ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет информатики, математики и компьютерных наук**

**Направление**

**«Интеллектуальный анализ данных»**

Шадрина Алина Михайловна

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

На тему «Исследование Методов Обработки Данных Системы Искусственного Обоняния»

|  |  |
| --- | --- |
|  | Научный руководитель  д-р технических наук, проф.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  В.В. Крылов |

Нижний Новгород, 2017

**Содержание**

[ВВЕДЕНИЕ 3](#_Toc473902434)

[**1.** **ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ** 6](#_Toc473902435)

[**1.1** **Принципы построения систем «Электронный нос»** 6](#_Toc473902436)

[**1.2 Обзор существующих систем и их приложений** 6](#_Toc473902437)

[**1.3 Описание данных и подходы к их обработке** 6](#_Toc473902438)

[**2.** **МЕТОДЫ ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ ИСКУССТВЕННОГО ОБОНЯНИЯ «МАГ-8»** 10](#_Toc473902439)

[**2.1 Описание входных данных и формирование датасетов** 10](#_Toc473902440)

[**2.2 Подходы к симуляции откликов массива сенсоров «МАГ-8»** 13](#_Toc473902441)

[***2.2.1 Добавление константы*** 13](#_Toc473902442)

[***2.2.2*** ***Модели Ленгмюра и Поляни*** 15](#_Toc473902443)

[**2.3 Исследование данных в задаче оценки безопасности полимерных материалов** 15](#_Toc473902444)

[***2.3.1 Автокорреляция и радиус автокорреляции*** 15](#_Toc473902445)

[***2.3.2 Графы кросс-корреляции*** 16](#_Toc473902446)

[***2.3.3 Тест стационарности*** 18](#_Toc473902447)

[***2.3.4 Сглаживание полиномом N степени*** 19](#_Toc473902448)

[***2.3.5 Авторегрессионные модели*** 21](#_Toc473902449)

[**2.4 Применение алгоритмов машинного обучения** 22](#_Toc473902450)

[**2.5 Применение нейронных сетей** 22](#_Toc473902451)

[ЗАКЛЮЧЕНИЕ 22](#_Toc473902452)

[СПИСОК ЛИТЕРАТУРЫ 23](#_Toc473902453)

[**ProjectProposal: СПИСОК ЛИТЕРАТУРЫ** 23](#_Toc473902454)

[Statistics: Ссылки 23](#_Toc473902455)

# ВВЕДЕНИЕ

Предметом исследования выпускной квалификационной работы на тему “Исследование Методов Обработки Данных Системы Искусственного Обоняния” являются методы интеллектуального анализа данных, которые позволили бы автоматически интерпретировать показания сенсоров в системах искусственного обоняния. Объект исследования — матрицы откликов пьезокварцевых сенсоров системы «Маг-8». При написании работы основное внимание было уделено изучению специализированной литературы и ресурсов Интернета, список источников состоит из 31 пункта.

В настоящее время в аналитической химии для исследования многокомпонентных смесей летучих соединений значительное развитие получает технология «электронный нос», благодаря низкой цене (в сравнении с газовыми хроматографами) и универсальности прибора. Такие приборы строятся на основе массива сенсоров нескольких видов, высокочувствительных к наиболее распространённым соединениям. Количество и вид сенсоров в этих системах может варьироваться в зависимости от решаемой задачи. В качестве выходных данных, доступных для дальнейшей обработки и анализа, прибор предоставляет матрицу изменений откликов сенсоров, время исследования, рабочие частоты сенсоров и некоторую служебную информацию.

Система «электронный нос» получает все больше приложений во многих областях, от оценки качества пищевых продуктов до обнаружения запрещенных грузов и диагностики некоторых заболеваний. Благодаря развитию электроники, появляются портативные приборы, пригодные для экспресс-анализа. Однако, интерпретация результатов требует значительного времени и присутствия эксперта.

Большой интерес в настоящее время представляет разработка интегрированных аналитических систем, образующих единый конвейер, начиная с измерения, через обработку и интеллектуальный анализ данных и заканчивая принятием решения. При этом, необходимо учитывать специфику подобных исследований: сначала исследователь определяется с характером веществ-маркеров, с которыми он планирует работать, исходя из этого подбирает селективные покрытия электродов пьезокварцевых резонаторов, затем тестирует этот массив, проводя первичные пробы отдельных веществ и их возможных смесей, и только после этого приступает к оценке тех объектов, которым посвящено исследование. Таким образом, несмотря на то, что методы анализа данных могут многое рассказать о каждом веществе из набора, применять их к отдельным матрицам откликов нецелесообразно из-за большого числа этих матриц. Этот этап должен быть либо скрыт от конечного пользователя, либо от него следует отказаться в пользу более высокоуровневых подходов. В качестве основного подхода здесь может быть использована классификация на N классов.

Новизна и актуальность данного исследования следуют из уникальности источника данных — прибора «МАГ-8». В данный момент, группой разработчиков используются исключительно графический метод анализа, к недостатку которого нужно отнести трудоемкость и требование к высокой квалификации человека, который работает с прибором. Данная работа позволит шагнуть от этапа измерений до этапа интерпретации результата, минуя рутинный анализ отдельных маркеров и их смесей, а также требующий повышенного внимания этап анализа объектов, на которые нацелено исследование.

Основная цель работы состояла в том, чтобы построить прототип системы анализа данных для «МАГ-8». Для этого были выполнены следующие задачи:

1. Подбор способа генерации достоверных искусственных данных
2. Обучение алгоритма на сгенерированных данных
3. Обучение на исключительно исходных данных
4. Апробация на матрицах проб, взятых с 75 детских игрушек, с целью обнаружения присутствия наиболее полного набора веществ-маркеров.

Оценка результатов будет проведена путем сравнения полученных меток классов с результатами исследования группы профессора Кучменко.

Таким образом, данная работа находится на стыке аналитической химии и машинного обучения. Логика исследования обуславливает структуру работы, состоящую из введения, двух глав, заключения, библиографии и приложения. В 1 главе сделан обзор предметной области [] . Вторая глава посвящена применению [] . В заключении подводятся итоги исследования и рассматриваются направления дальнейшего развития. Приложения содержат

# **ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ**

## **Принципы построения систем «Электронный нос»**

## **1.2 Обзор существующих систем и их приложений**

Зарубежный рынок систем «электронный нос» представлен следующими приборами: FOX с массивом из 6 металл-оксид-электронных сенсоров, и более продвинутая его версия GEMINI (до 18 сенсоров), комбинированная система «электронный нос» с газовым хроматографом HERACLES производства компании Alpha MOS, портативный газовый хроматограф zNose GS/SAW от Electronic Sensor Technology Inc, MOSES II немецкой компании GSG Meß- und Analysengeräte Vertriebsgesellschaft mbH и KAMINA (Германия). В России подобный прибор разработан в Воронежском государственном университете инженерных технологий группой под руководством доктора химических наук, профессора Татьяны Анатольевны Кучменко и называется «МАГ-8».

## **1.3 Описание данных и подходы к их обработке**

Следующий обзор основан на изучении исследований ученых-химиков, описывающих применение системы «электронный нос» и анализ результатов, и статьях из области анализа данных, описывающих общие подходы к работе в системах искусственного обоняния.

Основной подход, применяемый в данный момент профессором Кучменко, состоит в анализе визуальных отпечатков откликов сенсоров (кинетических и максимумов) в равновесной газовой фазе. Эти графики имеют вид полярной диаграммы, где осями являются временные метки, а факторами – значения сенсоров в момент времени t. В [1] делается вывод о том, что «электронный нос» «МАГ-8» превосходит физико-химические показатели в задаче оценки органолептических характеристик вина.

Сравнительный анализ возможностей интегрального анализатора газа “VOCmeter” (Германия) и дифференциального анализатора «МАГ-8» приведен в [2] на примере задачи количественной и качественной оценки легколетучей фракции ароматических добавок для мясного сырья. Сделан вывод о том, что результат, получаемые с использование отечественной разработки превосходят результаты “VOCmeter” и в большей степени коррелируют с результатами газохроматографии.

Для идентификации веществ по визуальным отпечаткам используется расчет таких геометрических параметров фрагментов фигуры визуальных отпечатков, как площади под кривыми i-х пьезосенсоров Si, площадь «визуального отпечатка» массива сенсоров, соотношение проекций сигналов сенсоров I и j на сигнал сенсора n и угол между этими проекциями (в радианах). Диссертация [3] на соискание степени кандидата химических наук Дроздовой Е.В. под руководством Кучменко Т.А. полностью посвящена апробации данного подхода в задаче оценки безопасности изделий из полимерных материалов на основе проб воздуха в локальных точках вблизи их поверхности. Кроме того, в данной работе показана возможность применения PCA и кластеризации как методов обработки данных, получаемых с помощью электронного носа «Маг-8».

Большое количество работ [4][5] посвящено подбору массивов сенсоров для решения определенных задач, что говорит о специфичности этих сенсоров и позволяет предположить, что обучение универсального алгоритма классификации невозможно – для решения каждой отдельной задачи он должен обучаться на отдельном наборе.

Основной упор в рассмотренных выше работах делается на применение данной системы в экспресс-анализе, что так же обосновывает необходимость разработки такой системы анализа данных, которая бы позволила ускорить процесс принятия решения исследователем.

Такие системы, построенные на основе двух компонентов - «электронного носа» и автоматизированной системы распознавания, нашли применение в медицине, охране окружающей среды и пищевой промышленности. В работе Келлера [6] описан прототип такой системы и показан успешный пример применения как нейронных сетей, обученных методом обратного распространения ошибки, так и сетей fuzzy ARTMAP, сочетающих в себе аппарат нечеткой логики и адаптивной резонансной теории. Обе архитектуры показали близкую точность – 92.9% и 93.4%, соответственно. Необходимо так же заметить, что обучение проводилось на сравнительно небольшой для нейронных сетей выборке в 619 объектов, а тестирование – на 196 объектах.

Еще в 1989 году в своей работе [7] Hoffheins показал, что, благодаря использованию многоканальных массивов сенсоров, нейронные сети успешно решают задачу распознавания паров летучих соединений, поскольку количество распознаваемых химических веществ в общем случае больше числа сенсоров. Эта фундаментальная работа дала важные результаты как по организации массива сенсоров, так и основу для дальнейших исследований архитектур нейронных сетей и способов представления входных данных. В работе предложен способ представления данных и показано, что худший результат показали сети Хопфилда, вероятно, и-за малой размерности входных данных, а наилучший – сеть Больцмана, которая не только возвращала лучшую метку, но также показывала следующего подходящего кандидата, что позволяло использовать эту сеть в задаче отображения концентрация газа. Сеть Хэмминга показала наилучший результат в распознавании смесей многих компонентов. Кроме того, показано, что алгоритмы кластеризации способны успешно разделять гексан и этанол, а также высокие и низкие концентрации смесей воды и этанола. Еще один важный результат данной работы состоит в том, что было найдено важное ограничение: сеть не способна распознавать неизвестные ей смеси веществ, присутствовавших в обучающей выборке.

Еще один успешный пример применения системы «электронный нос» описан в [8]: рассматривается целый набор задач по проверке грузов, которые каждый день решают сотрудники службы безопасности и таможенной службы в портах – обнаружение наркотических веществ, споров грибов и плесени, которые могут угрожать сельскохозяйственным культурам, опасных химикатов. В качестве метода выбрано построение ароматических профилей каждого контейнера в виде полярных графиков. Интересным так же является предложение использовать эти профили как своеобразные «контрольные суммы» контейнеров, изменение которых можно было бы отслеживать на протяжении всего маршрута и таким образом выявлять, в каком из портов к содержимому контейнеров был добавлен контрабандный товар. Важным отличием от прочих работ является использование единственного сенсора, что существенно сокращает стоимость применения такого устройства.

Работа [9] описывает все существующие подходы к анализу данных в задаче распознавания запахов, включая широкий спектр графических методов таких, как полярные диаграммы и иерархическая кластеризация, PCA, алгоритмы кластеризации и классификации, линейный и квадратичный дискриминантный анализ, нейронные сети, методы нечеткой логики и генетические алгоритмы. Делается особый акцент на необходимость нормализации данных и приводится около десятка способов нормализации, рассматриваются способы отбора признаков (сенсоров). В целом, данная работа скорее обобщает существующие подходы, чем привносит что-то принципиально новое.

Недостатком предыдущих работ по анализу данных является то, что они не освещают возможности многослойных нейронных сетей. Статья [10] восполняет этот пробел. Авторы рассматривают влияние смесей на качество распознавания (для простоты, берут смеси только двух веществ) и сравнивают данные от двух видов сенсоров (пьезокварцевых и металл-оксидных) и делают вывод о том, что электронные носы можно применять во многих областях (не может быть!).

# **МЕТОДЫ ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ ИСКУССТВЕННОГО ОБОНЯНИЯ «МАГ-8»**

## **2.1 Описание входных данных и формирование датасетов**

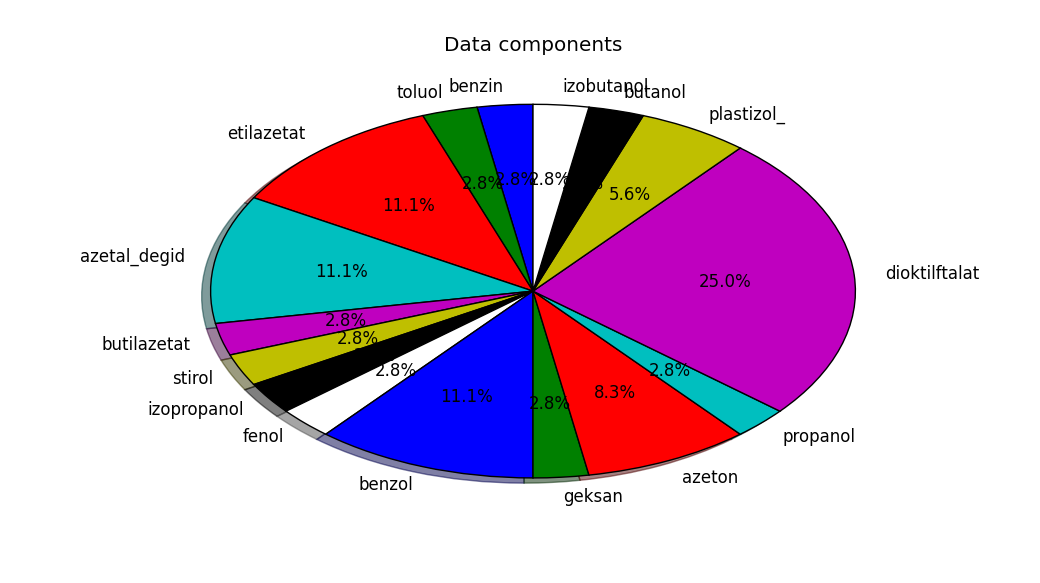
### ***2.1.1 Источники данных***

Для начального исследования и предобработки получены 33 объекта тренировочной выборки, которые представляют собой отдельные вещества, 4 объекта валидационной – смеси двух веществ, и 75 новых объектов, которые необходимо классифицировать. Каждый объект хранится в файле вида название\_вещества.XLS и представляет собой таблицу, содержащую следующие блоки:

1. Шапка: название (вещества или игрушки, или состав смеси), продолжительность (всегда 120 с), тип (обычно значение «измерение», назначение этого поля не исследовалось), статистические данные (обычно значение «нет», назначение этого поля не исследовалось), начало (число-время начала измерения).
2. Информация о сенсорах: 8 пар вида «название сенсора – базовая частота сенсора».
3. Матрица 120 х 8, где столбцы соответствуют сенсорам, а строки – временным отсчетам. Таким образом, каждый элемент матрицы отражает изменение частоты сенсора i (i=[1,8]) в момент времени j (j=[0,120])

Для обучения получены следующие вещества (см.рисунок 1):

* диоктилфталат – 9 шт. в разных концентрациях на разных носителях;
* ацетальдегид, ацетон, бензол, этилацетат - 4 шт. в разных концентрациях;
* пластизоль – 2 шт.;
* бензин, бутанол, бутилацетат, гексан, изобутанол, изопропанол, пропанол, стирол, толуол, фенол – 1шт.;

Метки классов извлекаются автоматически из названий файлов. Правило именования файлов выглядит следующим образом: «название\_вещества [концентрация] мкл на [носитель]» (носитель и концентрация опциональны). Для формирования датасета было решено не делать различий между одним и тем же веществом в разной концентрации или на разных носителях, поэтому алгоритм извлечения меток классов состоит в том, чтобы разрезать название файла по пробелам и сохранять первый (в индексах списков Python - нулевой) элемент.

**Рисунок 1 – Состав тренировочного множества**

В процессе исследования тренировочных данных выявлены следующие проблемы и предложены пути их решения:

1. **Проблема:** опечатки в названия файлов

**На что влияет:** на автоматическое извлечение меток классов

**Решение:** переименовать файл вручную

1. **Проблема:** кириллица в названиях файлов

**На что влияет:** бинаризация меток классов не работает с кириллицей => обучение с учителем невозможно

**Решение:** выполнить примитивную транслитерацию (см. файл xls\_parser.py, метод cyrillic2latin)

1. **Проблема:** Не все файлы содержат матрицу подходящего размера

**На что влияет:** невозможно сформировать датасет

**Решение:** включит проверку при считывании матриц (см. файл xls\_parser.py, метод main)

1. **Проблема:** датасет слишком маленький и несбалансированный

**На что влияет:** невозможно качественное обучение с учителем

1. **Решение:** а) работа с базой данных sniff (см. п.1.2) и б) генерация искусственных данных 8 (см. п. 1.4)

В процессе работы над данными изменился подход к формату датасета, поэтому в репозиторий содержит более старые данные (test\_data, train\_data, train\_labels). В частности, в новом датасете смеси вынесены в отдельный файл, изменились названия датасетов, а также изменился формат файлов на более удобный для дальнейшего парсинга: теперь столбцы матриц разделены прямым слешем «|», а отсчеты внутри столбцов – точкой с запятой «;».

Предполагалось, что из базы данных sniffdb.sdf удастся извлечь дополнительные данные и таким образом расширить множества веществ и смесей. Файл базы данных создан в Microfost SQL Server Compact Edition, что само по себе является серьезным недостатком, так как в команде были ноутбуки не только под ОС Windows, но также под OS X и Linux. Кроме того, это устаревший формат, несовместимый с прочими инструментами Microsoft для работы с базами данных. Для обеспечения кросс-платформенного доступа база с помощью инструмента SDF Viewer была конвертирована в sql-файл, который, в свою очередь был скорректирован для работы с mysql5.5.

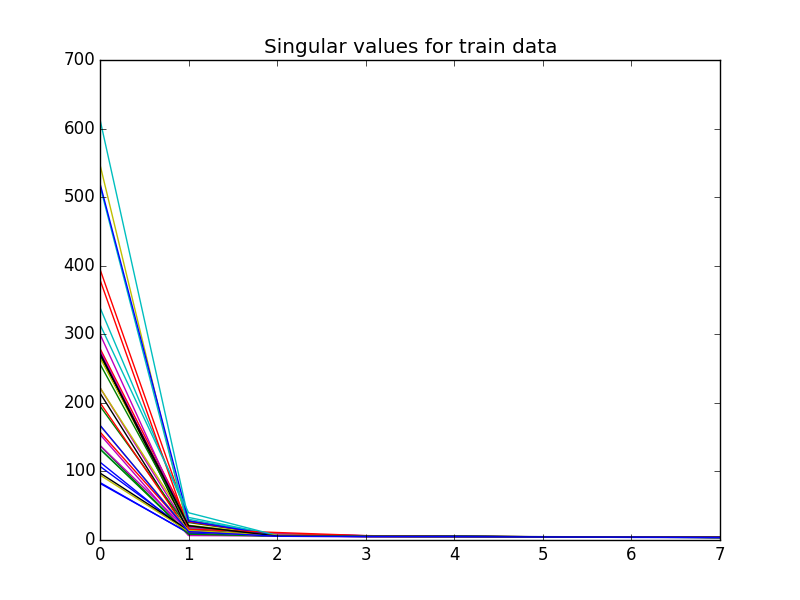
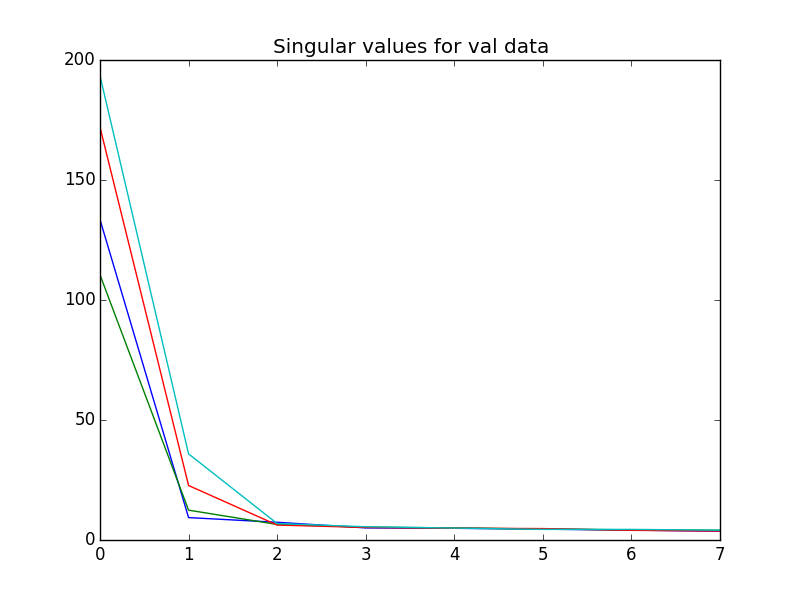
База данных содержит:

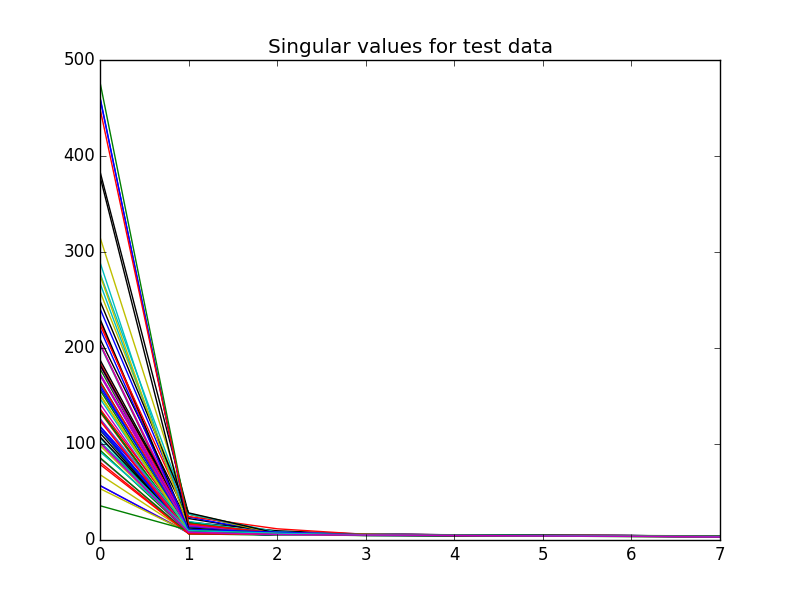
* 8 таблиц: Data, GroupTree, Mask, MaskData, MeasureProfile, MeasureProfileData, Measures, Sensors;
* 430 записей измерений и прочистки сенсоров;
* 250 различных объектов;
* 103 объекта длины 120 (объекты были отфильтрованы по полю FullLength как по самому критичному, дальнейший анализ извлеченных измерений предполагалось провести вручную);
* Огромное количество мусора.

Для автоматизации работы с БД был написан скрипт на Python parse\_sql.py. В процессе анализа извлекаемой информации было обнаружено, что в таблице Data хранятся не изменения частот сенсоров, а значения частот. Вычисление необходимых матриц показало, что частоты изменяются «ступенькой», что отличается от уже имеющихся данных из XLS, где они изменяются плавно. Возможная причина состоит в том, что система «электронный нос» совершенствовалась, поэтому данные из БД сделаны более старой версией анализатора, а данные в XLS - более новые. Таким образом, было принято решение отказаться от дальнейшей работы с этой базой.

### ***2.1.2 Исследование исходных данных***

В качестве начального шага было проведено исследование данных методов сингулярного разложения. На рисунке 2 видно, что во всех трёх датасетах основную информацию несёт только 1 компонента. Остальные можно игнориоровать, таким образом превратив датасет в множество векторов, а не в множество матриц.





**Рисунок 2 – Графики сингулярных чисел для тренировочного, валидационного и тестового датасетов**

## **2.2 Подходы к симуляции откликов массива сенсоров «МАГ-8»**

### ***2.2.1 Добавление константы***

Генерация искусственных данных преследовала 2 цели:

1. Получить б**о**льшее тренировочное множество
2. Сбалансировать классы

В данной работе реализован простейший алгоритм, вдохновлённый генерацией искусственных данных в обработке изображений: нужно просто добавлять константу из эталонных данных. В качестве альтернативных вариантов рассматривались остальные арифметические операции.

Кроме того, подбирался следующий параметр – количество генерируемых данных каждого класса. Чтобы удовлетворить требованию сбалансированности, это число рассчитывалось следующим образом:

N\_new = N\_exp - N\_real,

где N\_new – количество данных одного класса, которые необходимо сгенерировать, N\_exp – количество данных одного класса, которое мы хотели бы получить, и N\_real – количество «настоящих» данных в датасете.

Оба параметра – арифметическая операция и количество новых данных – подбирались по результатам обучения нескольких классификаторов. Описанные методы не являются лучшими, но в условиях ограниченности времени они работают.

Способ генерации смесей не найден, так как мы не располагаем достаточным для анализа количеством смесей. Однако, этот способ значительно повысил бы точность классификации новых данных.

Перейдём непосредственно к задаче классификации и обучению. Для подбора подходящего метода генерации данных в обучении и тестировании использовались только данные тренировочного датасета. В таблице 1 приведены результаты подбора наилучшего метода – добиваемся присутствия 10 объектов каждого класса с помощью операций прибавления или вычитания констант.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Данные | SVM | Knn | Rand.Forest | GussianNB |
| 1 | Начальные данные | 20% / 36% | 48% / 0% | 100% / 9% | 76% / 9% |
| 2 | Train – начальные,  Test - сгеренированные | 2% / 0% | 5% / 0% | 100% / 0% | 61% / 0% |
| 3 | Cross validation, 10 (+,-) | 18% / 20% | **92% / 81%** | 100% / 64% | 20% / 22% |
| 4 | Cross validation, 20 (+,-) | 15% / 17% | 87% / 78% | 100% / 62% | 20% / 20% |
| 5 | Cross validation, 10 (+,-,\*,\) | 13% / 7% | 84% / 74% | 97% / 66% | 6% / 7% |

**Таблица 1 – Результаты подбора метода генерации данных**

Затем, когда был выбран наилучший метод генерации, он был протестирован на смесях. Результаты приведены в таблице 2. Наилучший результат показал наивный байесовский классификатор – при обучении на смеси настоящих и искусственных веществах 83% и при тестировании на смесях точность 82%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Данные | SVM | Knn | Rand.Forest | GussianNB |
| Train – начальные + сгенерированные методом 10 (+,-),  Test - смеси веществ | **83% / 87%** | 98% / 81% | 100% / 81% | 83% / 82% |

**Таблица 2 – Результаты тестирования метода генерации данных**

Последним шагом в тренировочное множество были объединены начальные вещества, данные, сгенерированные методом 10 (+,-), и смеси и проведена классификация новых данных. Так как нет возможности проверить точность классификации, было принято решение задействовать больше алгоритмов и считать верным тот результат, который показало наибольшее число алгоритмов.

### ***Модели Ленгмюра и Поляни***

## **2.3 Исследование данных в задаче оценки безопасности полимерных материалов**

### ***2.3.1 Автокорреляция и радиус автокорреляции***

По автокоррелограмме можно сделать вывод о периодичности рассматриваемых рядов и выявить наличие взаимосвязанных изменений ряда.

Начнем с веществ-маркеров. Коррелограмма негладкая для: ацетальдегида 0 (S7), ацетальдегида 3 (S2, S3, S7), ацетон 6 (S1, S7), бензин 7 (S2, S7), диоктилфталат 15 (S1), диоктилфталат 16 (S7), диоктилфталат 17 (S7), диоктилфталат 18 (S1, S7), диоктилфталат 20 (S7), диоктилфталат 21 (S2, S4, S7, S8), диоктилфталат 22 (S1, S4, S7, S8), диоктилфталат 23 (S7), этилацетат 32 (S1), этилацетат 33 (S2), этилацетат 35 (S2), гексан 14 (почти весь), пластизоль 26 (S2, S4, S5), пластизоль 27 (S3, S4, S7) — всего 32 вектора. Можно предположить наличие скрытых синусоидальных зависимостей.

Анализ радиусов автокорреляции показывает, что переход через 0 в большинстве графиков происходит в районе 40 отсчёта. Однако существуют такие графики, в которых поведение автокорреляционной функции резко отличается: ацетальдегид 0 (S7), бензин 7 (S2), гексан 14 (S2, S7), диоктилфталат 15 (S1, S7), диоктилфталат 17 (S7), диоктилфталат 18 (S7), диоктилфталат 22 (S1, S4, S7, S8), пластизоль 26 (S2, S4, S8), пластизоль 27 (S3, S5, S6, S8), этилацетат 32 (S1).

Наблюдаются пересечения между множеством векторов, которые имеют сильно отличающуюся от остальных коррелограмму, и множеством векторов с сильно отличающимся вектором автокорреляции.

Теперь перейдем к смесям. Коррелограмма негладкая для: ДОФ+ацетальдегида 3 (S1, S2, S3), ДОФ+ацетон 0 (S1, S2, S3, S4, S7, S8), ДОФ+бензол 2 (S1) — всего 10 векторов. Можно предположить наличие скрытых синусоидальных зависимостей. Все коррелограммы. ДОФ+этилацетат 1 гладкие.

Анализ радиусов автокорреляции показывает, что переход через 0 в большинстве графиков происходит в районе 43-44 отсчёта. Однако существуют такие графики, в которых поведение автокорреляционной функции резко отличается: ДОФ+ацетон 0 (S1), но отклонение здесь не значительно (32 отсчет).

### ***2.3.2 Графы кросс-корреляции***

Для каждого вещества построены полносвязные взвешенные графы с 8 вершинами (S1, …, S8), где вершина sj (j=1…8) соответствует сенсору, а вес ребра ij присваиваются согласно значению корреляции между сенсорами si и sj.

Цвет ребра показывает, какому промежутку значений функции взаимной корреляции принадлежит данное ребро. Для наглядности рассматриваются следующие промежутки: вес ребра больше 0.9 – это сильно положительно коррелированные сенсоры, вес в интервале (0.3, 0.9] – некое промежуточное, среднее значение корреляции, (0, 0.3] – слабая положительная корреляция. Аналогично для отрицательной корреляции. Положительная корреляция отражена оттенками красного, отрицательная – оттенками синего.

|  |  |  |  |
| --- | --- | --- | --- |
| **Вещество** | **N** | **Тип** | **Характеристики** |
| Ацетальдегид | 4 | + | 2 графа, содержащие исключительно сильную положительную корреляцию (№1 и №2);  2 графа, содержащие сильную положительную и преобладающую среднюю корреляцию (№0 и №3); |
| Ацетон | 3 | + | 2 графа, содержащие исключительно сильную положительную корреляцию (№4 и №5);  1 граф, содержащий преобладающую сильную положительную и среднюю корреляцию (№6) от сенсора **S7** до каждого сенсора в виде **остовного дерева**; |
| Бензин | 1 | + | 1 граф, преобладает средняя положительная корреляция (№7); |
| Бензол | 4 | + | 2 графа, исключительно сильную положительную корреляцию (№9 и №11);  2 графа, содержащие преобладающую сильную положительную и среднюю корреляцию (№8 и №10), в обоих графах присутствуют ребра со слабой корреляцией (S3, S6) и (S3, S1); |
| Бутанол | 1 | + | 1 граф, только сильная положительная корреляция (№12); |
| Бутилацетат | 1 | + | 1 граф, только сильная положительная корреляция (№13); |
| ДОФ | 9 | +\- | Во всех рассматриваемых графах присутствуют ребра всех цветов, кроме темно-синего цвета, соответствующего максимальной отрицательной корреляции;  В 7 графах из 9 (№№ 15, 16, 17, 19, 20, 21, 23) наблюдается явное **остовное дерево** из вершины **S7**, аналогичное ацетону, и представленное одним из трех видов корреляции (слабой положительной, средней отрицательной). В 2 оставшихся графах данное остовное дерево так же представлено, но, видимо, рассматриваются некие пограничные значения, поэтому отдельные ребра этого дерева ребра попадают в смежные интервалы (№18 – средняя и слабая положительная, №22 – слабая положительная и слабая отрицательная). Тем не менее, остовное дерево явно выделяется. |
| Этилацетат | 4 | + | 2 графа, содержащий сильную положительную и преобладающую среднюю корреляцию (№32 и №35);  2 графа, содержащие исключительно сильную положительную корреляцию (№33, №34);  Средняя положительная корреляция всегда представлена ребрами (S2, S5), (S2, S8).  В графе №33 представлено почти полное **остовное дерево** из вершины **S2**, состоящее из ребер со слабой положительной корреляцией и одного ребра со средней (пограничные значения?) |
| Фенол | 1 | + | 1 граф, только сильная положительная корреляция (№31); |
| Гексан | 1 | +\- | 1 граф, содержащий сильную положительную, среднюю положительную, слабую положительную, слабую отрицательную (№14) корреляции;  наблюдается явное **остовное дерево** из вершины **S7**, аналогичное ацетону, и представленное слабой отрицательной корреляцией;  наблюдается неполное **остовное дерево** из вершины **S2**, представленное слабой и средней положительной корреляцией (вершины S3 и S4 имеют среднюю корреляцию, возможны пограничные значения);  наблюдается клика (S4, S5, S6); |
| Изобутанол | 1 | + | 1 граф, только сильная положительная корреляция (№24); |
| Изопропанол | 1 | + | 1 граф, только сильная положительная корреляция (№25); |
| Пластизоль | 2 | +\- | 2 графа, преобладает средняя положительная корреляция;  Наблюдается **остовное дерево** из вершины **S7**, состоящее из ребер со слабой и средней отрицательной корреляцией (№26) слабой и средней отрицательной и слабой положительной корреляцией (№27) |
| Пропанол | 1 | + | 1 граф, только сильная положительная корреляция (№28); |
| Стирол | 1 | + | 1 граф, преобладает сильная положительная корреляция (№29); |
| Толуол | 1 | + | 1 граф, только сильная положительная корреляция (№30); |

**Таблица 1 – Сравнение графов кросс-корреляций сенсоров**

Согласно таблице 1, для нас абсолютно неразличимы некоторые виды ацетальдегида, ацетона, бензола, бутанол, бутилацетат, фенол, изобутанол, изопропанол, пропанол, толуол, поскольку сенсоры у них имеют сильную взаимную корреляцию, а графы кросс корреляции идентичны.

**ТУТ БУДЕТ КАРТИНКА**

При анализе графов кросс-корреляций для смесей выявлены 2 остовных дерева – и вершины S2 и из вершины S7 (рис.1).

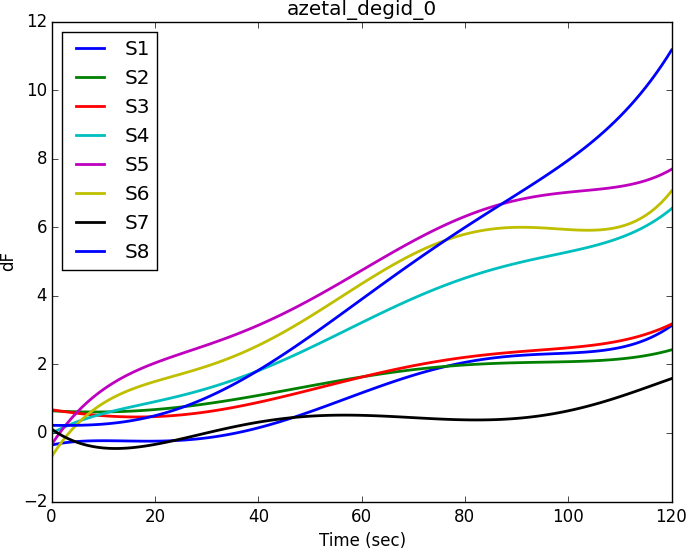
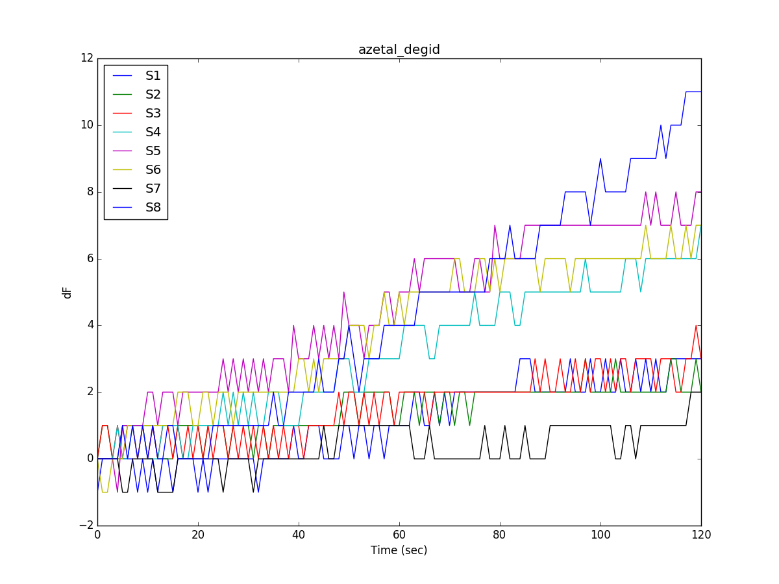
Для каждого вещества построены полносвязные взвешенные графы с 8 вершинами (S1, …, S8), где вершина sj (j=1…8) соответствует сенсору, а вес ребра ij присваиваются согласно значению корреляции между сенсорами si и sj.

|  |  |  |  |
| --- | --- | --- | --- |
| **Вещество** | **N** | **Тип** | **Характеристики** |
| ДОФ+ацетон 0 | 1 | + | В графе преобладает средняя положительная корреляция; |
| ДОФ+этилацетат 1 | 1 | + | Сложно определить, какой вид ребер преобладает – сильная положительная или средняя положительна корреляция.  Наблюдается остовное дерево из сенсора **S2** с одним ребром пограничной корреляции. |
| ДОФ+бензол 2 | 1 | + | В графе преобладает средняя положительная корреляция; |
| ДОФ+ацетальдегида 3 | 1 | + | В графе преобладает средняя положительная корреляция;  Наблюдаются остовное дерево из сенсора **S2;** |

### ***2.3.3 Тест стационарности***

Выполним вычисление скользящих статистик (среднего и стандартного отклонения) и тест на единичные корни. В пакете statsmodels реализован расширенный тест Дики-Фуллера (ADF — Augmented Dickey-Fuller) [6]. Его преимущество перед обычным тестом Дики-Фуллера (DF) состоит в том, что, благодаря включению первых разностей, он даёт возможность работать с авторегрессиями не только первого, но и более высоких порядков, так как авторегрессия еще не была исследована. В результате обнаружено, что среди векторов откликов сенсоров для веществ-маркеров существует 228 нестационарных рядов из 288 — то есть можно сделать вывод, что большинство из них нестационарны, и, следовательно, большинство рядов будут соответствовать моделям авторегрессий более высокого порядка. Для веществ существует 24 нестационарных ряда из 32 — то есть можно сделать вывод, что большинство из них нестационарны, и, следовательно, большинство рядов будут соответствовать моделям авторегрессий более высокого порядка. Смесь диоктилфталата и этилацетата полностью нестационарна.

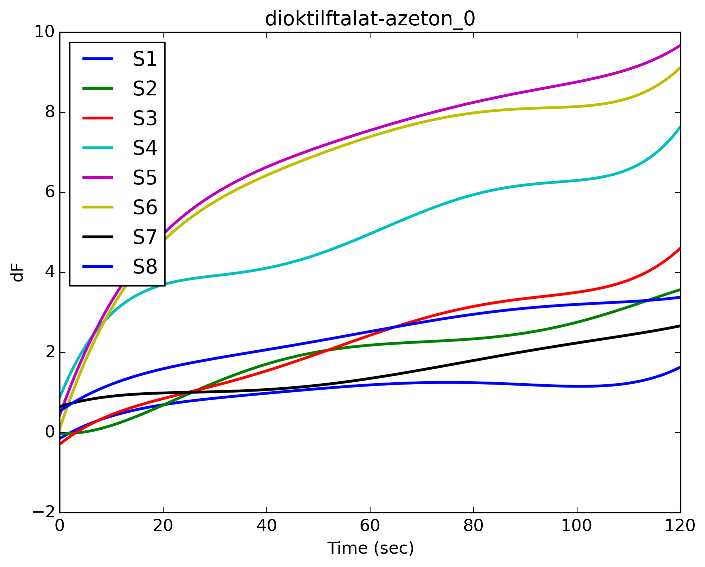
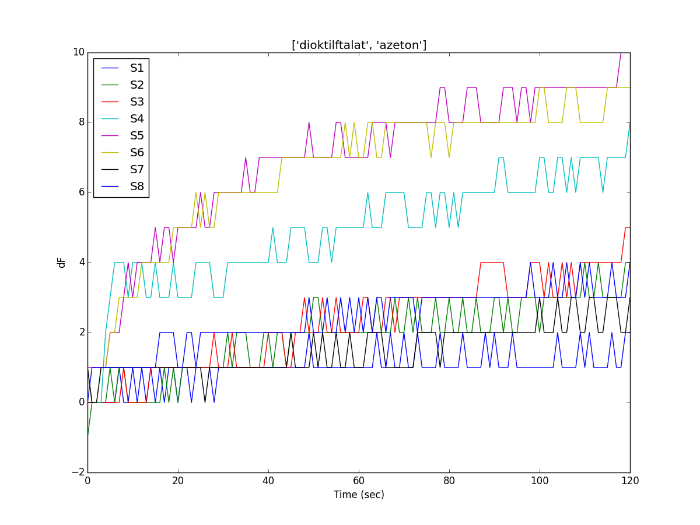
### ***2.3.4 Сглаживание полиномом N степени***

Исходные матрицы визуализированы с помощью кода visualization.py для каждого рассматриваемого вещества. Анализ результатов показал, что из-за постоянных колебаний пьезосенсора в пределах 1 Гц, полученные графики непригодны для работы. Приближение полиномом N степени позволило сгладить такой «шум» и оставить только общий вид графиков. Улучшение заметно на следующем примере. Справа исходный график, слева – сглаженный полиномом 5 степени.

Анализ полученных графиков приведен в таблице 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Вещество** | **N** | **Тип** | **Характеристики** |
| Ацетальдегид | 4 | + | Визуально очень похожи графики №1 и №2;  Между графиками №0 и №3 сходство проследить сложно; ы |
| Ацетон | 3 | + | Визуально похожи графики №4 и №5;  График №6 значительно отличается от №4 и №5; |
| Бензин | 1 | + | **Сложно сказать (№7);** |
| Бензол | 4 | + | Визуально похожи графики (№8, №9 и №10);  **Сложно сказать (№11);** |
| Бутанол | 1 | + | **Сложно сказать (№12);** |
| Бутилацетат | 1 | + | **Сложно сказать (№13);** |
| ДОФ | 9 | +\- | Среди рассматриваемых графиков нет сильно отличающихся от остальных.  Есть сходство между ДОФ |
| Этилацетат | 4 | + | (№32 и №35);  (№33, №34);  В графе №33 представлено почти полное **остовное дерево** из вершины **S2** |
| Фенол | 1 | + | **Сложно сказать** (№31); |
| Гексан | 1 | +\- | **Сложно сказать** (№14)  наблюдается явное **остовное дерево** из вершины **S7**,  наблюдается неполное **остовное дерево** из вершины **S2** |
| Изобутанол | 1 | + | **Сложно сказать** (№24); |
| Изопропанол | 1 | + | **Сложно сказать** (№25); |
| Пластизоль | 2 | +\- | Наблюдается **остовное дерево** из вершины **S7**,  (№26) (№27) |
| Пропанол | 1 | + | **Сложно сказать** (№28); |
| Стирол | 1 | + | **Сложно сказать** (№29); |
| Толуол | 1 | + | **Сложно сказать** (№30); |

**Таблица 2 – Сравнительный анализ кривых веществ**

Исходные матрицы визуализированы с помощью кода visualization.py для каждого рассматриваемого вещества. Анализ результатов показал, что из-за постоянных колебаний пьезосенсора в пределах 1 Гц, полученные графики непригодны для работы. Приближение полиномом N степени позволило сгладить такой «шум» и оставить только общий вид графиков. Улучшение заметно на следующем примере. Справа исходный график, слева – сглаженный полиномом 5 степени.

Анализ полученных графиков приведен в таблице 2.

|  |  |  |
| --- | --- | --- |
| **Вещество** | **N** | **Характеристики** |
| ДОФ+ацетон 0 | 1 |  |
| ДОФ+этилацетат 1 | 1 |  |
| ДОФ+бензол 2 | 1 |  |
| ДОФ+ацетальдегида 3 | 1 |  |

**Таблица 2 – Сравнительный анализ кривых смесей**

### ***2.3.5 Авторегрессионные модели***

Поскольку не все векторы стационарны, целесообразно остановить свой выбор на модели Бокса-Дженсинса [7] или ARIMA (Autoregressive Integrated Moving Average) [8]. Модель, реализованная в пакете statsmodels, предлагает удобный интерфейс как для задания параметров модели [9], так и для применения этой модели к данным [10]. Функция fit позволяет настроить такие глобальные параметры: “на лету” приводить исследуемый ряд к стационарному (*transparams*), выбрать один из трех методов работы с максимальным правдоподобием (*method*), учет тренда (*trend*), солвер (*solver*), количество итераций (*maxiter*).

Текущие глобальные параметры: transparams=True (defalut), так как не все исследуемые временные ряды стационарны, method=css-mle (default) – условный метод максимального правдоподобия, trend=c (default) – учитывать константу, solver=newton.

Подбор параметров p, d, q будет осуществляться автоматически из следующих интервалов: , поскольку более интеллектуальный выбор параметров затрудняется большим количеством векторов (8 векторов из 36 матриц веществ – 288 векторов), поведение которых отличается. В качестве критерия отбора выбрана минимизация информационного критерия Акаике.

Дальнейшие эксперименты будут проводиться с увеличением порядка модели.

ДЛЯ СМЕСЕЙ ДОБАВИТЬ

## **2.4 Применение алгоритмов машинного обучения**

## **2.5 Применение нейронных сетей**

# ЗАКЛЮЧЕНИЕ

При написании данной курсовой работы было проведено исследование электроэнцефалограмм 10 испытуемых, которые прослушивали 48 музыкальных последовательностей, представляющих собой ближние и дальние модуляции. Данные были получены в Лаборатории когнитивной психофизиологии ННГУ им. Лобачевского. Основной интерес с точки зрения нейрофизиологии представляла проверка гипотезы о том, что разные модуляции будут восприниматься слушателями по-разному.

В рамках бакалаврской выпускной квалификационной работы был проведен спектральный и статистический анализ ЭЭГ, однако из-за огромного количества полученных на выходе данных был сделан вывод о необходимости перехода к машинному обучению и выбора такого алгоритма, который позволил бы работать с «сырыми» данными и извлекать значимые признаки без участия исследователя. Подобные возможности предоставляют нейронные сети глубокого обучения.

В качестве основного инструмента анализа в настоящей работе были выбраны рекуррентные нейронные сети с модулями долго-краткосрочной памяти. В предобработке задействовано сингулярное разложение и линейный дискриминантный анализ.

Ставилась задача классификации ЭЭГ на 4 класса в зависимости от типа модуляций, происходящих в прослушиваемой испытуемыми музыкальной последовательности. На основе изучения специальной литературы, разработки и реализации практической части были сделаны следующие выводы: все модуляции испытуемые воспринимали сходным образом. Для дальнейшего развития работы существуют следующие пути:

1. Расстояния между классами малы, следовательно необходимо провести дополнительную предобработку, причем, сделать это таким образом, чтобы природа данных не претерпевала радикальных изменений.
2. Расширить датасет новыми объектами и снова подать его на вход описанной в настоящей работе сети.

Таким образом, цели и задачи, поставленные во введении, были достигнуты, поэтому выполненная курсовая работа является действительно актуальной и имеет практическое значение.

# СПИСОК ЛИТЕРАТУРЫ

# **ProjectProposal: СПИСОК ЛИТЕРАТУРЫ**

1. Т.А. Кучменко, Р.П. Лисицкая, А.А. Шуба, Информативность анализатора газов «электронный нос» для оценки качества вина. - Аналитика и контроль. - №4. – 2014.
2. Т.А. Кучменко, Д.А. Погребная, Сравнительная оценка возможностей интегрального и дифференциального анализаторов газа типа «электронный нос» для исследования мясных продуктов. - Аналитика и контроль. - №3. – 2011.
3. Е.В. Дроздова, Определение органических легколетучих токсикантов массивом пьезосенсоров для оценки безопасности полимерных материалов: диссертация ... кандидата химических наук: 02.00.02 / Дроздова Евгения Викторовна; [Место защиты: Воронеж. гос. ун-т]. - Воронеж, 2016. - 263 с.: ил
4. An integrated chemical sensor array using carbon black polymers and a standard cmos process, год??
5. Chemical Sensors for Electronic Nose Systems, 2004
6. Electronic noses and their applications, 1995
7. Using Sensor Arrays and Pattern Recognition to Identify Organic Compounds, 1989
8. Homeland Security, Olfactory Images, and Virtual Chemical Sensors
9. Data analysis for electronic nose systems, 2005
10. Intelligent Classification of Odor Data Using Neural Networks, 2012

# Statistics: Ссылки

1. <https://www.scipy.org/>

2. <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.normaltest.html>

3. D’Agostino, R. B. (1971), “An omnibus test of normality for moderate and large sample size,” Biometrika, 58, 341-348

4. <http://statsmodels.sourceforge.net/devel/generated/statsmodels.graphics.gofplots.qqplot.html>

5. <http://statsmodels.sourceforge.net/>

6. <http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.stattools.adfuller.html>

7. Box, G.E.P., and Jenkins, G., (1970) Time Series Analysis, Forecasting and Control, Holden-

Day, San Francisco.

8. Box, G.E.P., and Pierce, D.A., (1970) "Distribution of the Residual Autocorrelations in

Autoregressive-Integrated Moving-Average Time Series Models", Journal of the American

Statistical Association, 65, 1509-1526.

9. <http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.arima_model.ARIMA.html>

10. <http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.arima_model.ARIMA.fit.html>

**План**

1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ
   1. Принципы построения систем «Электронный нос»
   2. Обзор существующих систем и их приложений
   3. Описание данных и подходы к их обработке
2. МЕТОДЫ ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ ИСКУССТВЕННОГО ОБОНЯНИЯ «МАГ-8»
   1. Подходы к симуляции откликов массива сенсоров «МАГ-8»
   2. Исследование данных в задаче оценки безопасности полимерных материалов
   3. Применение алгоритмов машинного обучения
   4. Применение нейронных сетей

***В основной части*** магистерской диссертации должно быть систематизировано и полно изложено состояние вопроса, которому посвящена данная работа. Основная часть состоит, как правило, из двух – трех глав, каждая из которых делится на параграфы в зависимости от темы исследования и его целей. В каждой главе должно быть не менее двух параграфов. Объем каждого параграфа должен составлять не менее 6-7 страниц.

Основная часть работы состоит из теоретической (теоретико-методологической) и эмпирической (аналитической) составляющей.

В теоретической части отражается умение магистранта систематизировать существующие разработки и теории по выбранной проблеме, критически их рассматривать, выделять существенное, оценивать опыт других исследователей, определять главное в изученности темы с позиций современных подходов, аргументировать собственное мнение.

Поскольку магистерская диссертация обычно посвящается достаточно узкой теме, то обзор работ следует делать только по вопросам выбранной темы, а не по всей проблеме в целом. В нем должны быть названы, проанализированы и критически оценены наиболее известные и ценные публикации, имеющие непосредственное отношение к теме магистерской диссертации (с обязательным оформлением ссылок на источники). При изложении спорных вопросов необходимо приводить мнения различных авторов. После чего следует обосновать свое мнение или согласиться с одной из уже имеющихся точек зрения, однако в любом случае необходимо выдвинуть соответствующие аргументы.

*Практическая (аналитическая) часть* работы должна содержать описание объекта исследования, анализ изучаемой проблемы, анализ эмпирических (фактических) данных, выполненный с помощью современных методик и представленный в виде аналитических выкладок. Кроме того, должны быть приведены расчеты отдельных показателей, используемых в качестве характеристик объекта исследования.

В качестве эмпирической базы в магистерской диссертации могут выступать данные исследования, проведенного лично или при участии магистранта; российские и международные социологические и статистические исследования (например, РМЭЗ-ВШЭ, Европейское социальное исследование ESS, Комплексное наблюдение условий жизни Росстата и др.), микроданные которых имеются в свободном доступе; данные исследований, предоставленные магистранту научным руководителем.

Эмпирическая часть может завершаться *проектными предложениями,* которые представляют собой разработку рекомендаций и мероприятий по решению изучаемой проблемы (например, по совершенствованию практики адаптации, обучения, оплаты труда и т.д.), а также подтвержденный расчетами анализ результатов использования предложенных мер или обоснование предполагаемых результатов использования предложенных мер.