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](#_Toc478420751)

[**1.** **ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ** 6](#_Toc478420752)

[**1.1** **Архитектура и обзор систем «Электронный нос»** 6](#_Toc478420753)

[***1.1.1 Принципы построения систем «Электронный нос»*** 6](#_Toc478420754)

[***1.1.2 Примеры с систем «Электронный нос» и их приложения*** 9](#_Toc478420755)

[**Ссылки** 12](#_Toc478420756)

[**1.2 Обзор литературы в области анализа данных систем искусственного обоняния** 12](#_Toc478420757)

[**Ссылки** 15](#_Toc478420758)

[**2.** **МЕТОДЫ ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ ИСКУССТВЕННОГО ОБОНЯНИЯ «МАГ-8»** 16](#_Toc478420759)

[**2.1 Описание входных данных и формирование датасетов** 16](#_Toc478420760)

[***2.1.1 Источники данных*** 16](#_Toc478420761)

[***2.1.2 Исследование исходных данных*** 19](#_Toc478420762)

[***2.2 Предобработка*** 20](#_Toc478420763)

[**2.3 Выбор между сэмплированием и работой с исходными данными** 20](#_Toc478420764)

[***2.3.1 Добавление константы*** 20](#_Toc478420765)

[***2.2.2 Oversampling-алгоритмы*** 22](#_Toc478420766)

[***2.2.3 Результаты только на исходных данных*** 22](#_Toc478420767)

[**2.3 Применение алгоритмов машинного обучения и нейронных сетей** 22](#_Toc478420768)

[ЗАКЛЮЧЕНИЕ 22](#_Toc478420769)

[СПИСОК ЛИТЕРАТУРЫ 23](#_Toc478420770)

[**ПРИЛОЖЕНИЕ 1 «Некоторые коммерческие системы «Электронный нос»: модели и технологии».** 24](#_Toc478420771)

[**ОПИСАНИЕ МАГИСТЕРСКОЙ** 26](#_Toc478420772)

# 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», разработанной в Воронежском государственном университете инженерных технологий группой под руководством доктора химических наук, профессора Татьяны Анатольевны Кучменко. При написании работы основное внимание было уделено изучению специализированной литературы и ресурсов Интернета, список источников состоит из 31 пункта.

Система «электронный нос» получает все больше приложений при решении многих задач аналитической химии, от оценки качества пищевых продуктов до обнаружения запрещенных грузов и диагностики некоторых заболеваний. Благодаря развитию электроники, появляются портативные приборы, пригодные для экспресс-анализа, и существенным их преимуществом является низкая цена в сравнении с газовыми хроматографами.

Такие системы строятся на основе массива сенсоров нескольких видов, высокочувствительных к заданному набору соединений и веществ. 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. 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 набором веществ-маркеров, с которыми он планирует работать, исходя из этого подбирает селективные покрытия датчиков, 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 сенсоров 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 использовано решение задачи бинарной классификации для отдельных веществах и их смесей на этапе обучения и решение задачи классификации на N классов для новых объектов, которые всегда в своем «ароматическом отпечатке» будут содержать следы многих веществ в непредсказуемых концентрациях, часть из которых никогда не будет включена в обучающую выборку по причине ограниченности временных ресурсов исследователя и, как следствие, узости решаемой им задачи.

Новизна и актуальность данного исследования следуют из уникальности источника данных — прибора «МАГ-8». В данный момент группой разработчиков «МАГ-8» используется исключительно графический метод анализа. Данный подход успешно применяется при решении широкого круга задач – от анализа качества колбасных изделий до диагностики некоторых заболеваний. Недостатком данного подхода является трудоемкость и требование к высокой квалификации человека, который интерпретирует результаты. 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 расширить обучающее множество путем размножения наименее представленных объектов ИЛИ Подбор способа предобработки сравнительно небольшого и несбалансированного массива исходных данных
2. Подбор и обучение алгоритма бинарной классификации веществ-маркеров и их смесей
3. Решение задачи обнаружения веществ-маркеров и их смесей в пробах 75 детских игрушек

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 рассмотрению подходов к анализу данных прибора «МАГ-8» и построению прототипа системы анализа данных [] . В заключении подводятся итоги исследования и рассматриваются направления дальнейшего развития. Приложения содержат

# **ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ**

## **Архитектура и обзор систем «Электронный нос»**

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

Концепция «электронного носа” как инструмента, состоящего из датчиков, используемых для классификации запахов, была впервые введена Персо и Доддом в 1982 году [-1]. В ходе своих экспериментов эти два исследователя поставили перед собой цель создать инструмент, способный эмулировать обонятельную систему млекопитающих, распознавая различные запахи и давая воспроизводимые результаты. В частности, разработанный ими электронный нос состоял из: (I) матрицы датчиков для имитации рецепторов обонятельной системы человека; (II) блок обработки данных, который будет выполнять ту же функцию, что и обонятельная лампа; (III) система распознавания образов, которая распознавала бы обонятельные паттерны испытуемого вещества, подобно обонятельной системе человека [0].

Понятие «Электронный нос» было введено в 1988 году Гарднером и Бартлеттом, которые определили его как «инструмент, который содержит множество электронных химических датчиков с частичной специфичностью и соответствующей системой распознавания образов, способных распознавать простые или сложные запахи» [1].

В процессе развития систем «электронный нос» возникла перспектива создания портативных приборов, который позволили бы заменить газовую хроматографию в задачах, критичных к скорости распознавания летучих органических соединений. Одновременно с этим стали активно развиваться математические методы анализа данных, получаемых с этих устройств [2].

Системы «электронный нос» строятся на основе массива сенсоров, которые способны не только обнаруживать, но и идентифицировать летучие органические соединения, представляя их в виде уникальных комбинаций откликов всех сенсоров. Обычно для построения систем общего назначения используются высокоселективные сенсоры – то есть такие сенсоры, которые способы реагировать на многие классы органических веществ. Этот подход позволяет представлять большее число веществ с помощью меньшего числа сенсоров; в случае же применения низкоселективных сенсоров, их потребовалось бы ровно столько штук, сколько веществ необходимо обнаружить, что лишает устройство портативности и добавляет сложность при анализе. Управление селективностью осуществляется с помощью подбора пленок-сорбентов, наносимых на поверхность сенсоров. Структурная схема системы представлена на рисунке 1.

Система анализа данных

Преобразователь сигнала

Массив сенсоров

Газ\летучее органическое соединение

Рисунок 1 – Системы «Электронный нос»

Химические датчики обнаруживают молекулы запахов на основании реакции между молекулами запаха и целевыми чувствительными материалами на поверхности датчика, называемыми сорбентами. Эта реакция вызывает определенное изменение в массе, объеме или других физических свойствах. Затем это изменение преобразуется в электронный сигнал с помощью преобразователя. Существуют различные типы преобразователей для химических сенсоров: оптические, электрохимические, термочувствительные и чувствительные к массе. Далее будут рассмотрены некоторые из наиболее распространенных химических сенсоров: датчик поверхностных акустических волн (surface acoustic wave, SAW), датчик микробаланса кварцевого кристалла (quartz crystal microbalance sensor, QSM), полупроводниковый датчик на основе оксида металла (metal oxide semiconductor sensor, MOX) и полимерный композитный датчик (polymer composite-based sensor) [3].

Преобразователь датчика поверхностных акустических волн (SAW) чувствителен к массе. Датчик состоит из подложки из кварца и химически чувствительной тонкой пленки. Поскольку кварц является пьезоэлектрическим материалом, он преобразует поверхностные акустические волны в электрические сигналы. Когда химически чувствительная тонкая пленка адсорбирует определенные молекулы, масса пленки увеличивается, что приводит к более медленному перемещению акустических волн. Это изменение может быть обнаружено сенсорной микроэлектроникой, как только акустическая волна преобразуется в электрический сигнал. Существенным ограничением таких сенсоров является необходимость контролировать температуру окружающей среды

Датчик микробаланса кварцевого кристалла (QCM) представляет еще один тип датчиков, в основе которых лежит микровзвешивание. Подобно датчику SAW, преобразователь датчика QCM также чувствителен к массе. Основное различие между SAW и QCM заключается в том, что первый использует датчик поверхностных акустических волн, а второй использует датчик объемных акустических волн. Его чувствительный механизм основан на сдвиге резонансной частоты кварцевого резонатора (QC) за счет адсорбции газовых молекул на поверхности чувствительной пленки [3]. Приборы на основе этих датчиков – одни из самых часто встречающихся на рынке в среднем ценовом сегменте.

Датчики на основе оксидов металлов представляют собой принципиально новое устройство, которые переводит изменение концентрации паров химических веществ в электрические сигналы. На поверхности датчика находится чувствительный металл-оксидный полупроводник с изолирующим слоем под ним, нагреватель и схема для измерения сопротивления. Когда молекулы летучих органических соединений собираются на поверхности оксида металла, начинается процесс их окисление при повышенной температуре, обычно от 250 до 450 ° С. Реакция приводит к переносу электронов от молекул соединения к структуре оксида металла. В результате, регистрируется изменение проводимости. Эти датчики существенно отличаются от кварцевых, поскольку, благодаря такой высокой температуре реакции, индифферентны к изменениям температуры окружающей среды [3]. Однако, это преимущество так же является и недостатком – повышение рабочей температуры ведет к повышению энергопотребления, что является критичных параметром для портативных приборов. Существенные усилия были направлены на устранения этого недостатка, и в результате появился ряд работ, которые предлагают значительные улучшения существующих систем на основе MOX [4]. В результате, приборы на основе таких датчиков стали очень популярны на зарубежном рынке.

### ***1.1.2 Примеры с систем «Электронный нос» и их приложения***

Благодаря способности «электронного носа» различать и распознавать множество различных газов и запахов, используя только небольшое количество датчиков, и первым многообещающим результатам, полученным в результате исследований в этой области, к этому предмету возник огромный интерес в научном сообществе и за его пределами.

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), Cyranose-320 (Америка). 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». Далее будут рассмотрены подробнее некоторые наиболее коммерчески успешные продукты. Более полный перечень существующих систем см. в Приложении 1 «Некоторые коммерческие системы «Электронный нос»: модели и технологии» [5-5].

Система E-nose KAMINA была разработана Гошником [5] и коммерциализирована Cистемами и службами химического анализа (Systems and Services for Chemical Analysis, SYSCA). Система работает на уникальной микросхеме, состоящей из 38 градиентных датчиков окиси олова (SnO2) и вольфрама (WO3), и помещается в ладони, так как её размеры примерно совпадают с размерами стандартной рации. При более высоких температурах (около 300° C) электропроводность оксида металла зависит от состава окружающего газа и, следовательно, может быть использована для обнаружения газов. Кроме того, каждый датчик нагревается до разных температур, а его толщина мембраны отличается от толщины его соседних датчиков. Как следствие, отдельные датчики имеют различный спектр чувствительности к газам, и все датчики имеют различную реакцию на один газ. Таким образом достигается высокая точность обнаружения газов и их смесей.

Другой коммерческой системой E-nose является система Cyranose 320 E-nose от Smiths Detection [6]. Это портативная система, состоящая из массива из 32 химических датчиков, пробоотборника ЛОС и встроенного системы обработки данных. Она обнаруживает и идентифицирует летучие органические соединения (ЛОС) на основе изменения электрического сопротивления за счет поглощения ЛОС. Сенсорная поверхность представляет собой полимерную сетку с трехмерной непрерывной пористой структурой, заполненной проводящим углеродом. Когда молекулы ЛОС попадают на чувствительную поверхность, реакция между молекулами ЛОС и функциональной группой (группами) в полимерах вызывает расширение объема в полимерной сети. Как следствие, связь между блоками углерода, заполняющими полимерную сетчатую структуру, становится рыхлой, а электропроводность уменьшается. Тип и плотность функциональной группы (групп) в макромолекулах адаптированы для каждого типа датчика, с тем чтобы каждый датчик реагировал на разные ЛОС по-разному.

Система обработки данных Cyranose-320 протестирована в [3], где показано, что точность детектирования резко уменьшается, когда количество компонентов в смеси превышает 3. Это показывает, что данная система не обладает достаточной способностью идентифицировать сложные смеси летучих органических соединений. Однако, идентификация отдельных ЛОС довольно успешна при условии, что концентрация известных веществ в новых образцах ниже, чем концентрация в объектах тренировочной выборки. Отличные результаты показал данный прибор в задаче классификации бактерий [7]

Alpha-MOS (Тулуза, Франция) Fox electronic nose был разработан в сотрудничестве с университетами Warwich и Southampton. Он использует шесть (Fox 2000), 12 (Fox 3000), либо 18 (Fox 4000) металлоксидных газовых сенсоров и может использоваться как впрыска газов из внешних баллонов, так и с штатным внутренним насосом и контроллером расхода масс.

Российская разработка - прибор «МАГ-8» - содержит 8 пьезокварцевых сенсоров, разработке прототипа интегрированной системы обработки и анализа данных для этого прибора посвящена данная работа. Существующий подход, прошедший успешную апробацию на ряде типичных для таких систем задач, описан далее. Основные выводы, сделанные в результате ряда исследований, таковы: «электронный нос» «МАГ-8» превосходит физико-химические показатели в задаче оценки органолептических характеристик вина [8]. Сравнительный анализ возможностей интегрального анализатора газа “VOCmeter” (Германия) и дифференциального анализатора «МАГ-8» приведен в [9] 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.

Таким образом, первоначальные исследования были направлены на применение систем «электронный нос» в пищевой и косметической промышленности, где они в настоящее время по-прежнему широко используются для оценки качества продуктов питания, контроля вкусовых характеристик и качественного ранжирования сортов вин и пива [10]. Системы «Электронный нос» используются также в экологическом мониторинге для идентификации токсичных отходов, выявления опасных химических веществ в грунтовых водах и мониторинга качества воздуха и промышленных выбросов. В последнее время достигнут прогресс также в применении к мониторингу здоровья и медицинской диагностике. Развитие методов обработки данных шло параллельно с развитием технологической составляющей; рассмотрению эволюции подходов, с помощью которых удалось достичь настолько широкого круга решаемых задач, посвящен параграф 1.2 настоящего исследования.

**Ссылки:**

[-1] Persaud K., Dodd G., Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. - Nature. - №282. - 1982. - p. 352–355.

[0] Gardner, J.W., Bartlett, P.N., Electronic Noses: Principles and Applications. - Oxford University Press: New York. - NY, USA. - 1999.

**[1]** Gardner J.W.; Bartlett, P.N. A brief history of electronic noses. Sens. Actuat. B: Chem. **1994**, 18, 211-220.

[2] Hoffheins, B. Using Sensor Arrays and Pattern Recognition to Identify Organic Compounds // M.Sc. Thesis. - University of Tennessee. – Knoxville. - TX, USA. - 1989.

(ЕСТЬ ДУБЛЬ В 1.2)

[3] Li S., Overview of Odor Detection Instrumentation and the Potential for Human Odor Detection in Air Matrices, - 2009.

[4] Elmi I., Zampolli S., Cozzani E., Mancarella F. and Cardinali G. C., Development of  
ultra-low-power consumption MOX sensors with ppb-level VOC detection capabilities  
for emerging applications. - Sensors and Actuators. – 2008.

[5-5] Wilson A.D., Baietto M. Applications and Advances in Electronic-Nose Technologies. - Sensors. - №9. - 2009.

[5] Arnold C., Haeringer D., Kiselev I. and Goschnick J., Sub-surface probe module  
equipped with the Karlsruhe Micronose KAMINA using a hierarchical LDA for the  
recognition of volatile soil pollutants. - Sensors and Actuators. – pp. 90-94. – 2006.

[6] The Cyranose 320 E-nose User Manual, Smiths Detection // User Manual.

[7] Dutta R., Hines E. L. Gardner J.W. and Boilo P. t, Bacteria classification using Cyranose 320 electronic nose, - Bio Med Central Ltd., - 2002.

[8] Кучменко Т.А., Лисицкая Р.П., Шуба А.А., Информативность анализатора газов «электронный нос» для оценки качества вина. - Аналитика и контроль. - №4. – 2014.

[9] Кучменко Т.А., Погребная Д.А., Сравнительная оценка возможностей интегрального и дифференциального анализаторов газа типа «электронный нос» для исследования мясных продуктов. - Аналитика и контроль. - №3. – 2011.

[10] Rolfe B., Toward Nanometer-Scale Sensing Systems: Natural and Artificial Noses as Models for Ultra-Small, Ultra-Dense Sensing Systems // Nanosystems Group, The MITRE Corporation, - McLean, Virginia, - 2004.

## **1.2 Обзор литературы в области анализа данных систем искусственного обоняния**

Работа [1] обобщает все существующие подходы к анализу данных в задаче распознавания паров химических веществ и летучих органических соединений: 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 несколько формул нормализации, рассматриваются способы отбора признаков. В целом, данная работа скорее обобщает существующие подходы, чем привносит что-то принципиально новое. С момента её написание прошло 11 лет, однако она не перестаёт быть актуальным источником, который позволяет охватить широкий набор методов.

Применение нейронных сетей описал Хоффхайнс в 1989 году в своей работе [2]. Он показал, что, благодаря использованию массивов сенсоров, нейронные сети успешно решают задачу распознавания паров летучих органических 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 следующее ограничение: сеть не способна распознавать неизвестные ей смеси веществ, присутствовавших в обучающей выборке.

Большое количество работ [3][4] 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.

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 [5] 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. Однако, для задачи распознавания летучих органических соединений такой объем выборки достаточно велик.

Еще один успешный пример применения системы «электронный нос» описан в [7]: 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.

Недостатком предыдущих работ по анализу данных является то, что они не освещают возможности многослойных нейронных сетей. Статья [8] 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).

В работе [9] показана комбинация графического метода (полярные диаграммы) и дендрограмм с расстояние Чебышева для выбора сенсоров в задаче классификации сортов сыра, сделан вывод о том, что для устойчивого различения веществ, характеризующих сорта, достаточно выбрать уникальную пару из набора сенсоров.

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. 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 [10] 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».

**Ссылки**:

[1] Scott, S., James, D. & Ali, Z, Data analysis for electronic nose systems, - Microchim Acta ,-2006.

[2] Hoffheins, B. Using Sensor Arrays and Pattern Recognition to Identify Organic Compounds // M.Sc. Thesis. - University of Tennessee. – Knoxville. - TX, USA. - 1989.

(ДУБЛЬ в 1.1!)

[3] J.A. Dickson, et al., “An Integrated Chemical Sensor Arrays Using Carbon Black Polymers and a Standard CMOS Process”, Proc. Solid-State Sensors and Actuators Workshop, Hilton Head Island, SC, June 2000, pp. 162-165.

[4] James, D., Scott, S.M., Zulfiqur, A., O'Hare, W.T., 2005. Chemical sensors for electronic nose systems. Microchimica Acta 149, 1-17.

[5] Keller P.E., Kangas L.J., Liden L.H., Hashem S., Kouzes R.T., Electronic noses and their applications // Proceedings of the IEEE Technical Applications Conference (TAC’95) at Northcon. - Portland, Oregon, - 10–12 October, 1995.

[6]

[7] Staples E.J., Viswanathan S., Homeland security, olfactory images, and virtual chemical sensors // Proceedings of the AIChE Annual Meeting. - pp. 41-49. - 2004.

[8] Omatu S., Araki H., Fujinaka T., Yano M., Intelligent Classification of Odor Data Using Neural Networks, // ADVCOMP 2012 : The Sixth International Conference on Advanced Engineering Computing and Applications in Sciences. - 2012

[9] Pais, V. P, Oliveira J .A. B. P, Gomes M. T. S.R., An Electronic Nose Based on Coated Piezoelectric Quartz Crystals to Certify Ewes’ Cheese and to Discriminate between Cheese Varieties, - Sensors (Basel) 2012; 12(2): 1422–1436. Published online 2012 Feb 1.

[10] 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

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

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

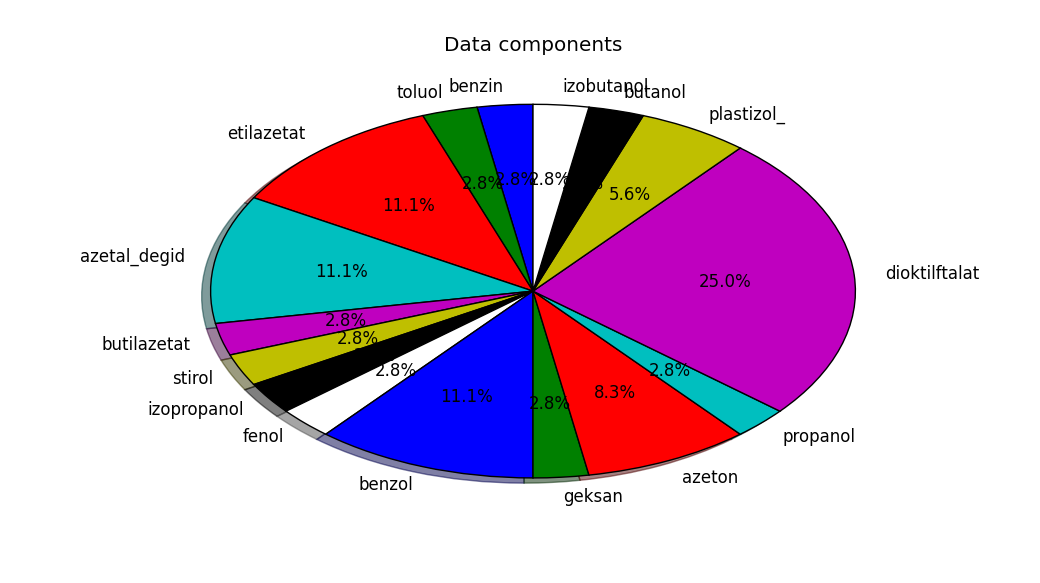
### ***2.1.1 Источники данных***

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.

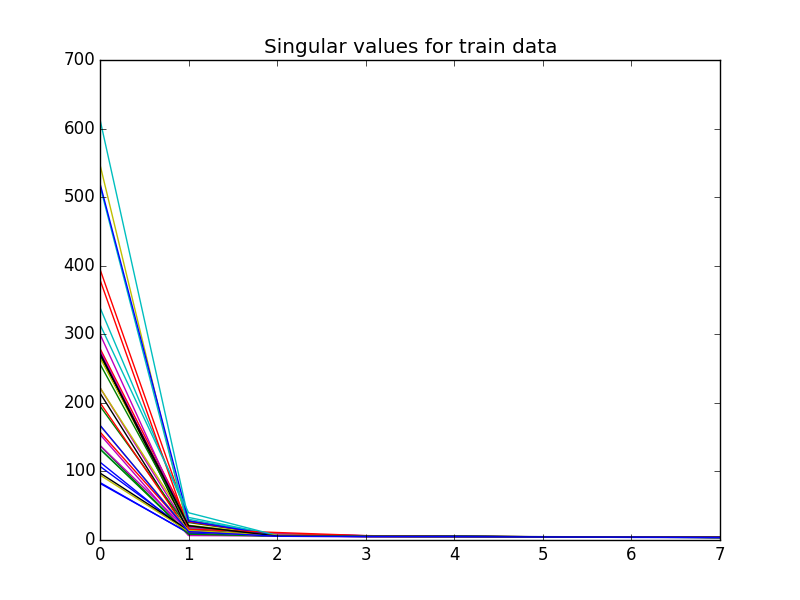
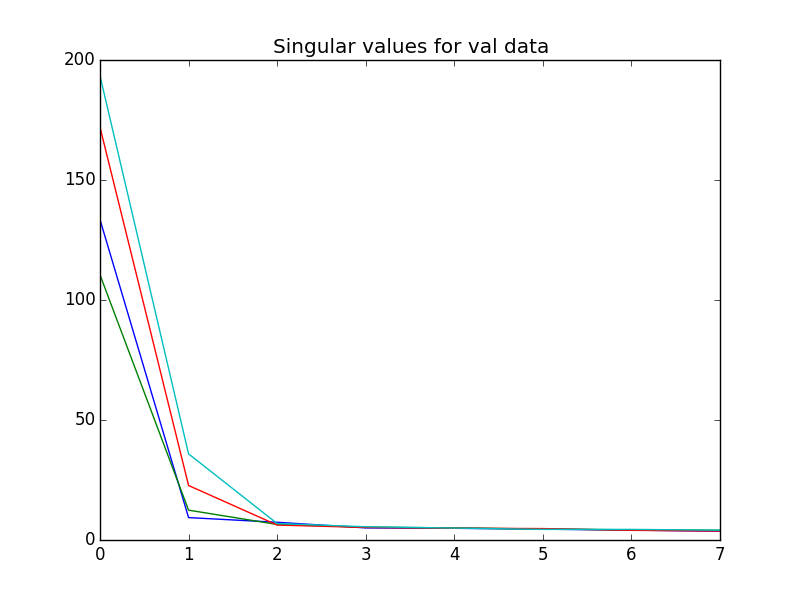
База данных содержит:

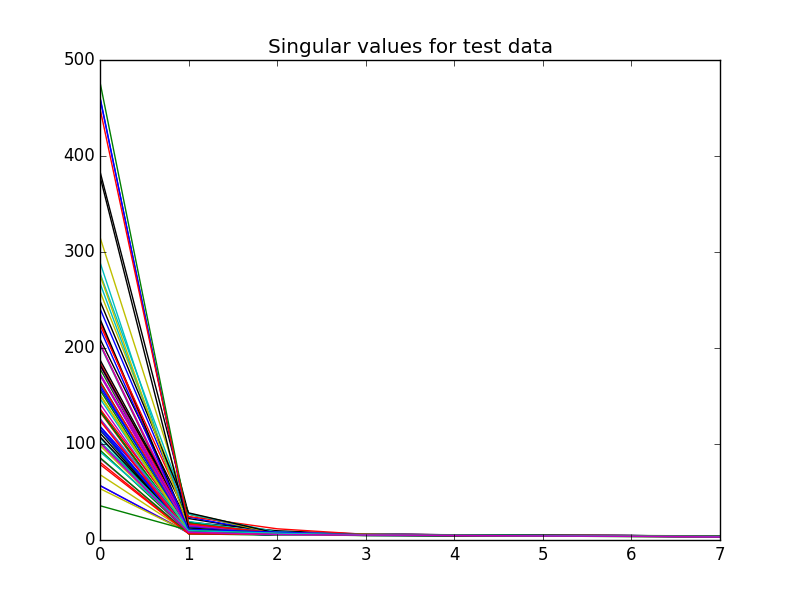
* 8 таблиц: Data, GroupTree, Mask, MaskData, MeasureProfile, MeasureProfileData, Measures, Sensors;
* 430 записей измерений и прочистки сенсоров;
* 250 различных объектов;
* 103 объекта длины 120 (объекты были отфильтрованы по полю FullLength как по самому критичному, дальнейший анализ извлеченных измерений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

Нормализация, удаление тренда, scaling

## **2.3 Выбор между сэмплированием и работой с исходными данными**

### ***2.3.1 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.

### ***2.2.2 Oversampling-алгоритмы***

### ***2.2.3 Результаты только на исходных данных***

## **2.3 Применение алгоритмов машинного обучения и нейронных сетей**

1. Сравнение SVM, KNN, RF, Bagging

2. Нейросети LSTM,

3. Изображения - CNN

# 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

# **ПРИЛОЖЕНИЕ 1 «Некоторые коммерческие системы «Электронный нос»: модели и технологии».**

|  |  |  |  |
| --- | --- | --- | --- |
| **Тип** | **Производитель** | **Модели** | **Технология** |
| Моно-технология  (только «электронный нос») | Airsense Analytics | i-Pen, PEN2, PEN3 | MOS sensors |
| Alpha MOS | FOX 2000, 3000, 4000 | MOS sensors |
| Applied Sensor | Air quality module | MOS sensors |
| Chemsensing | ChemSensing Sensor array | Colorimetric optical |
| CogniScent Inc. | ScenTrak | Dye polymer sensors |
| Dr. Födisch AG | OMD 98, 1.10 | Receptor-based array |
| Forschungszentrum Karlsruhe | SAGAS | MOS sensors |
| Gerstel GmbH Co. | QSC | SAW sensors |
| GSG Mess- und Analysengeräte | MOSES II | Modular gas sensors |
| Illumina Inc. | oNose | Fluorescence optical |
| Microsensor Systems Inc | Hazmatcad, Fuel Sniffer, SAW MiniCAD mk II | SAW sensors |
| Osmetech Plc | Aromascan A32S | Conducting polymers |
| Sacmi | EOS 835, Ambiente | Gas sensor array |
| Scensive Technol. | Bloodhound ST214 | Conducting polymers |
| Smiths Group plc | Cyranose 320 | Carbon black-polymers |
| Sysca AG | Artinose | MOS sensors |
| Technobiochip | LibraNose 2.1 | QMB sensors |
| Комбинированная технология  (“электронный нос” + другие типы) | Airsense Analytics | GDA 2 | MOS, EC, IMS, PID |
| Alpha MOS | RQ Box, Prometheus | MOS, EC, PID, MS |
| Electronic Sensor Technology | ZNose 4200, 4300, 7100 | SAW, GC |
| Microsensor Syst. | Hazmatcad Plus CW Sentry 3G | SAW, EC |
| Rae Systems | Area RAE monitor IAQRAE | Thermistor, EC, PID, CO2, humidity |
| RST Rostock | FF2, GFD1 | MOS, QMB, SAW |

# **ОПИСАНИЕ МАГИСТЕРСКОЙ**

***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