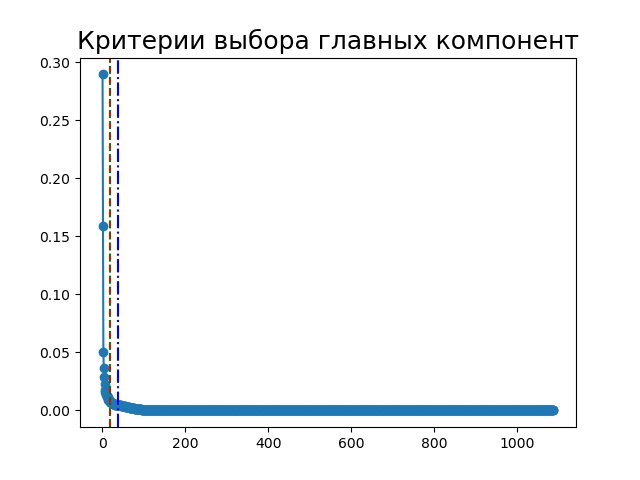


Рис. 3. Реализация критериев выбора числа главных компонент для датасета «DetectingMalwareUsingAHybridApproach\_Datasets\_1087»

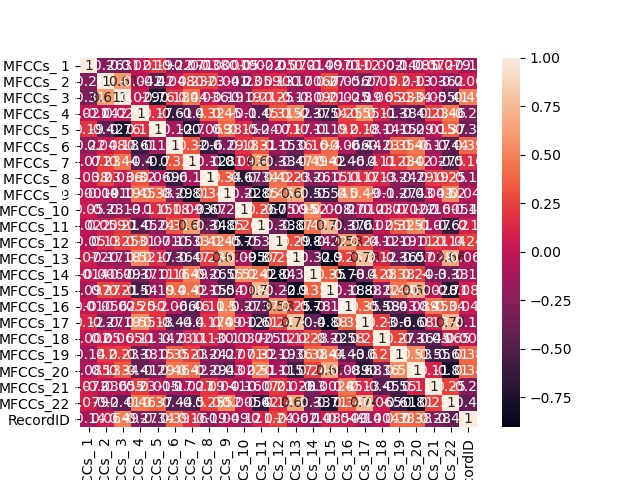


Рис. 4. Матрица корреляции для датасета «Anuran\_Calls\_Dataset\_22»

\* Красная линия – критерий сломанной трости.

Синяя линия – критерий Кайзера.

Зелёная линия – критерий каменистой осыпи.

Список остатков с датасетов «Anuran\_Calls\_Dataset\_22» и «Diabetic\_Dataset\_55» представлены в таблицах 1-2. Исходя из представленных остатков, после применения метода главных компонент к датасету «Anuran\_Calls\_Dataset\_22» остаётся достаточно большая остаточная дисперсия (среднее: -3,87), что говорит о слабой применимости метода главных компонент для данного датасета, а после применения метода главных компонент к датасету «Diabetic\_Dataset\_55» остаётся небольшая остаточная дисперсия (среднее: -0,46), что говорит о достаточно большой применимости метода главных компонент для данного датасета.

|  |  |
| --- | --- |
| Компонента | Остаток |
| comp\_0 | -7,49737 |
| comp\_1 | -3,3251 |
| comp\_2 | -2,32118 |
| comp\_3 | -2,3249 |
| mean | -3,86714 |

Таблица 1. Список остатков для датасета «Anuran\_Calls\_Dataset\_22»

|  |  |
| --- | --- |
| Компонента | Остаток |
| comp\_0 | -1,34242 |
| comp\_1 | -1,30537 |
| comp\_2 | -0,79088 |
| comp\_3 | -1,0039 |
| comp\_4 | -0,50046 |
| comp\_5 | -0,7678 |
| comp\_6 | -0,23747 |
| comp\_7 | 0,017228 |
| comp\_8 | -0,00259 |
| mean | -0,45644 |

Таблица 2. Список остатков для датасета «Diabetic\_Dataset\_55»

Остальные результаты работы программы представлены в папке «Results».

# Список используемой литературы

1. UCI Machine Learning Repository – Anuran Calls (MFCCs) Data Set. URL: <https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29> (дата обращения: 05.01.21).
2. UCI Machine Learning Repository – Diabetes 130-US hospitals for years 1999-2008 Data Set. URL: <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008> (дата обращения: 05.01.21).
3. UCI Machine Learning Repository – Malware static and dynamic features VxHeaven and Virus Total Data Set. URL: <https://archive.ics.uci.edu/ml/datasets/Malware+static+and+dynamic+features+VxHeaven+and+Virus+Total> (дата обращения: 05.01.20).