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

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 35 пунктов.

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 определяет набор веществ-маркеров, с которыми он планирует работать, затем подбирает селективные покрытия датчиков, 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 классификации на N классов на отдельных веществах и их смесях на этапе обучения и решение задачи классификации на N классов с назначением многих меток 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 специалиста, 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. Выбор из двух альтернатив: 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 алгоритмов классификации веществ-маркеров и их смесей
3. Решение задачи обнаружения веществ-маркеров и их смесей в ароматических отпечатках 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 главе 1 сделан обзор предметной области – рассматривается история развития и принципы работы систем «Электронный нос», приводится обзор рынка этих систем. Глава 1 также содержит обзор литературы по методам анализа данных для систем «Электронный нос», начиная с конца 80х годов и до настоящего времени. Вторая глава посвящена рассмотрению подходов к анализу данных прибора «МАГ-8» и построению прототипа системы анализа данных – рассматриваются основным методы предварительной обработки, методы расширения массива данных искусственными объектами, приводятся сравнительные таблицы производительности алгоритмов машинного обучения и нейронных сетей. В заключении подводятся итоги исследования и рассматриваются направления дальнейшего развития. Приложения содержат сравнительные таблицы приборов «Электронный нос» и результаты работы алгоритмов, а также графики архитектур нейронных сетей.

1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Архитектура и обзор систем «Электронный нос»

1.1.1 Принципы построения систем «Электронный нос»

Концепция «электронного носа» (Electronic nose, e-nose) как массива датчиков, используемых для классификации запахов, была впервые введена Персо и Доддом в 1982 году [1]. Целью их исследований было создание инструмента, способного эмулировать обонятельную систему млекопитающих, распознавая различные запахи и давая воспроизводимые и интерпретируемые реакции. В частности, разработанный ими электронный нос состоял из: массива датчиков для имитации рецепторов обонятельной системы человека; блока обработки данных, который выполнял ту же функцию, что и подкорковые центры головного мозга человека; системы распознавания образов, которая распознавала бы ароматические отпечатки вещества подобно корковому центру обоняния головного мозга человека, располагающемуся в височной доле [2].

Понятие «Электронный нос» было сформулировано позже, 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» [3].

В процессе развития систем «электронный нос» возникла необходимость создания портативных приборов, который позволили бы заменить газовую хроматографию в задачах, критичных к скорости распознавания летучих органических соединений и паров химических веществ. Одновременно с этим стали активно развиваться математические методы анализа данных, получаемых с этих устройств [4].

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 низкоселективные 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 высокоселективных сенсоров, их потребовалось бы ровно столько штук, сколько веществ необходимо обнаружить, что лишает устройство портативности и универсальности. Управление селективностью осуществляется с помощью подбора пленок-сорбентов, наносимых на поверхность сенсоров. Структурная схема системы представлена на рисунке 1.

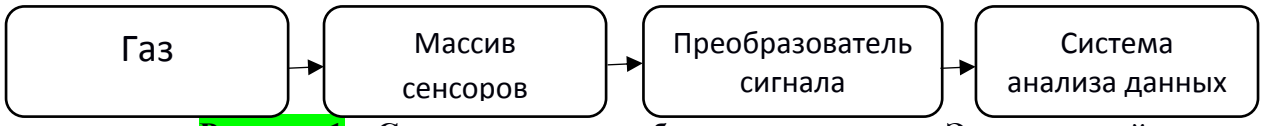


Рисунок 1 – Схематическое изображение системы «Электронный нос»

Датчики регистрируют присутствие молекул запахов на основании реакции между ними и чувствительными материалами на поверхности датчика, называемыми сорбентами. Эта реакция вызывает определенные изменения в массе, объеме или других физических характеристиках плёнки, 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 датчиков: 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) [5].

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 пленка адсорбирует определенные молекулы, масса пленки увеличивается, что приводит к более медленному перемещению акустических волн. Это изменение регистрируется микроэлектроникой, которая преобразует акустическую волну в электрический сигнал. Существенным ограничением таких сенсоров является необходимость соблюдения температурного режима и режима влажности 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. Принцип работы преобразователя QCM основан на сдвиге резонансной частоты кварцевого резонатора (QC) во время адсорбции газовых молекул на поверхности чувствительной пленки [5].

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 (газов) собираются на поверхности оксида металла, начинается окислительная реакция с повышением температуры, обычно в интервале от 250 до 450° С. Реакция приводит к переносу электронов от молекул газа на оксид металла. В результате, регистрируется изменение проводимости. Эти датчики существенно отличаются от кварцевых, поскольку, благодаря такой высокой температуре реакции, индифферентны к изменениям температуры окружающей среды [5]. 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 [6]. В результате, приборы на основе таких датчиков стали очень популярны на зарубежном рынке.

1.1.2 Примеры систем «Электронный нос» и их приложения

Благодаря способности «электронного носа» различать и распознавать множество различных запахов с помощью небольшого количества датчиков, а также благодаря первым многообещающим результатам 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» [7].

Система E-nose KAMINA была разработана Гошником [8] и коммерциализирована Системами и службами химического анализа (Systems and Services for Chemical Analysis, SYSCA). Система работает на микросхеме, состоящей из 38 градиентных датчиков окиси олова (SnO2) и вольфрама (WO3), и её размеры примерно совпадают с размерами стандартной рации.

Другой коммерческой системой E-nose является система Cyranose 320 E-nose от Smiths Detection [9]. Это портативная система, состоящая из массива из 32 химических датчиков, пробоотборника газов и встроенной 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) полимеров вызывает расширение объема 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 [5], где показано, что точность детектирования резко уменьшается, когда количество компонентов в смеси превышает 3. Следовательно, 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 [10].

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

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. «Электронный нос» «МАГ-8» превосходит физико-химические показатели в задаче оценки органолептических характеристик вина [11]. Сравнительный анализ возможностей интегрального анализатора газа “VOCmeter” (Германия) и дифференциального анализатора «МАГ-8» описан в [12] на примере задачи количественной и качественной оценки легколетучей фракции ароматических добавок для мясного сырья. Сделан вывод о том, что результаты, 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 [13]. 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.

1.2 Обзор литературы в области анализа данных систем искусственного обоняния

Работа [14] 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. С момента написания данной работы прошло 11 лет, однако она не перестаёт быть актуальным источником, который позволяет охватить широкий набор методов.

Применение нейронных сетей в задаче анализа данных с массивов «обонятельных» сенсоров описал Хоффхайнс в 1989 году в [4]. Он показал, что, благодаря использованию массивов нескольких 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 по подбору массива сенсоров и заложила основу для дальнейших исследований архитектур нейронных сетей и способов представления входных данных. В работе показано, что худший результат показали сети Хопфилда, вероятно, из-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 концентраций. 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, даже если компоненты этой смеси в обучающей выборке присутствовали.

Большое количество работ [15][16] посвящено подбору массивов сенсоров для решения определенных задач, что говорит о специфичности этих сенсоров и позволяет предположить, что обучение универсального алгоритма классификации, который успешно работал бы с разными видами датчиков, невозможно – для решения каждой отдельной задачи он должен обучаться индивидуально.

Интегрированные системы, состоящие из массива датчиков и автоматизированной системы распознавания, нашли применение в медицине, охране окружающей среды и пищевой промышленности. В работе Келлера [17] 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. Многоклассовая классификация и задача назначения многих меток классов в данной работе не ставилась.

Еще один успешный пример применения системы «электронный нос» описан в [18]: 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 такого устройства.

Недостатком предыдущих работ по анализу данных является то, что они не освещают возможности многослойных нейронных сетей, поскольку написаны ранее. Статья [19] 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 [20] показана комбинация графического метода (полярные диаграммы) и дендрограмм с расстоянием 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.

В более новых работах много внимания уделяется отбору признаков, поскольку крайне важно извлекать полезную и надежную информацию из характеристического отклика сенсоров, избегая избыточности. Исчерпывающий обзор и сравнение современных методов сделан в [21]: представлена классификация методов отбора признаков (извлечение признаков из оригинальных кривых, из методов сглаживания кривых, из преобразования – FFT, CWT и т.д., факторный анализ), сравниваются различные алгоритмы машинного обучения и исследуется зависимость точности классификации от архитектуры нейронной сети, рассматриваются границы применимости методов. Авторы делают вывод о важности нормализации как таковой и правильности выбора метода, говорят о том, что характеристики переходных процессов несут больше информации, чем стационарные, и интегральные характеристики обычно дают лучшую производительность, чем максимальные значения. Кроме того, DWT обычно показывает лучшие результаты, чем любые другие методы преобразования. Авторы обращают особое внимание на то, что из-за различий в селективности, чувствительности и специфичности датчиков оптимальные характеристики

являются строго индивидуальными для каждого прибора и иногда даже для задачи из-за различного протекания процессов сорбции для разных сорбентов. Более того, внутри одной задачи показания датчиков неоднородны, и поэтому выбранные методы могут быть не оптимальны для каждого датчика в отдельности.

Так, в [22] представлен принципиально новый подход к распознаванию летучих органических соединений и газов с использованием интегрированного нейрогенетического алгоритма классификации (NGCA). Предложенная авторами процедура распознавания состоит из сбора данных с массива датчиков, предварительной обработки сигналов алгоритмом скользящего среднего (SMMA) и последующего запуска NGCA. Предварительная обработка обеспечивает сглаживание данных, фильтрацию и устранение шума, а также извлечение вариаций паттернов. На следующем этапе работает NGCA – результат интеграции генетических алгоритмов (GA) и искусственных нейронных сетей (ANN). Сначала процедура GA производит отбор поколений признаков, которые подаются на вход ANN, которая обучается с помощью алгоритма обратного распространения ошибки. Эксперименты показывают, что предложенный авторами NGCA даёт лучшую производительность по сравнению с классическими генетическими алгоритмами (GA) и искусственными нейронными сетями (ANN). Так, отдельно ANN работают с точностью 82%, отдельно GA – 91%, только ANN-процедура NCGA – 92%, только GA-процедура NCGA – 72%, и, наконец, вся система NGCA – 95%.

Однако обычно массивы данных, используемые в рутинных задачах оценки качества пищевых продуктов, напитков и предметов быта, чрезвычайно малы для работы с такими мощными инструментами, как нейронные сети. Чтобы решить эту проблему, в последние годы стали развиваться подходы к разработке методов генерации массивов данных. Совершенно очевидно, что опираться такие методы должны на понимание физико-химических процессов, происходящих на границе раздела сред – газа

и твердых плёнок сорбентов. Изучением этих процессов занимается коллоидная химия, и, следовательно, исследовательская работа теперь выходит за рамки только области обработки данных или химико-инженерной задачи выбора массива сорбентов – теперь это построение математической модели на пересечении химии и анализа данных. Значительного успеха в этом достигли в университете Каталонии: на основе однокомпонентной модели сорбции Ленгмюра группой исследователей был создан пакет для R chemosensors, который моделирует работу сенсоров, начиная с модели сорбции и заканчивая моделированием шума, обязательно возникающего в реальных электрических системах [23]. В дальнейшем эта работа вылилась в разработку бенчмарков в области распознавания летучих органических соединений и газов [24].

Основной подход, применяемый в данный момент профессором Кучменко для работы с данных “МАГ-8”, 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 [25] 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».

2. МЕТОДЫ ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ ИСКУССТВЕННОГО ОБОНЯНИЯ «МАГ-8»

2.1 Описание входных данных и формирование массивов данных

2.1.1 Источники данных

Для начального исследования были получены матрицы откликов сенсоров «МАГ-8» на набор веществ-маркеров и смесей на основе диоктилфталата – эти объекты, называемые ароматическими отпечатками, составляют обучающее множество (36 отпечатков веществ и 4 смесей). Кроме того, предоставлено 75 новых объектов – это матрицы откликов сенсоров на пробы, взятые с детских игрушек, изготовленных из полимерных материалов. 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 121 x 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=[-1,121])

Для обучения получены следующие вещества (см. рисунок 2):

- Диоктилфталат (ДОФ) – 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.;
- ДОФ с ацетоном, ДОФ с ацетальдегидом, ДОФ с бензолом, ДОФ с этилацетатом – 1 шт.;

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 обучающего множества 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 сохранить первый элемент.

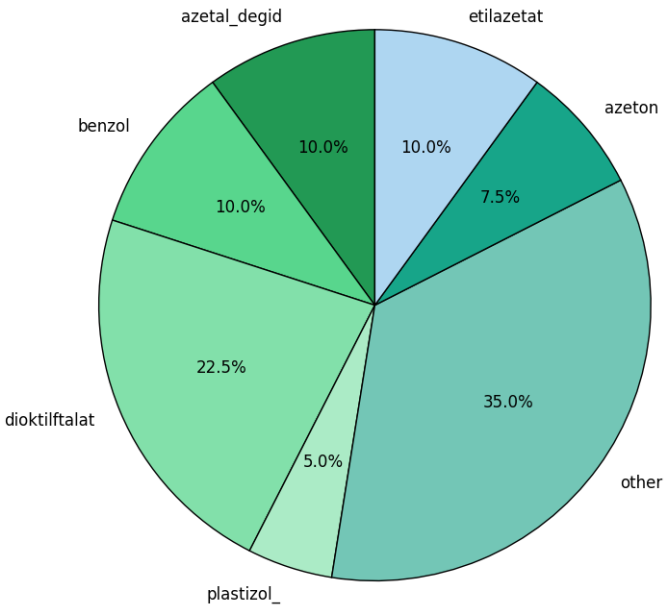


Рисунок 2 – Состав тренировочного множества

В процессе формирования массивов данных выявлены и решены некоторые проблемы. Все проблемы можно разделить на две группы: те, решение которых можно автоматизировать, и те, которые требуют более сложного решения.

Первая среди технически решаемых проблем – опечатки в названия файлов, которые мешают автоматическому извлечению меток классов – всё, что делается вручную, подвержено ошибкам, поэтому при необходимости

нужно просматривать и исправлять названия файлов вручную. Вторая проблема с названиями файлов состоит в том, что названия написаны по-русски, а автоматическая бинаризация меток в Sklearn не работает с кириллицей – решением этой проблемы является добавление метода, который выполняет примитивную транслитерацию по принципу ближайшей соответствующей буквы. Третья проблема – не все файлы содержат матрицу подходящего размера, некоторые – только каждый двадцатый отсчёт. Для решения этой проблемы в функцию чтения включена проверка соответствия матрицы заданной размерности.

Более сложная проблема, которая решается неочевидно, состоит в том, что массив данных слишком маленький и несбалансированный (см.рисунок 2). Это значит, в действительности, что качественное обучение с учителем на таких данных невозможно, поскольку алгоритмам свойственно ошибаться и предсказывать метки для новых данных в соответствии с их распределением – более классы, которые имеют большую долу в тренировочной выборке, будут иметь большую вероятность. Кроме того, поскольку некоторые классы присутствуют в единственном экземпляре, то невозможно разбить выборку случайным образом так, чтобы можно было обучить и проверить алгоритмы на всех существующих классах, то есть автоматический подбор признаков по методу grid-search также осложняется. Решений этой проблемы было три: поработать с базой данных sniffdb.sdf, которая содержит все эксперименты, которые проводили разработчики «МАГ-8», пойти по пути расширения набора искусственными данными или попытаться заставить алгоритмы научиться надежно различать классы (возможно, с помощью более сложных методов отбора признаков).

Первый способ решения проблемы рассмотрим ниже, остальные два вынесены в отдельные параграфы и будут рассмотрены далее. Предполагалось, что из базы данных sniffdb.sdf, содержащей все эксперименты, которые проводила кафедра, удастся извлечь дополнительные данные и таким образом расширить классы веществ и смесей. Для обучения в

рамках одной задачи релевантны данные одного эксперимента, снятые в один день, поскольку поведение кривых ΔF отражает реакцию, происходящую на поверхности сорбента, и, следовательно, напрямую зависит от используемых сорбентов и исследуемых соединений. Кроме того, пьезокварцевые датчики чувствительны к изменения влажности и температуры окружающей среды, следовательно, для сбора корректных данных необходимо стабильно воспроизводить эти условия.

Файл базы данных создан в Microsoft SQL Server Compact Edition, что является серьезным недостатком, так как это устаревший формат, несовместимый с прочими инструментами Microsoft для работы с базами данных. Для обеспечения кроссплатформенного доступа к базе с помощью инструмента SDF Viewer была конвертирована в sql-файл, который, в свою очередь был скорректирован для работы с mysql5.5, и для дальнейшей автоматизации работы с БД был написан скрипт на Python.

В процессе анализа извлекаемой информации было обнаружено, что в таблице Data хранятся не изменения частот колебаний сенсоров, а значения частот. Вычисление необходимых матриц ΔF показало, что частоты изменяются «ступенькой», что отличается от уже имеющихся данных из XLS, где они изменяются плавно. Возможная причина состоит в том, что система «МАГ-8» 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 и предварительная обработка исходных данных

Пример исходных данных представлен на рисунке 3 – это график колебаний частоты сенсоров при замере пробы ацетальдегида. На рисунке 3 слева видно, насколько «шумный» полученный сигнал – можно предположить, что причина появления такого шума в 1 Гц состоит в каких-либо инженерных несовершенствах устройства, однако похоже, что эти колебания описывают движение молекул в процессе сорбции – когда, согласно однослойной модели Ленгмюра, молекулы сначала занимают сорбционный

центр (стабильный рост частоты колебаний), но при этом одни молекулы уверенно закрепляются на пленке, в то время как другие имеют более слабую связь, вследствие чего улетают. Затем, когда вся поверхность пленки оказывается занята, молекулы налетают на слой других молекул – частота колебаний все еще растет, однако, согласно модели, эти молекулы не могут закрепиться. Таким образом, мы имеем отражение поведения только тех молекул, которые образовали единственный слой на сорбенте – именно они в итоге будут влиять на то, какие вещества будут обнаружены. Справа на рисунке 3 график ацетальдегида после сглаживания полиномом 3 степени, который выступал в качестве простого фильтра, позволяющего убрать скачки в 1 Гц, и увидеть общий вид графиков.

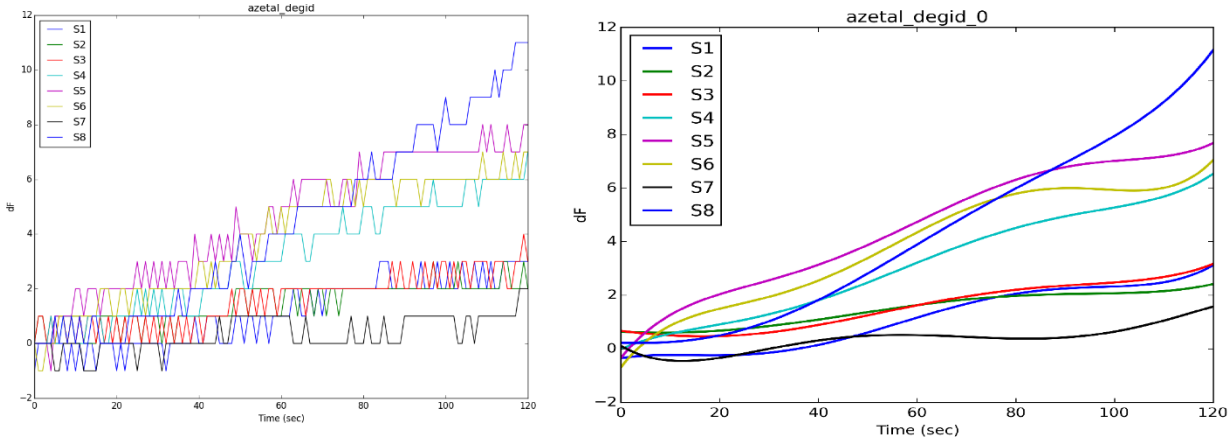


Рисунок 3 – График изменений колебаний частоты сенсора для ацетальдегида

В качестве начального шага было проведено исследование каждой матрицы откликов методом сингулярного разложения. По рисунку 4 видно, что во всех трёх наборах данных основную информацию несёт только 1 компонента. Таким образом, одного сенсора должно быть достаточно для корректного обнаружения веществ-маркеров и их смесей, а также единственный компонент ароматического отпечатка игрушек несет информацию обо всех содержащихся веществах.

Расширенный текст Дики-Фуллера для выявления стационарности показал, что, вопреки ожиданиям, векторы откликов отдельных сенсоров являются стационарными. Однако, выявить какие-либо закономерности в этом не удалось. Тем не менее, это ограничивает нас в выборе методов

преобразования для извлечения признаков – дискретное преобразование Фурье для нестационарных сигналов не подходит, остается вейвлет-преобразование.

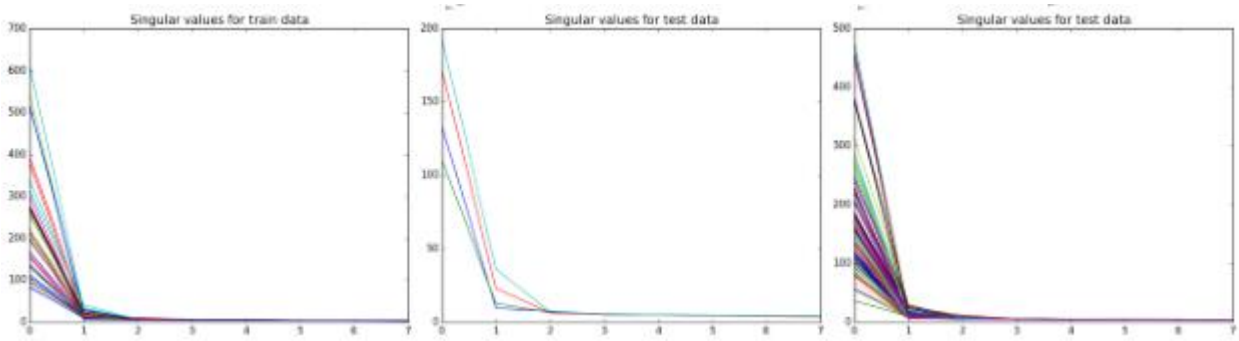


Рисунок 4– Графики сингулярных чисел для веществ, смесей и игрушек

Анализ графов кросс-корреляции сенсоров подтверждает гипотезу о том, что сенсоры почти всегда сильно коррелированы (их коэффициент кросс-корреляции – более 0.8), иногда возникает остовное дерево от одной вершины к нескольким, а также отрицательная кросс-корреляция (см. рисунок 5). Вероятно, эта информация косвенно подтверждает результаты, описанные в [21] о том, что необходимо выбирать методы извлечения признаков, опираясь на свойства отдельных матриц, а не данных вообще.

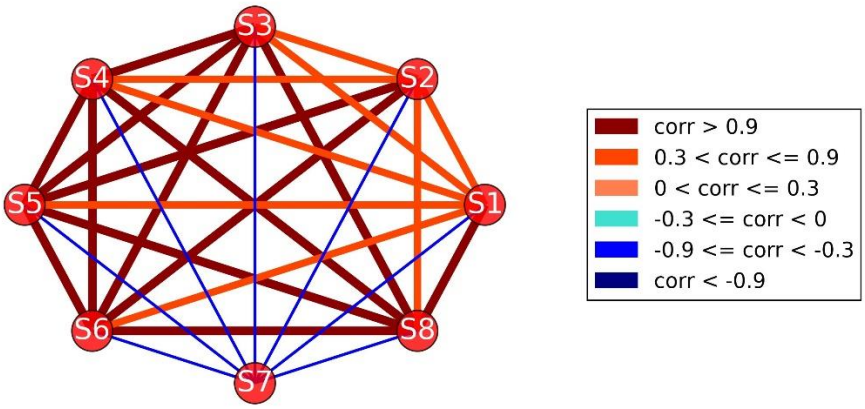


Рисунок 5 – Граф кросс-корреляции сенсоров

Далее был изучен подход, разработанный группой под руководством Т. А. Кучменко – для этого было реализовано построение радар-диаграмм максимальных изменений частоты для каждого сенсора. Визуально в построенных графиках легко выделяются характерные для каждого вещества элементы. Так, отличительной чертой ацетальдегида является правая половина графика и левая верхняя треть – острый угол в правой верхней

четверти и большой треугольник в правой нижней четверти, а также два треугольника в левой верхней трети (рисунок 6). Следовательно, преобразованные таким образом матрицы можно использовать в качестве первого метода отбора признаков. Кроме того, в диссертации Дроздовой [25] обращается внимание, что не только ΔF_{max} характеризует матрицу откликов, но и $\Delta F_{\text{равн}}$ – равновесная частота, то есть момент, когда процесс сорбции вошел в равновесную фазу.

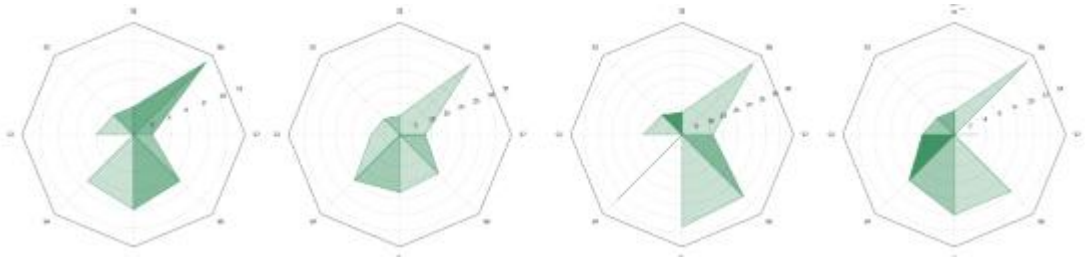


Рисунок 6 – Радар-диаграммы ΔF_{max} для ацетальдегида

Проверим предположение с помощью машины опорных векторов (SVM) с радиальной базисной функцией в качестве ядра, которая будет предсказывать вероятность того, что объект относится к какому-либо классу (таблица 1). Выбору и применению алгоритмов машинного обучения посвящен параграф 2.3. В параграфах 2.1 и 2.2 применяется машина опорных векторов со всеми параметрами по умолчанию как наиболее устойчивый классификатор в смысле переобучения.

Данные	Предсказанное	Истинное
Только исходные данные (нормализация + удаление тренда)	ДОФ – 44%, ацетон – 31%	ДОФ, ацетон
Только dFmax	Этилацетат - 39%, ДОФ – 31%	ДОФ, ацетон
Исходные данные + dFmax	Этилацетат - 31%, ДОФ – 23%	ДОФ, ацетон
Только dFравн	ДОФ – 22%, Ацетальдегид – 10%	ДОФ, ацетон
Исходные данные + dFравн	ДОФ 17%, бензол 11%	ДОФ, ацетон
Исходные данные + dFmax + dFравн	ДОФ – 23%, этилацетат - 12%	ДОФ, ацетон

Таблица 1 – Влияние состава признаков на предсказание меток

Для этого обучимся на этих данных и подадим на вход отпечаток первой игрушки, в которой графическим методом были обнаружены ацетон и

диоктилфталат. Поскольку решается задача классификации на N классов, то применим подход «Один против всех» и будем предсказывать вероятность появления тех или иных классов.

Рассмотрим подробнее, что в действительности значат такие результаты (см. рисунок 2). Диоктилфталат (ДОФ) – это наиболее представленный класс (22.5%), один из самых распространенных пластификаторов, который применяется для придания эластичности полимерным материалам; бензол, ацетальдегид и этилацетат представлены поровну и занимают второе место – 10% в обучающем наборе данных; ацетон – на третьем месте (7.5%), далее – пластизоль, обнаружение которой на самом деле не означает ничего, так как она является нетоксичным пластификатором, который обычно применяется в изготовлении детских игрушек. Для простоты метки пластизоли будем использовать в случаях, если прочих веществ не обнаружено – в эталонных результатах, полученных группой профессора Кучменко, присутствуют такие объекты. Прочие вещества и все смеси представлены 1 объектом – с их верным распознаванием могут возникнуть трудности.

Поскольку метрику точности (accuracy) нельзя считать корректной в случае сильно несбалансированного многоклассового массива данных, рассмотрим влияние компоновки обучающего датасета на такие метрики качества как coverage error (CE) и средняя точность ранжирования меток (label ranking average precision, LRAP). Обе эти метрики корректны для оценки качества многоклассовой классификации, при которой назначаются многие метки.

Coverage error показывает среднее количество меток классов, которое необходимо включить в финальное решение, чтобы были предсказаны все верные метки. Если задана матрица истинных меток классов $y \in \{0,1\}^{n_{\text{samples}} \times n_{\text{labels}}}$, и для каждой метки существует оценка вероятностей $\hat{f} \in \mathbb{R}^{n_{\text{samples}} \times n_{\text{labels}}}$, то эта метрика вычисляется следующим образом (1):

$$coverage(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \max_{j:y_{ij}=1} \text{rank}_{ij}$$ (1) [26]

, где $\text{rank}_{ij} = \left|\left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}\right|$. Вторая метрика - средняя точность ранжирования метки (label ranking average precision score, LRAP). Это среднее значение по каждой истинной метке, присвоенной каждому объекту, из соотношения истинных и суммарных меток с более низким рангом. Полученный результат всегда строго больше 0, а наилучшее значение равно 1. Если имеется ровно одна метка, средняя точность ранжирования метки эквивалентна среднему обратному рангу.

Метрика LRAP связана с метрикой average precision score, но основана на понятии ранжирования метки вместо precision и recall. При заданной матрице истинных меток классов $y \in \{0, 1\}^{n_{\text{samples}} \times n_{\text{labels}}}$ и оценках каждой метки $\hat{f} \in \mathcal{R}^{n_{\text{samples}} \times n_{\text{labels}}}$, метрика LRAP рассчитывается по следующей формуле (2):

$$LRAP(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{1}{|y_i|} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{\text{rank}_{ij}}$$ (2) [26]

, где $\mathcal{L}_{ij} = \left\{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$, $\text{rank}_{ij} = \left|\left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}\right|$ и $|\cdot|$ - это мощность множества.

Данные	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Только исходные данные (нормализация + удаление тренда)	19.52	**9.56**	0.07	**0.48**
Только dFmax	2.42	11.82	0.92	0.41
Исходные данные + dFmax	6.7	10.5	0.71	0.46
Только dFравн	***1.95***	11.54	***0.95***	0.41
Исходные данные + dFравн	5.27	11.02	0.78	0.40
Исходные данные + dFmax + Fравн	2.9	15.13	0.90	0.37

Таблица 2 – Влияние состава признаков на метрики CE и LRAP

Рассмотрим таблицу 2, где представлены эти метрики в зависимости от того, как собран обучающий набор данных. Несмотря на то, что оценки для тренировочного множества наихудшие, исходные данные дают наилучший результат для итоговой классификации по обеим метрикам - LRAP (0.48) и CE

(9.56). Метрики для тренировочной выборки обманчиво подводят нас к тому, что удачным решением будет формирование датасета только на основе равновесных частот, однако необходимо заметить, что на этапе обучения каждому объекту необходимо присвоить только одну метку, в то время как на этапе работы с новыми данными – несколько. Отсюда можно сделать вывод о том, что нет необходимости добавлять информацию о максимальных или равновесных частотах – достаточно работать с исходными данными, уделив большое внимание предварительной обработке.

Алгоритмы машинного обучения чувствительны к масштабированию данных, поэтому необходимо нормировать каждую отдельную матрицу. Кроме того, необходимо удалить тренд в векторах матриц откликов сенсоров. Другие приемы предварительной обработки, которые могут быть использованы – масштабирование матриц относительно друг друга и применение сглаживания полиномом N степени для удаления низкочастотного шума. Применение полосовых фильтров неоправданно в данном случае, так как неизвестна полоса пропускания, в которой лежат информативные частоты.

Метод	**CE(train)**	**CE(new)**	**LRAP(train)**	**LRAP(new)**
Только нормализация	15.72	**8.96**	0.26	**0.53**
Только удаление тренда	20.0	10.36	0.05	0.41
Нормализация + удаление тренда	19.52	**9.56**	0.07	**0.48**
Полином 3 степени	**4.32**	***13.16***	**0.83**	***0.45***
Полином 5 степени	**4.32**	12.30	**0.83**	***0.45***
Полином 7 степени	**4.32**	12.36	**0.83**	0.42
Нормализация + удаление тренда + полином	20.0	12.05	0.05	0.38
Нормализация + удаление тренда + масштабирование	19.52	**9.84**	0.07	**0.45**
Нормализация + удаление тренда + полином + масштабирование	20.0	11.85	0.05	0.41

Таблица 3 – Влияние методов предобработки на метрики CE и LRAP

Исследуем, как комбинация разных методов предварительной обработки влияет на рассмотренные выше метрики (таблица 3). По метрикам обучающего множества видно, что наилучшим методом предварительной обработки является сглаживание полиномом N степени, причем существенной

разницы в выборе степени N не наблюдается – остановимся на 3 степени как наилучшей. Однако для итоговой классификации наилучший метод – это нормализация. Необходимо учесть, что top-2 предсказанных меток оказались верными для удаления тренда (ДОФ – 33%, ацетон – 30%), для нормализации и удаления тренда (ДОФ – 52%, ацетон – 26%), что ненамного хуже, чем только нормализация. Хорошие результаты дает комплексный подход «нормализация + удаление тренда + масштабирование» (ацетон – 32%, ДОФ – 27%). Кроме того, он ненамного хуже при итоговой классификации, чем нормализация (CE=9.84 против 8.96). Все прочие подходы в top-2 дали предсказания в соответствии с распределением классов. Таким образом, лучше всего будет остановиться на подходе «нормализация + удаление тренда + масштабирование».

Метод	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Без PCA	19.52	**9.84**	0.07	**0.45**
121 компонента	19.52	11.25	0.07	0.42
8 компонент	19.52	11.36	0.07	0.41
2 компонента	19.52	10.02	0.07	0.48
1 компонента	19.52	**9.61**	0.07	**0.46**

Таблица 4 – Влияние методов предобработки на метрики CE и LRAP

Посмотрим, можно ли улучшить эти показатели с помощью сокращения размерности методом главных компонент (таблица 4). Видно, что на обучение применение PCA не влияет, а на итоговую классификацию влияет незначительно. Посмотрим top-2 предсказанных меток и их вероятности: без PCA – ацетон 32%, ДОФ 27%, 121 компонента – ацетон 33% и ДОФ 20%, 8 компонент – ДОФ 40% и ацетон 20%, 2 компоненты – ДОФ 67% и ацетон 62%, 1 компонента – ацетон 34% и ДОФ 26%.

Таблица 5 показывает, что PCA плохо применим к исследуемым данным, так как в самом лучшем случае, на исходных данных без предварительной обработки, первая главная компонента объясняет только 68% дисперсии, а при наилучшем способе обработки – всего 14%. Затем значение объяснённой дисперсии начинает стремительно падать, и уже 3 компонента объясняет лишь 9%, дальнейшие компоненты объясняют менее

1% модели. Таким образом, можно сделать вывод о том, что метод главных компонент для уменьшения размерности данных не подходит.

Метод	EVR(1)	EVR(2)	EVR(3)	EVR(4)	EVR(5)
Сырые данные	0.6875	0.1677	0.0974	0.0074	0.005
Только нормализация	0.3062	0.0810	0.0663	0.0554	0.0460
Только удаление тренда	0.0573	0.0438	0.0417	0.0396	0.0381
Нормализация + удаление тренда	0.0861	0.0852	0.0770	0.0642	0.0591
Полином 3 степени	**0.6927**	0.1688	0.0980	0.0158	0.0073
Полином 5 степени	**0.6920**	0.1686	0.0979	0.0159	0.0073
Полином 7 степени	**0.6917**	0.1686	0.0979	0.0159	0.0073
Нормализация + удаление тренда + полином	0.1657	0.1475	0.1246	0.1163	0.0984
Нормализация + удаление тренда + масштабирование	0.1406	0.1157	0.0994	0.0782	0.0642
Нормализация + удаление тренда + полином + масштабирование	0.1367	0.1104	0.0953	0.0878	0.0852

Таблица 5 – Объясненная дисперсия остатков для первых 5 главных компонент PCA в зависимости от выбора модели предварительной обработки

Рассмотрим возможность применения линейного дискриминатного анализа для уменьшения размерности данных (таблица 6): для этого исследуем объяснённую дисперсию остатков для первых пяти факторов. Видно, что наилучшим способом предварительной обработки в этом случае является нормализация и удаление тренда – тогда первый фактор объясняет 99.95% модели. Это наилучший результат для LDA, который значительно превосходит PCA в задаче уменьшения размерности.

Метод	EVR(1)	EVR(2)	EVR(3)	EVR(4)	EVR(5)
Сырые данные	0.6127	0.2003	0.0588	0.0466	0.0233
Только нормализация	0.4071	0.1743	0.1365	0.0824	0.0528
Только удаление тренда	0.1953	0.1323	0.1151	0.1062	0.099
Нормализация + удаление тренда	**0.9995**	0.0004	0.0001	0	0
Полином 3 степени	***0. 7827***	0.1313	0.0321	0.321	0.0184
Полином 5 степени	0.6820	0.1748	0.0544	0.0302	0.0198
Полином 7 степени	0.6471	0.2040	0.0539	0.0312	0.0150
Нормализация + удаление тренда + полином(5)	0.4181	0.2081	0.1193	0.0929	0.051
Нормализация + удаление тренда + масштабирование	0.6846	0.3137	0.0017	0	0
Нормализация + удаление тренда + полином + масштабирование	0.4181	0.2081	0.1193	0.0929	0.0510

Таблица 5 – Объясненная дисперсия остатков для первых 5 факторов LDA в зависимости от выбора модели предварительной обработки

Таким образом, выстраивается следующий конвейер предварительной обработки: нормализация каждой матрицы, которая включает в себя

центрирование с помощью вычитания среднего и деление на стандартное отклонение, удаление тренда по методу скользящего среднего в каждом канале и масштабирование матриц относительно друг друга.

2.2 Подходы к решению проблемы несбалансированности данных

2.2.1 Балансировка массива данных для предсказания редких классов

Очевидным решение проблемы несбалансированности является удаление «лишних» объектов и, как следствие, сокращение размера обучающей выборки до 20 объектов – по количеству классов. Вариацией этого подхода является замена всех объектов одного класса усредненным объектом. Противоположный вариант – простое копирование объектов. Более интеллектуальные методы предлагает пакет imbalanced-learn [27] для Python – он содержит under-sampling и over-sampling алгоритмы и алгоритмы, реализующие комбинацию этих методов. Сравним все рассмотренные выше альтернативы (таблица 6).

Метод	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Удаление лишних	14.3	16.14	0.33	0.20
Усреднение	3.85	16.09	0.85	0.21
Under-sampling (ClusterCentroids)	17.15	16.81	0.19	0.35
Копирование	1.0	***10.01***	1.0	***0.43***
Over-sampling (ADASYN)	3.47	**11.22**	0.87	**0.36**

Таблица 6 – Влияние методов балансировки на метрики CE и LRAP

Несмотря на простоту, копирование объектов до 9 штук (по наиболее представленному классу) оказалось наиболее удачным для распознавания отдельных веществ: мы, действительно, должны предсказать ровно 1 метку для каждого объекта, а средняя точность ранжирования при этом равна 1, то есть единственная верная метка находится на 1 месте. Однако, для назначения многих меток для игрушек этот подход в корне не верный – копии веществ не несут новой никакой информации, которая позволила бы извлечь заданное вещество из ароматического отпечатка многих веществ, собранных в концентрациях.

В действительности, наилучшим методом является искусственное дополнение выборки методом ADASYN (Adaptive Synthetic Sampling Approach

for Imbalanced Learning), предложенным в [28]. Идея подхода состоит в использовании взвешенного распределения объектов редких классов пропорционально сложности обучения на них: больше синтетических данных будет сгенерировано для тех классов, при обучении на которых возникли наибольшие трудности. В результате подход ADASYN улучшает обучение двумя способами: уменьшает смещение, вызванное дисбалансом классов и адаптивно сдвигает границу принятия решения к трудным примерам.

Объекты каждого класса генерировались в противопоставлении ко всем прочим, чтобы добиться максимального различия между всеми классами. Таким образом получилось, что экземпляров каждого класса от 35 до 39 штук. Однако, вопреки ожиданиям, возникли серьезные проблемы с генерацией двух наиболее представленных классов – диоктилфталата и ацетальдегида. По рисунку 7 видно, что многие объекты лежат очень тесно на границе классов, и ни одна из четырех машин опорных векторов не способна разделить эти объекты. Кроме того, существенного повышения точности распознавания не произошло: диоктилфталат по-прежнему стоит на первом месте среди назначаемых меток для большинства новых объектов.

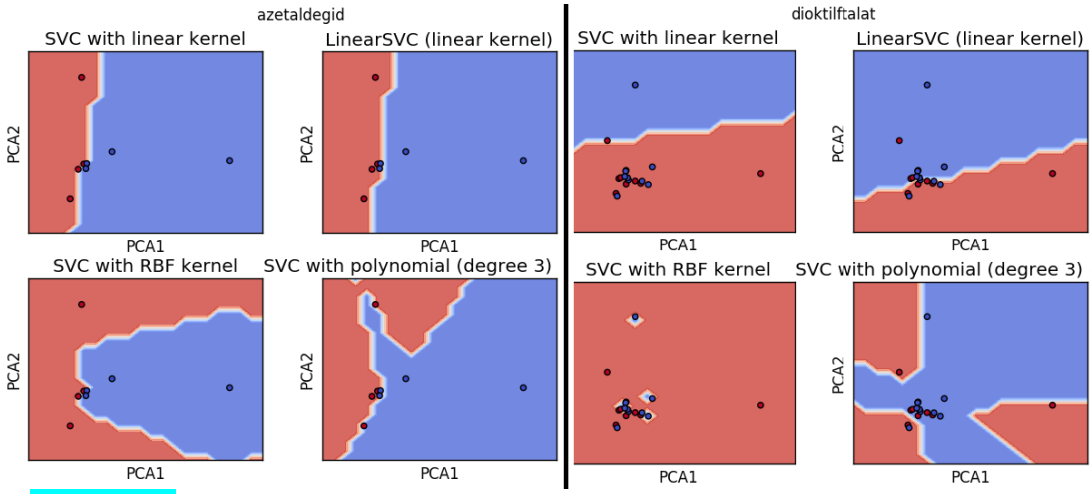


Рисунок 7 – Синтетические объекты диоктилфталата и ацетальдегида, сгенерированные методом ADASYN

Посмотрим, подтверждается ли вывод о том, что PCA в качестве метода понижения размерности для этих данных не подходит, а LDA, напротив, несёт 99% информации в первом факторе (таблица 7). Действительно, PCA плохо подходит для работы с этими данными. Однако, LDA ненамного лучше – 18%

объяснённое дисперсии модели. Вероятно, генерация новых объектов изменила природу данных.

Метод	EVR(1)	EVR(2)	EVR(3)	EVR(4)	EVR(5)
PCA	0.1002	0.0662	0.0557	0.0511	0.0493
LDA	0.1837	0.1356	0.1	0.0856	0.0765

Таблица 7 – Сравнение объяснённой дисперсии для PCA и LDA

Поскольку теперь выборка сбалансирована, можно подобрать алгоритм классификации по кросс-валидации, а в качестве новых данных, для проверки информативности сгенерированных данных, подавать исходные вещества и смеси. Таблица 8 содержит среднюю точность (accuracy) и стандартное отклонение для логистической регрессии, k ближайших соседей, машины опорных векторов, ExtraTrees и AdaBoost с параметрами по умолчанию. Для выбора наилучшей модели необходимо минимизировать среднее значение по кросс-валидации на семи запусках, а затем, если таковых окажется несколько, выбрать алгоритм с минимальным стандартным отклонением. Видно, что наилучшую точность показывает машина опорных векторов и логистическая регрессия (92.4% +\- 1.9%).

Метод	Mean accuracy	std
LogisticRegression	**0.924**	**0.019**
KNeighborsClassifier	0.888	0.034
RandomForestClassifier	0.908	0.022
SVC	**0.924**	**0.019**
ExtraTreesClassifier	0.908	0.022
AdaBoostClassifier	0.892	0.030

Таблица 8 – Подбор алгоритма классификации

После подбора параметров, запустим модели на реальных данных. Логистическая регрессия после подбора параметров (penalty = 'l2', C = 1) достигла точности 92.3%, но на исходные («настоящих») данные оказалась абсолютно немасштабируемы (точность 0%). Машина опорных векторов после подбора параметров (C = 1, gamma = 0.001) достигла точности 92.8% и на новых данных показала точность 22.5%. Таким образом, можно сделать вывод о том, что искусственные объекты веществ и смесей, генерируемые методом ADASYN, не являются достоверными для обучения.

2.2.2 Моделирование объектов с помощью Generative adversarial network

Генеративные состязательные модели (generative adversarial network, GAN) предложены сотрудником Facebook Яном Гудфеллоу (Ian Goodfellow) в 2014 году [29]. В оригинальной работе были представлены 2 сети прямого распространения, которые обучаются одновременно без учителя: генеративная модель G (генератор), которая на вход принимает случайный шум и затем учится имитировать распределение реальных данных, и дискриминативная модель D (дискриминатор), которая принимает на вход наборы реальных и сгенерированных данных и оценивает вероятность того, являются ли данные настоящими или искусственными. Процедура обучения для G заключается в максимизации вероятности того, что D совершит ошибку. В терминах теории игр эта модель соответствует минимаксной игре двух игроков.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right] \quad (3)$$

В пространстве произвольных функций G и D существует единственное решение, при котором G восстанавливает распределение обучающей выборки, а распределение D равно 1/2 везде. В случае, когда G и D являются многослойными перцептронами, вся система может обучаться с помощью алгоритма обратного распространения ошибки.

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \quad (4)$$

Градиентный спуск дискриминатора описывается формулой 3. Первое слагаемое увеличивает вероятность того, что реальные данные (x) оцениваются как хорошие, второе слагаемое должно понижать вероятности того, что сгенерированные данные G(z) распознаются как «подделка». Цель обучения генератора, напротив, состоит в том, чтобы сгенерированные данные с высокой вероятностью оценивались как хорошие. Градиентный спуск генератора описывается формулой 4.

Поочерёдно оптимизируя градиент каждой из сетей на новых наборах реальных и сгенерированных данных, GAN будет медленно сходиться к генерации данных, которые так же реалистичны, как и предъявленные в обучающей выборке объекты.

GAN получили широкое распространение в обработке изображений в виде архитектуры DCGAN (Deep Convolutional GAN) [30], которая состоит из сверточной нейронной сети (convolutional neural network, CNN) в качестве генеративной модели и развёртывающей нейронной сети (deconvolutional neural network, CNN) в качестве дискриминативной.

В 2016 году в работе [31] представлена архитектура InfoGAN - информационно-теоретическое расширение GAN. InfoGAN представляет собой генеративную состязательную сеть, которая максимизирует взаимную информацию между подмножеством скрытых переменных. Авторы показали, что InfoGAN способна успешно отличать стили написания от форм цифр в наборе данных MNIST, генерировать номера домов с различной освещенностью и имитировать «обрезанные» изображения с номерами домов в наборе SVHN. InfoGAN также способна обнаруживать визуальные концепции, включая стили стрижек, наличие / отсутствие очков и эмоции в наборе данных CelebA face. Эксперименты показывают, что InfoGAN обучается интерпретируемым представлениям, которые способны конкурировать с представлениями, на которых обучаются существующие методы обучения с учителем.

Опыт успешного применения InfoGAN к генерации одномерных временных рядов двух классов проиллюстрирован на Github-репозитории [32]. Разработчик не предоставил исходные данные и информацию о размерности обучающей выборки, поэтому сложно говорить о том, насколько трудно повторить его успех и расширить его на многоканальные сигналы, которыми являются наши данные.

Так же остаётся открытым вопрос о том, как различать классы в искусственно сгенерированных данных: это не является проблемой для

изображений, так как объекты разных классов визуально различимы, в то время как для нашей задачи визуально различать исходные матрицы очень тяжело, и полагаться на косвенные методы, такие, как радиальные диаграммы ΔF_{max}, не представляется возможным.

Метод	G_loss (0)	D_loss (0)	G_loss (100)	D_loss (100)
Только нормализация	0.693193	0.692456	0.000064	0.000000
Только удаление тренда	0.692932	0.695542	**0.00002**	**0.00003**
Нормализация + удаление тренда	0.693598	0.693190	0.344400	0.000009
Нормализация + удаление тренда + LDA	0.691807	0.692945	3.708233	0.000006
Полином 3 степени	0.692991	0.677187	7.246695	0.000069
Полином 5 степени	0.691710	0.722930	6.977266	0.000067
Полином 7 степени	0.692559	0.681715	4.650308	0.000031
Нормализация + удаление тренда + полином	0.693081	0.693099	0.347795	0.002469
Нормализация + удаление тренда + масштабирование	0.692997	0.686141	0.001910	0.000697
Нормализация + удаление тренда + полином(3) + масштабирование	0.693046	0.694560	0.004515	0.000073
Нормализация + удаление тренда + полином(5) + масштабирование	0.693109	0.701059	0.000113	0.000224
Нормализация + удаление тренда + полином(7) + масштабирование	0.693423	0.694273	0.000730	0.003580

Таблица 9 – Влияние методов предварительной обработка на обучение DCGAN в течение 100 эпох, batch_size=4

Рассмотрим саму возможность обучения GAN в течение 100 эпох на примере DCGAN (таблица 9) при выборе различных методов предварительной обработки. Видно, что сглаживание полиномом не подходит, поскольку функция потерь генератора значительно превосходит функцию потерь дискриминатора – это значит, что дискриминатор выбраковывает сгенерированные данные с высокой надёжностью. Примерно равны функции потерь, когда из данных удаляется только тренд, но поскольку эту функции ничтожно малы, есть основания предполагать переобучение. Кроме того, невозможно сказать, какие именно объекты сгенерировала сеть - вещества ли это или смеси. Очевидное решение этой проблемы – переобучать GAN для каждого класса веществ, чтобы одна сеть могла генерировать только одно вещество. Существенный недостаток этого подхода состоит в том, что обучить

сеть на единственном примере невозможно, а именно в этом – увеличении числа наименее представленных классов и состоит задача.

Альтернативный способ генерации объектов определённых классов – это так называемые «условные» генеративные модели (conditional GAN), описанные в 2014 году в [33]. Кроме того, можно изменить внутреннее представление классов – вместо бинарной задачи «настоящие» и «поддельные» объекты, перейти к N-мерной задаче «настоящий объект класса 1», «настоящий объект класса N-1» и общая группа «поддельных» объектов – этот подход описан в [34], одним из автором которой является Ян Гудфеллоу. Однако предполагается, что обучение такой сети на единственном «настоящем» примере так же затруднительно, поэтому развитие этой идеи применительно к нашей задаче не имеет смысла .

2.3 Применение алгоритмов машинного обучения и нейронных сетей

2.3.1 Сравнение основных алгоритмов машинного обучения

Большинство алгоритмов во фреймворке Sklearn поддерживают многоклассовую классификацию: Naive Bayes, LDA and QDA, Decision Trees, Random Forests, Nearest Neighbors способны работать с многими классами по умолчанию. При этом, назначение более одной метки класса поддерживается в Decision Trees, Random Forests, Nearest Neighbors. [35] Для этого реализованы подходы «каждый-против-каждого» (one-vs-one) и «один-против-всех» (one-vs-all). SVC поддерживает мультикласс-мультилэйбл подход с помощью мета-алгоритма One-Vs-One, стратегия One-Vs-All реализована для всех линейных моделей, за исключением SVC.

Обучение производится на выборке из 40 объектов (36 веществ и 4 смесей), каждый объект обучающей выборки принадлежит единственному классу. Для исходной выборки провести подбор алгоритмов методом кросс-валидации не представляется возможным из-за присутствия редких классов, представленных единственным объектом. Проведение тестирования так же затруднительно, поскольку в этом случае придется уменьшить обучающую выборку.

Ожидается, что новым данным, которые представляют из себя ароматические отпечатки игрушек, алгоритм будет назначать несколько меток класса – согласно тому, какие вещества-маркеры обучающей выборки обнаружены в пробах.

В настоящей работе рассматриваются следующие алгоритмы: метод опорных векторов (Support Vector Machine, SVM), ансамбль деревьев решений «случайный лес» (Random Forest, RF) и усиление «случайного леса» с помощью мета-алгоритма Bagging (bootstrap aggregating), основная идея которого состоит в обучении на N выборках с повторениями. Алгоритмически, для SVM и Bagging преодолеть дисбаланс классов можно, выставив параметр class_weight в «balanced».

Кроме того, использовались следующие параметры, отличные от стандартных: для SVM ядро – радиальная базисная функция, для Random Forest – 5000 деревьев в ансамбле, и Bagging 25 Random Forest из 1500 деревьев.

Дополнительно было проведено исследование выбросов с помощью рядя алгоритмов (рисунок 8). Отбор алгоритмов производился на основе анализа объектов, признанных выбросами: во-первых, поскольку массив данных слишком мал, мы не можем себе позволить удалить больше число объектов, значит выбирать следует алгоритм, который предложил наименьшее число выбросов, во-вторых, необходимо, чтобы эти выбросы принадлежали классу, который представлен многими объектами, в противном случае анализ выбросов неинформативен – мы не можем удалять единственный объекты. В результате: One-Class SVM с радиальным ядром – 40 выбросов (не подходит), эмпирическая ковариация – 11 выбросов (не подходит), One-Class NuSVM – 32 объекта (не подходит), робастная ковариация с минимальным определителем обнаружила 11 выбросов (не подходит), робастная ковариация – 3 (ацетальдегид, стирол, ДОФ с этилацетатом) – можем удалить только ацетальдегид, Isolation Forest – 4 объекта (ацетальдегид, гексан, 2 ДОФ) –

можем удалить 3 из 4 выбросов, причем, удаление ДОФ частично сбалансирует массив данных.

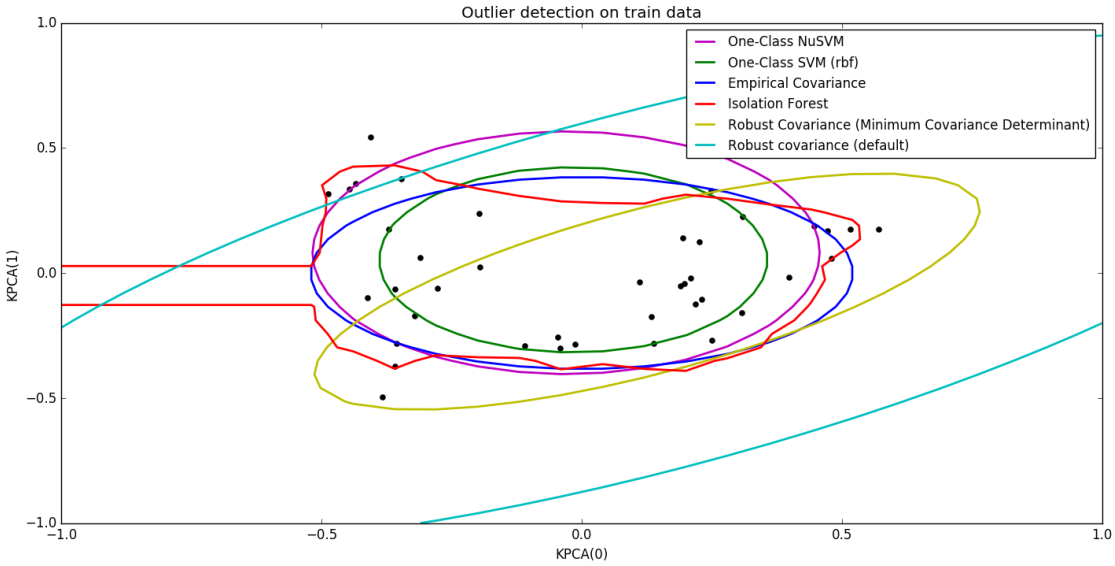


Рисунок 8 – Выбросы в исходном наборе данных

Выбор был остановлен на алгоритме Isolation Forest. Выбросами были признаны: ацетальдегид 1 мкл (ацетальдегид 0), единственный экземпляр гексана, диоктилфталат (диоктилфталат 15) и диоктилфталат 1 мкл на пластизоле (диоктилфталт 21). Мы не можем себе позволить удалить гексан, однако удаление ацетальдегида и ДОФ могут быть полезны.

Кроме рассмотренных ранее подходов – обучения на всех веществах-маркерах или обучения на сгенерированных данных – возможны альтернативные способы уменьшения «перевеса» ДОФ и более точной классификации единичных классов.

Первый подход состоит в удалении ДОФ из обучающей выборки, обучения и классификации на 19 классов, а затем второй цикл обучения и бинарной классификации по принципу «один-против-всех». В этом случае первый датасет будет содержать: 4 бензола и этилацетата, 3 ацетальдегида и ацетона, 2 пластизоли по одному – бензин, бутанол, бутилацетат, гексан, изобутанол, изопропанол, стирол, толуол, фенол, ДОФ с ацетоном, ДОФ с этилацетатом, ДОФ с бензолом и ДОФ с ацетальдегидом. Наиболее представленные классы превосходят по числу объектов наименее представленный класс всего в 4 раза (до удаления ДОФ – в 9), а в задаче

детектирования ДОФ – также примерно в 3.5 раза (9 к 31). То есть полученные датасеты гораздо более сбалансированные, чем исходный.

Второй подход – это объединение всех «единичных» классов в класс «other» и последующее детектирование их отдельно. В первом случае выборка будет составлять 40 объектов и 7 классов с дисбалансом в пользу other – тех самых «редких» объектов. Во втором случае – 14 «редких» и 26 other для обозначения прочих веществ или их отсутствия.

Алгоритм	**CE(train)**	**CE(new)**	**LRAP(train)**	**LRAP(new)**
Исходные данные				
Support Vector Machine	19.52	**9.92**	0.07	**0.48**
Bagging with Random Forest	20.0	9.96	0.05	0.42
Random Forest	19.52	***10.18***	0.07	***0.48***
Исходные данные без выбросов				
Support Vector Machine	1.0	**9.9**	1.0	**0.33**
Bagging with Random Forest	0.94	19.0	0.05	0.32
Random Forest	1.0	**9.82**	1.0	**0.32**
Исходные данные без ДОФ				
Support Vector Machine	1.0	10.83	1.0	0.30
Bagging with Random Forest	20.0	**9.46**	0.05	**0.35**
Random Forest	1.0	10.16	1.0	0.30
Распознавание ДОФ				
Support Vector Machine	1.0	***1.42***	1.0	***0.97***
Bagging with Random Forest	1.22	1.42	0.88	0.97
Random Forest	1.0	***1.42***	1.0	***0.97***
Исходные данные с OTHER				
Support Vector Machine	1.0	**4.84**	1.0	**0.56**
Bagging with Random Forest	6.4	5.09	0.22	0.55
Random Forest	1.0	5.09	1.0	0.56
Распознавание объектов из OTHER				
Support Vector Machine	5.55	5.82	0.69	0.80
Bagging with Random Forest	5.9	**5.0**	0.67	**0.84**
Random Forest	1.0	5.41	1.0	0.80
Over-sampling (ADASYN)				
Support Vector Machine	3.47	***13.65***	0.87	***0.25***
Bagging with Random Forest	7.93	17.25	0.65	0.16
Random Forest	3.47	***13.38***	0.11	***0.24***

Таблица 10 – Подбор алгоритмов классификации

Таблица 10 содержит результаты экспериментов с различными массивам и данных. Беглый взгляд на неё позволяет сделать следующее наблюдение: если мы добиваемся идеального результата на тренировочном множестве (см. SVM и RF для данных без ДОФ и для данных без выбросов), то есть в задаче

распознавания веществ-маркеров и их смесей, то мы получаем не самые лучше результаты в финальной задаче детектирования веществ в отпечатках игрушек. Если же результаты идеальны для игрушек, то этот алгоритм не справляется с распознаванием чистых веществ: например, см. SVM на исходных данных – CE = 19.52 для веществ, то есть мы должны вместо одного назначать почти все 20 меток класса одному веществу.

Видно, худший результат, как и ожидалось, получен для искусственных объектов. Видно так же, что результат для распознавания всех веществ за исключение ДОФ не очень хорош, в то время как ДОФ в задаче бинарной классификации распознаётся отлично (обе метрики максимально близки к наилучшему значению). Кроме того, объединение редких классов в класс OTHER даёт лучший результат, в в отличие от отдельного распознавания ДОФ.

Остановимся подробнее на анализе результатов. Усреднённые метрики, приведённые в таблице 10, позволяют сравнивать алгоритмы между собой, однако гораздо более важным является оценка качества предсказания для каждого отдельного объекта – как именно назначаются классы, какова «уверенность» алгоритма в присутствии этих веществ в ароматическом отпечатке, каково распределение вероятностей в первой пятерке и первой десятке. Очевидно, что среди ошибок для нас предпочтительнее ложное срабатывание – лучше увидеть «лишний» токсичный компонент в ароматическом отпечатке детской игрушки, чем пропустить его, и, в случае, если он был единственный, считать игрушку безопасной.

Перед тем, как приступить к анализу общей производительности алгоритмов, нужно рассмотреть вопрос о том, сколько меток должны быть показаны в итоговом решении. Ранее мы оценивали качество работы алгоритмов по усреднённым на всю выборку метрикам CE и LRAP – теперь рассчитаем индивидуальные метрики для каждого нового объекта, отсортировав предсказанные вероятности классов по убыванию. В случае, если классификатор успешно обнаружил все компоненты, которые

присутствуют в смеси, coverage error покажет, сколько первых компонент вектора вероятностей нужно взять, а label ranking average precision – насколько мы можем доверять этому результату.

Однако, необходимо понимать одно важное ограничение в интерпретации этой метрики: пусть имеются 2 класса, 1 и 0, и два объекта, для которых верная метка – [1, 0], а предсказанная [0, 1]. В этом случае, из-за обратного порядка, метрика будет равняться 0.5, несмотря на то, что классы предсказаны верно. Это не позволяет в полной мере интерпретировать данную метрику как аналог multiclass и multilabel точности (accuracy).

Анализ меток проводился вручную с помощью таблицы MS Excel. Для этого вероятности назначения меток отсортируем по убыванию, а зачем ячейке назначим цвет в соответствии со следующими начальными критериями:

1. Категория 1 – CE совпадает с числом реальных меток, LRAP = 1
2. Категория 2 - CE не совпадает с числом реальных меток менее, чем в 2 раза, LRAP >= 0.8
3. Категория 3 - CE не совпадает с числом реальных меток в 2 и более раз, или LRAP >= 0.6 и LRAP < 0.8
4. Категория 4 - CE не совпадает с числом реальных меток в 2 и более раз, или LRAP >= 0.4 и LRAP < 0.6.
5. Категория 5 - CE не совпадает с числом реальных меток в 2 и более раз, и LRAP < 0.4.
6. Категория 6 – Присутствие веществ не обнаружено (CE = 20.0 и \ или LRAP =< 0.07).

Указанные выше категории для наглядности обозначаются цветом, а затем стандартный макрос MS Excel подсчитывает количество меток каждого цвета. Первые 4 категории представляют собой градации верных ответов классификатора, категория 5 нужна для того, чтобы найти критерий, по которому определяются объекты, в которых веществ-маркеры не обнаружены. Особое внимание уделяется количеству меток последней категории – это

именно те случаи, в которых реально присутствующие вещества не были обнаружены, то есть классификатор дал абсолютно неверный ответ. Количество объектов, попавших в первые 3 категории, будем максимизировать, 4 и 6 – минимизировать.

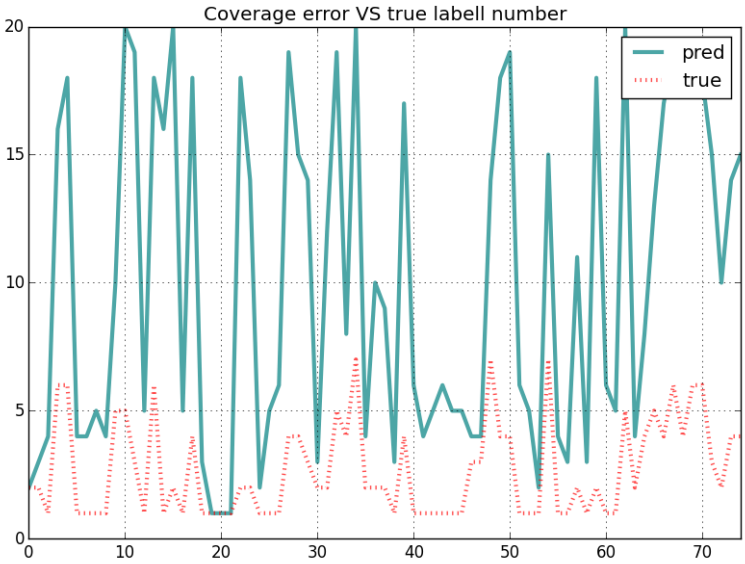


Рисунок 9 – Сравнение оценки количества меток, которые необходимо предсказать, и реального количества меток на исходных данных (SVM)

Рассмотрим результаты на исходных данных. Машина опорных векторов выдаёт результаты, полностью совпадающие с представленной выше классификацией. Bagging с Random Forest, как и сам Random Forest, отклоняются от этой классификации по критериям категории 6. Таким образом, после первой проверки результатов ни одному объекту категория 6 не назначена. Однако в их результатах наблюдается некоторая закономерность, которая позволит выделять вещества, которым не назначено меток.

В метках Bagging с Random Forest 5 раз возникла комбинация "CE = 7.0, LRAP = 0.14", 4 раза - "CE = 9.0, LRAP = 0.11", трижды повторялась комбинация "CE = 8.0, LRAP = 0.125", трижды - "CE = 6.0, LRAP = 0.16", дважды - "CE = 10.0, LRAP = 0.1" и по одному разу - "CE = 12.0, LRAP = 0.8" и "CE = 19.0, LRAP = 0.25". Назначим для результатов Bagging с Random Forest категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

В метках Random Forest 5 раз возникла комбинация "CE = 6.0 LRAP = 0.16", трижды повторялась комбинация "CE = 8.0 LRAP = 0.125" и "CE = 12.0 LRAP = 0.8", дважды - "CE = 11.0 LRAP = 0.09", "CE = 13.0 LRAP = 0.07" , "CE = 4.0 LRAP = 0.25" и по одному разу - "CE = 9.0 LRAP = 0.11", "CE = 10.0 LRAP = 0.1", "CE = 7.0 LRAP = 0.14". Назначим для результатов Bagging с Random Forest категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

В сравнении результатов Bagging и Bagging с Random Forest совпадают почти все метрики, за исключение: "CE = 11.0 LRAP = 0.09", "CE = 13.0 LRAP = 0.07", "CE = 4.0 LRAP = 0.25".

Категории	**SVM**	**Random Forest**	**Bagging with Random Forest**
Исходные данные			
Категория 1	5	5	4
Категория 2	4	6	4
Категория 3	7	17	17
Категория 4	20	76	10
Категория 5	14	40	40
Категория 6	25	0	0
Исходные данные без выбросов			
Категория 1	4	4	4
Категория 2	5	4	4
Категория 3	9	19	17
Категория 4	16	8	7
Категория 5	17	21	22
Категория 6	24	19	21

Таблица 11 – Распределение оценок качества алгоритмов машинного обучения

Теперь рассмотрим результаты на данных без выбросов. Здесь результаты всех трёх алгоритмов требуют пересмотра категории 6. Машина опорных векторов даёт очень стабильные результаты: 8 раз - "CE = 5.0 LRAP = 0.2", 7 раз - "CE = 4.0 LRAP = 0.25", 5 раз "CE = 6.0 LRAP = 0.16", 4 раза - "CE = 3.0 LRAP = 0.33" и 1 раз - "CE = 7.0 LRAP = 0.14". Назначим для результатов SVM категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим. Причина выбора трёх в качестве порога состоит в поддержании сравнимости алгоритмов.

Менее стабильные результаты даёт Bagging с Random Forest: 12 раз - "CE = 6.0 LRAP = 0.16", 5 раз - "CE = 7.0 LRAP = 0.14", 3 раза - "CE = 8.0 LRAP = 0.125", дважды - "CE = 9.0 LRAP = 0.11", "CE = 10.0 LRAP = 0.1" и 1 раз - "CE = 8.0 LRAP = 0.125" и "CE = 20.0 LRAP = 0.6". Назначим для результатов Bagging с Random Forest категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

Менее стабильные результаты даёт Random Forest: 6 раз - "CE = 6.0 LRAP = 0.16", 5 раз - "CE = 5.0 LRAP = 0.2", 4 раза - "CE = 10.0 LRAP = 0.1", 3 раза - "CE = 8.0 LRAP = 0.125", 2 - "CE = 11.0 LRAP = 0.09" и 1 раз - "CE = 7.0 LRAP = 0.14", "CE = 14.0 LRAP = 0.07" и "CE = 12.0 LRAP = 0.8" Назначим для результатов Random Forest категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

Рассмотрим таблицу 11, где приводится распределение категорий по всем алгоритмам. Для удобства сравнения введём следующие метрики (5 и 6):

$$\mu_1 = \frac{C1+C2+C3+C6}{N} \quad (5)$$

$$\mu_2 = \frac{C4+C5}{N} \quad (6)$$

Метрика 1 обозначает долю объектов, которые мы можем предсказывать с хорошей долей уверенности, метрика 2 – это доля объектов с низким качеством предсказания. C1…C6 – это присвоенные категории, а N – это общее число новых объектов (в нашем случае 75). Цель – максимизировать первую метрику и, соответственно, минимизировать вторую. Кроме того, интересно также исследовать площадь между кривыми, которые отображают реальное количество меток и coverage error (см. рисунок 7). Эту метрику можно интерпретировать как способность алгоритма предсказать наиболее близкое к реальному число меток. Несмотря на то, что ложные срабатывания мы считаем приемлемыми, это позволит их минимизировать: чем меньше площадь между кривыми, тем ближе к реальному число меток в итоговом решении. Рассчитаем значения метрик для каждого алгоритма (таблица 12).

Алгоритм	μ_1	μ_2	S
Исходные данные			
Support Vector Machine	0.54	0.45	888.0
Bagging with Random Forest	**0.62**	**0.37**	573.0
Random Forest	0.54	0.45	**558.5**
Исходные данные без выбросов			
Support Vector Machine	0.56	0.44	576.5
Bagging with Random Forest	**0.61**	**0.38**	**541.5**
Random Forest	**0.61**	**0.38**	552.5

Таблица 12 – Метрики μ_1, μ_2 и S

Видно, что на исходных данных меньшую площадь под кривой имеет Random Forest. Однако, он показывает не лучшие результаты по другим метрикам. Наилучшая площадь под кривой – у Bagging with Random Forest на данных без выбросов. Таким образом, благодаря введению метрики площади удаётся сравнить алгоритмы Bagging with Random Forest и Random Forest на данных без выбросов, где они дают одинаковый результат по другим метрикам.

2.3.2 Сравнение архитектур нейронных сетей

Многоклассовая классификация и назначение многих меток одному объекту – естественная задача для нейронных сетей, для этого достаточно указать желаемую размерность выходного слоя и в качестве функции потерь выбрать категориальную кросс-энтропию.

В настоящей работе применяется фреймворк для работы с нейронными сетями для Python Keras и новейшая версия фреймворка для машинного обучения Scikit-learn. Исследованы возможности работы с тремя различными видами нейронных сетей – многослойный перцептрон с обучением методом обратного распространения ошибки как пример простейшей нейронной сети, полносвязная нейронная сеть (Приложение 2) и LSTM (Приложение 3).

В результате для исходных данных оптимальными были признаны следующие параметры:

- Multilayer Perceptron: alpha 97.4 (перебор альфа вручную)
- Fully Connected Network: batch_size = 5, обучение в течение 100 эпох (Gridsearch)

- LSTM: batch_size = 5, look_back = 3, обучение в течение 100 эпох (без подбора).

Добавление Batch Normalization и серии dropout в FCN и LSTM позволяет избежать переобучения даже на такой маленькой выборке на протяжении 100 эпох. Видно, что на исходных данных нейронные сети способны достигать наилучшей точности при обучении на веществах и смесях, однако даже в этом случае детектирование этих веществ в ароматических отпечатках игрушек не является успешным. Более того, нейронные сети в задаче исследования ароматических отпечатков игрушек показывают худшую производительность в сравнении с SVM и Random Forest.

Алгоритм	**CE(train)**	**CE(new)**	**LRAP(train)**	**LRAP(new)**
Исходные данные				
Multilayer Perceptron (alpha 97.4)	20.0	**10.82**	0.05	**0.38**
Fully Connected Network	1.0	11.56	1.0	0.37
LSTM Stateful	1.0	12.33	1.0	0.34
Исходные данные без выбросов				
Multilayer Perceptron (alpha 97.4)	18.45	10.57	0.12	0.37
Fully Connected Network	1.0	11.68	1.0	0.36
LSTM	8.18	**11.17**	0.64	**0.40**
Исходные данные без ДОФ				
Multilayer Perceptron (alpha 97.4)	20.0	**9.96**	0.05	**0.36**
Fully Connected Network	3.75	11.13	0.88	0.32
LSTM	10.5	10.3	0.52	0.33
Распознавание ДОФ				
Multilayer Perceptron (alpha 97.4)	1.075	**1.42**	0.96	**0.97**
Fully Connected Network	1.05	1.69	0.97	0.84
LSTM Stateful	1.05	1.62	0.97	0.87
Исходные данные с OTHER				
Multilayer Perceptron (alpha 97.4)	3.55	5.44	0.63	0.50
Fully Connected Network	1.0	4.89	1.0	0.53
LSTM Stateful	1.46	**4.6**	0.93	**0.55**
Распознавание объектов из OTHER				
Multilayer Perceptron (alpha 97.4)	5.9	5.02	0.67	0.79
Fully Connected Network	1.0	7.21	1.0	0.65
LSTM (200 epoches)	1.35	**5.25**	0.97	**0.80**
Over-sampling (ADASYN)				
Multilayer Perceptron	18.69	13.44	0.11	0.24
Fully Connected Network	2.1	**12.93**	12.93	**0.29**
LSTM Stateful	5.45	13.38	0.77	0.30

Таблица 13 – Подбор нейронных сетей

Согласно критериям, определенным в параграфе 2.2.2, проанализируем результаты на исходных данных и данных без выбросов (**Таблица 14**). По умолчанию отнесём все объекты, которым не должно быть назначено меток, к категории 6. В метках многослойного перцептрона 8 раз появляется метка - "CE = 2.0 LRAP = 0.5", 6 раз - "CE = 3.0 LRAP = 0.33", 4 раза - "CE = 4.0 LRAP = 0.25", трижды - "CE = 5.0 LRAP = 0.2", 2 - "CE = 1.0 LRAP = 0.1" и 1 раз - "CE = 7.0 LRAP = 0.14" "CE = 6.0 LRAP = 0.16"". Назначим для результатов категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

В метках, назначенных полносвязной нейронной сетью, 4 раза появляется "CE = 3.0 LRAP = 0.33", 3 раза - "CE = 8.0 LRAP = 0.125", "CE = 14.0 LRAP = 0.7", "CE = 5.0 LRAP = 0.2", "CE = 7.0 LRAP = 0.14", дважды - "CE = 9.0 LRAP = 0.11", "CE = 13.0 LRAP = 0.7", и один раз - "CE = 4.0 LRAP = 0.25" "CE = 12.0 LRAP = 0.8" "CE = 10.0 LRAP = 0.1" "CE = 2.0 LRAP = 0.5". Эти результаты менее стабильны. Назначим для результатов категорию 6 только тем комбинациям метрик, которые повторяются 3 и более раза, а предыдущий критерий отменим.

Категории	**Multilayer Perceptron**	**Fully Connected Network**	**LSTM**
Исходные данные			
Категория 1	1	7	0
Категория 2	1	1	5
Категория 3	1	10	12
Категория 4	21	18	44
Категория 5	30	25	14
Категория 6	21	14	0
Исходные данные без выбросов			
Категория 1	1	5	-
Категория 2	1	2	-
Категория 3	1	8	-
Категория 4	16	15	-
Категория 5	30	44	-
Категория 6	26	1	-

Таблица 14 – Распределение оценок качества нейронных сетей

Интересны результаты LSTM – не наблюдается никакого отличия в результатах для объектов без меток. Кроме того, максимальная площадь под

кривой (Таблица 15) позволяет говорить о том, что эта архитектура работает наихудшим образом.

В результате, для исходных данных наблюдается то же явление, что и с алгоритмами машинного обучения: по паре метрикμ иμ_2наилучшие результаты показывает полносвязная сеть. Однако, площадь между кривыми реального количества меток и coverage error гораздо меньше у многослойного перцептрона. Если сравнить результаты результатами с алгоритмами машинного обучения (таблица 12), то вопрос выбора не возникает – выбранные архитектуры нейронных сетей значительно проигрывают и Bagging with Random Forest, Random Forest.

Алгоритм	μ_1	μ_2	S
Исходные данные			
Multilayer Perceptron	0.32	0.68	**572.5**
Fully Connected Network	**0.42**	**0.57**	581.5
LSTM	0.22	0.77	1128.5
Исходные данные без выбросов			
Multilayer Perceptron	**0.38**	**0.61**	**581.0**
Fully Connected Network	0.21	0.78	682.0
LSTM	-	-	-

Таблица 15– Метрики μ_1, μ_2 и S

Для работы с данными без выбросов, невозможно обучить statefull LSTM, поэтому было проведено сравнение двух архитектур. Многослойный перцептрон оказался более устойчив в выделении объектов без меток и таким образом оказался намного лучше полносвязных сетей, который оказались полностью неспособны предоставить критерий для различения объектов, в которых отсутствуют искомые вещества, и объектов с одним веществом.

2.3.3 Итоговая модель

Анализ, проведённый в предыдущих параграфах, показал, что хорошо работают подходы, в которых обучение разбивается на 2 задачи – детектировании двух наборов объектов против всех. Разовьем этот подход, разбив на три обучающих множества согласно плотности распределения:

- Группа 1: ДОФ против всех (9 против 31)

- Группа 2: ацетальдегид, ацетон, бензол, этилацетат, пластизоль против всех (17 против 21)
- Группа 3: бензин, бутанол, бутилацетат, гексан, изобутанол, изопропанол, пропанол, фенол, стирол, толуол, ДОФ с ацетальдегидом, ДОФ с ацетоном, ДОФ с бензолом, ДОФ с этилацетатом (20 к 20).

Это позволит уменьшить дисбаланс классов. Группы №1 и №3 уже рассмотрены в таблицах 10 и 13. Сравним алгоритмы для них и рассмотрим подход 2 (таблица 16). Задача различения ДОФ и прочих веществ решается наилучшим образом с помощью SVM, Random Forest и многослойного перцептрона – они показывают одинаковый результат. Однако, результат обучения многослойного перцептрона несколько хуже, остальные 2 алгоритма по этому параметру неразличимы. LSTM с сохранением состояний лучше всего решает задачу различения второй группы веществ, а Bagging с Random Forest – третьей.

Алгоритм	**CE(train)**	**CE(new)**	**LRAP(train)**	**LRAP(new)**
Группа 1				
Support Vector Machine	1.0	***1.42***	1.0	***0.97***
Bagging with Random Forest	1.22	*1.42*	0.88	*0.97*
Random Forest	1.0	***1.42***	1.0	***0.97***
Multilayer Perceptron (alpha 97.4)	1.07	*1.42*	0.96	*0.97*
Fully Connected Network	1.05	1.69	0.97	0.84
LSTM Stateful	1.05	1.62	0.97	0.87
Группа 2				
Support Vector Machine (linear)	1.0	**2.21**	1.0	**0.97**
Bagging with Random Forest	3.12	4.17	0.64	0.64
Random Forest	1.0	4.12	1.0	0.62
Multilayer Perceptron (alpha 97.4)	3.12	3.74	0.64	0.65
Fully Connected Network	1.12	4.09	0.97	0.58
LSTM Stateful	1.0	3.48	1.0	0.67
Группа 3				
Support Vector Machine (rbf)	5.55	5.82	0.69	0.80
Bagging with Random Forest	5.9	5.0	0.67	0.84
Random Forest	1.0	5.41	1.0	0.80
Multilayer Perceptron (alpha 97.4)	5.9	5.02	0.67	0.79
Fully Connected Network	1.0	7.21	1.0	0.65
LSTM (200 epoches)	1.35	5.25	0.97	0.80

Таблица 16 – Подбор алгоритмов для обнаружения объектов каждой группы

Проанализируем результаты. Для группы 1 запускались оба алгоритма – SVM и Random Forest. Меньшую площадь между кривыми реальных и назначаемых меток даёт метод опорных векторов – площадь в этом случае всего 4. Установим порог в 30% для попадания в решение первой группы. Для второй группы SVM с линейным ядром даёт площадь между кривыми. равную 29 – это неплохой результат, поскольку он явно ниже результатов на полном датасете. Установим порог вхождения в решение в 20% для веществ второй группы и 10% для третьей группы. Сводная таблица с результатами этого подхода и Bagging Random Forest, который обучался на всей выборке, представлены в Приложении 4 «Сравнительный анализ результатов обучение на трёх множествах и на всем обучающем множестве».

Интересный результат наблюдается для группы №2: присутствие пластизоли в обучающем множестве влияет на качество распознавания (см. рисунок 9). На рисунке 9 представлены матрицы ошибок справа – с пластизолью, слева – без. Мы можем позволить себе удалить этот класс из обучающего множества, так как за отсутствие известных нам веществ в ароматическом отпечатке игрушки мы можем теперь принять высокую вероятность other для всех групп при отсутствии иных классов.

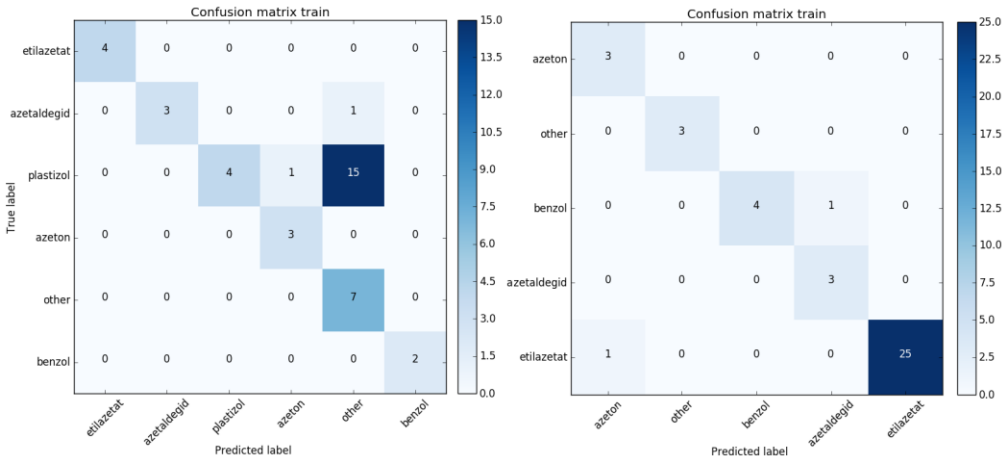


Рисунок 9 – Матрицы ошибок с пластизолью и без

Преимуществом разделения обучающего множества на 3 группы по «весу» классов является то, что появляется возможность обнаруживать «редкие» классы. Однако, качество распознавания существенно снижается для

всех групп. Частично за счет того, что мы назначаем много «лишних» классов, а также потому, что некоторые классы «выпадают» в силу изменения баланса весов. Кроме того, в ряде случаем удаётся сделать вывод о том, что известных нам веществ-маркеров в ароматическом отпечатке не обнаружено.

С другой стороны, при обучении на всём множестве мы стабильно обнаруживаем «частые» вещества – из-за того, что, несмотря на режим балансировки классов, алгоритм предсказывает их в соответствии с их плотностью распределения в обучающем множестве. Незначительно меняется порядок следования «частых» классов с одинаковым весом (ацетальдегида, этилацетата и бензола). При этом, «редкие» классы всегда остаются ниже порога в 10%. Даже если понизить порог до 8%, существенного улучшения не наблюдается.

Возникает вопрос: хотим ли мы «хорошо» предсказывать «частые» классы, которые присутствуют почти в каждом ароматическом отпечатке, но полностью игнорировать «редкие»? В случае положительного ответа на этот вопрос мы пропустим игрушки, которые содержат только редкие классы (например, 74 и 75). В случае отрицательного, нам удастся обнаружить некоторое количество редких веществ из третьей группы, однако существенно упадет производительность для второй группы.

ЗАКЛЮЧЕНИЕ

При написании данной выпускной квалификационной работы было проведено исследование 40 матриц откликов пьезоэлектрических сенсоров системы «Электронный нос» «МАГ-8», содержащих ароматические отпечатки 36 органических веществ-маркеров и 4 смесей, а также ароматических отпечатков 75 детских игрушек, изготовленных из полимерных материалов. Данные были получены на кафедре физической и аналитической химии Воронежского государственного университета инженерных технологий с помощью прибора «МАГ-8». Ставилась задача построения системы обработки и анализа данных для прибора «МАГ-8» на примере задачи детектирования наиболее полного набора представленных токсичных органических соединений в пробах игрушек. Основной интерес с точки зрения анализа данных представляла размерность обучающего множества, которое состояло из 40 многомерных объектов (матриц 121x8), содержало 20 классов и являлось несбалансированным.

В процессе исследования данных был проведен сравнительных анализ методов предварительной обработки и в результате был выстроен конвейер, состоящий из нормализации каждой отдельной матрицы с помощью центрирования и деления на стандартное отклонение, удаления тренда из каждого вектора матрицы методом скользящего среднего и масштабирования полученных матриц относительно друг друга.

Был исследован ряд подходов к балансировке обучающего множества, включающие как работу с исходными данным, так и методы генерации искусственных объектов с помощью алгоритма ADASYN и сравнительного новой архитектуры искусственных нейронных сетей GAN. Сделан вывод о том, что рассмотренные подходы не дают удовлетворительного решения проблемы несбалансированности тренировочного массива данных. На основании этого вывода было принято решение перейти к обучению исключительно на исходном массиве данных.

На втором шаге конвейера ставилась задача классификации на 20 классов, в результате ожидалось получить список веществ-маркеров и их смесей, обнаруженных в ароматическом отпечатке каждой из игрушек. В результате было выработано два подхода: один предполагал балансировку массива данных через разделение обучающего множества на три группы по степени представленности в обучающем множестве (группа 1 - ДОФ как наиболее представленный, затем, в группе 2, классы, которые представлены 2-4 объектами, и в группе №3 – «единственные» объекты, прочие объекты помечались общей меткой other), второй подход – обучение на всем массиве данных. Был произведен отбор алгоритмов для каждой группы – наилучшим образом себя показал метод опорных векторов. Оба подхода имеют свои достоинства – так, первый подход, несмотря на низкое качество распознавания позволяет всё же находить «редкие» классы, в то время как второй подход стабильно обнаруживает объекты 1 и 2 групп, и, поскольку именно этих веществ-маркеров – подавляющее большинство среди истинных меток, то создаётся иллюзия высокой точности распознавания.

Однако, в сравнении с эталонными значениями, полученными на кафедре физической и аналитической химии с помощью графического метода, оба эти подхода в значительной степени проигрывают. Это объясняется тем, что половину обучающего множества составляют объекты третьей группы, а обучение на единственных примерах затруднительно.

На основе изучения специальной литературы, разработки и реализации практической части были сделаны следующие выводы:

1. Сильная несбалансированность и небольшой размер обучающего множества являются помехой для качественного обучения моделей
2. В отличие от графического метода, максимальные изменения частоты колебаний пьезокварцевого датчика не несут никакой уникальной информации, необходимой для надёжного различения объектов разных классов

3. Вопреки [21], где утверждается, что нестационарность несёт в себе больше информации, в нашем случае именно удаление тренда с помощью фильтра скользящих средних позволяет существенно улучшить качество распознавания
4. Линейные методы понижения размерности не применимы

Для дальнейшего развития работы существуют следующие пути:

1. Поиск других методов генерации искусственных данных
2. Углубление исследования методов предварительной обработки для повышения расстояния между объектами разных классов

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

СПИСОК ЛИТЕРАТУРЫ

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

2. Gardner, J.W., Bartlett, P.N., Electronic Noses: Principles and Applications. — Oxford University Press: New York. — NY, USA. — 1999.

3. Gardner J.W., Bartlett, P.N., A brief history of electronic noses. Sens. Actuat. B: Chem. —№18. — 1994.

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

5. Li S., Overview of Odor Detection Instrumentation and the Potential for Human Odor Detection in Air Matrices, — 2009.

6. 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.

7. Wilson A.D., Baietto M., Applications and Advances in Electronic—Nose Technologies. — Sensors. — №9. — 2009.

8. 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.

9. The Cyranose 320 E— nose User Manual, Smiths Detection // User Manual.

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

11. Кучменко Т.А., Лисицкая Р.П., Шуба А.А., Информативность анализатора газов «электронный нос» для оценки качества вина. — Аналитика и контроль. — №4. — 2014.

12. 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.

13. 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.

14. Scott, S., James, D. & Ali, Z, Data analysis for electronic nose systems, — Microchim Acta. —2006.

15. Dickson J.A., 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. — pp. 162-165. — June 2000.

16. James, D., Scott, S.M., Zulfiqur, A., O'Hare, W.T., Chemical sensors for electronic nose systems. — Microchimica Acta. №149. – pp. 1-17. — 2005.

17. 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.

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

19. 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

20. 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.

21. Yan J., Guo X., Duan S., Jia P., Wang L., Peng C., Zhang S., Electronic Nose Feature Extraction Methods: A Review. – Sensors. — №11. — 2015.

22. Kim E.G., Lee S, Kim J.H., Kim C., Byun Y.T., Pattern Recognition for Selective Odor Detection with Gas Sensor Arrays. — Sensors. — №12. — 2012.

23. Ziyatdinov A., Perera-Lluna A., Data Simulation in Machine Olfaction with the R Package Chemosensors. — PLoS ONE. — №9 (2). — 2013.

24. Ziyatdinov A., Perera A., Synthetic benchmarks for machine olfaction: Classification, segmentation and sensor damage. — Data in Brief. — №3. — pp. 126-130. — 2014.

25. Дроздова Е. В., 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

26. Scikit-learn: Model evaluation: quantifying the quality of predictions // Scikit-learn [Электронный ресурс]. — Режим доступа: http://scikit-learn.org/stable/modules/model_evaluation.html (Дата обращения: 02.04.2017)

27. Lemaitre G., Nogueira F., Aridas C. K., Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning // imblearn-Scikit-learn [Электронный ресурс]. — Режим доступа: http://contrib.scikit-learn.org/imbalanced-learn/about.html (Дата обращения: 12.04.2017)

28. He H., Bai Y., Garcia E.A., Li S., ADASYN: adaptive synthetic sampling approach for imbalanced learning // Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN'08). — pp. 1322-1328. — 2008.

29. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley,D., Ozair, S., Courville, A., Bengio, Y., Generative Adversarial Networks // Cournell University Library: arXiv.org [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1406.2661. (Дата обращения: 02.04.2017). — 2014.

30. Radford A., Metz L., Chintala S., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks // Cournell University Library: arXiv.org [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1511.06434. (Дата обращения: 02.04.2017). — 2016.

31. Chen X., Duan Y., Houthooft R., Schulman J., Sutskever I., Abbeel P., InfoGAN: Interpretable Representation Learning by Information Maximizing Generative

Adversarial Nets // Cournell University Library: arXiv.org [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1606.03657. (Дата обращения: 02.04.2017). — 2015.

32. Kim N., A tensorflow implementation of GAN (exactly InfoGAN or Info GAN) to one dimensional (1D time series data) // GitHub repository [Электронный ресурс]. — Режим доступа: https://github.com/buriburisuri/timeseries_gan. (Дата обращения: 02.04.2017). — 2016.

33. Mirza M., Osindero S., Conditional Generative Adversarial Nets // Cournell University Library: arXiv.org [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1411.1784. (Дата обращения: 30.04.2017). — 2014.

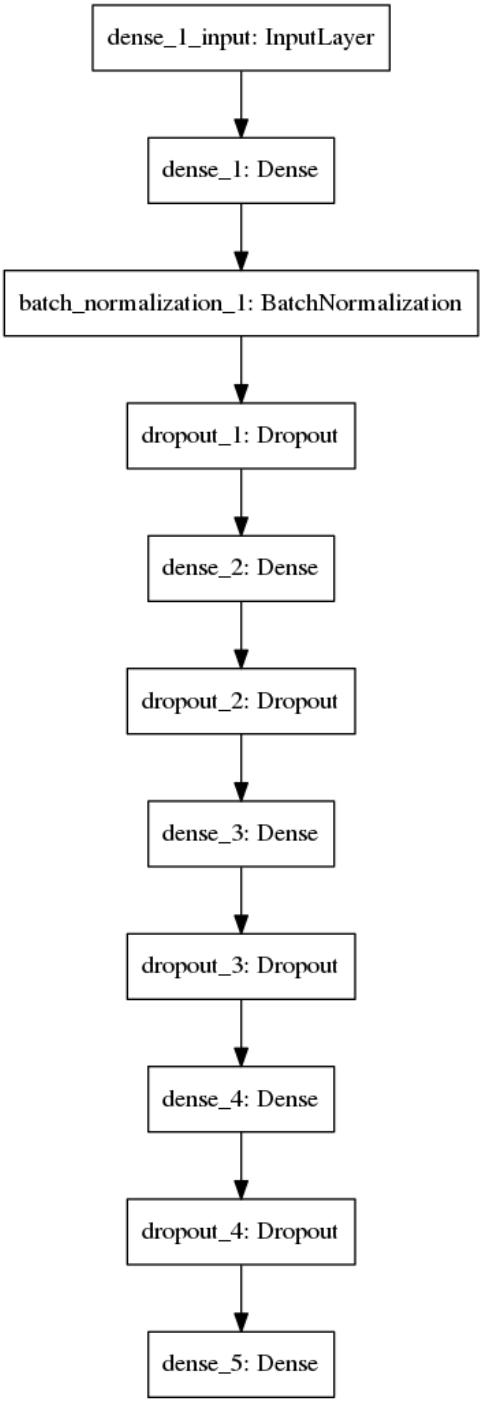
34. Salimans T., Goodfellow I., Zaremba W., Cheung V., Radford A., Checn X., Improved Techniques for Training GANs // Cournell University Library: arXiv.org [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1606.03498 . (Дата обращения: 30.04.2017). — 2016.

35. Multiclass and multilabel algorithms // Scikit-learn. [Электронный ресурс]. — Режим доступа: http://scikit-learn.org/stable/modules/multiclass.html. (Дата обращения: 11.04.2017).

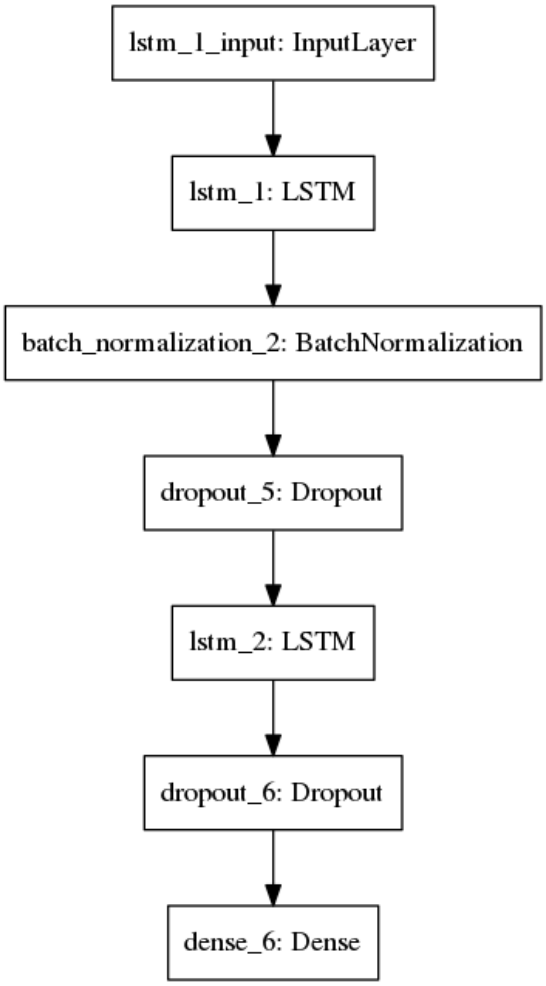
ПРИЛОЖЕНИЕ 1 «Некоторые коммерческие системы «Электронный нос»: модели и технологии».

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

ПРИЛОЖЕНИЕ 2 Архитектура полносвязной нейронной сети



ПРИЛОЖЕНИЕ 3 Архитектура LSTM



ПРИЛОЖЕНИЕ 4 Сравнительный анализ итоговых моделей

№	Группа 1 (20%)	Группа 2 (10%)	Группа 3 (10%)	Результат	Истинное значение	Общий алгоритм (8%)	Истинное значение
1	Other - 60% ДОФ – 28%	Other – 60% Ацетон — 41% Бензол — 13% Ацетальдегид — 11%	Other – 60% Бутилацетат — 13% ДОФ с бензолом — 13% Изобутанол — 10%	ДОФ Ацетон Бутилацетат ДОФ с бензолом Изобутанол	**ДОФ** **ацетон**	ДОФ – 21% Ацетальдегид – 14% Этилацетат – 12% Бензол – 12% Ацетон – 10%	**ДОФ** **ацетон**
2	Other - 64% ДОФ – 22%	Other – 63%	Other – 65% ДОФ с бензолом — 11%	ДОФ ДОФ с бензолом	**ДОФ** ацетон	ДОФ – 24% Ацетальдегид – 13% Этилацетат – 13% Бензол – 11% Ацетон – 8%	**ДОФ** **ацетон**
3	Other – 84%	Other – 61% Этилацетат – 41% Бензол — 12% Ацетальдегид — 10%	Other – 61%	Этилацетат Бензол Ацетальдегид	-	ДОФ – 21% Ацетальдегид – 12% Бензол – 9% Этилацетат – 9%	-
4	Other – 76%	Other – 62% Ацетон — 14% Бензол — 14%	Other – 55% *Бутилацетат — 9%* *Толуол — 3%*	Ацетон Бензол *Бутилацетат* *Толул*	Изобутанол Изопропанол **Толуол** **Бутилацетат** ДОФ с этилацетатом ДОФ с Ацетальдегидом	ДОФ – 18% Бензол – 12% Этилацетат – 10% Ацетальдегид – 10% Ацетон – 8%	Изобутанол Изопропанол Толуол Бутилацетат ДОФ с этилацетатом ДОФ с Ацетальдегидом
5	Other – 74% ДОФ – 22%	Other – 60% Этилацетат — 17%	Other – 54%	ДОФ Этилацетат	Изобутанол Изопропанол	ДОФ – 20% Бензол – 11%	Изобутанол Изопропанол

			Бутилацетат — 9%	*Бутилацетат*	**Толуол** **Бутилацетат** ДОФ с этилацетатом ДОФ с Ацетальдегидом	Ацетальдегид – 10% Этилацетат – 10% Ацетон – 8%	Толуол Бутилацетат ДОФ с этилацетатом ДОФ с Ацетальдегидом
6	Other – 64%	Other – 61%	Other – 65%	-	-	ДОФ – 22% Этилацетат – 13% Ацетальдегид – 10% Бензол – 10% Ацетон – 9%	-
7	Other – 57% ДОФ – 26%	Other – 61% Ацетон — 11%	Other – 70%	ДОФ Ацетон	-	ДОФ – 23% Этилацетат – 11% Ацетальдегид – 10% Бензол – 10% Ацетон – 8%	-
8	Other – 67%	Other – 61% Ацетон — 23% Ацетальдегид — 10%	Other – 54%	Ацетон Ацетальдегид	-	ДОФ – 22% Ацетальдегид – 12% Этилацетат – 11% Бензол – 10% Ацетон – 9%	-
9	Other – 77%	Other – 59% Этилацетат — 23%	Other – 51%	Этилацетат	-	ДОФ – 22% Бензол – 11% Ацетальдегид – 11% Этилацетат – 11% Ацетон – 10%	-
10	Other – 75%	Other – 61% Ацетон — 15%	Other – 60%	Ацетон	ДОФ Толуол Бензол Ацетальдегид **Ацетон**	ДОФ – 20% Этилацетат – 10% Ацетальдегид – 10% Бензол – 10%	**ДОФ** Толуол **Бензол** **Ацетальдегид** **Ацетон**

						Ацетон – 8%	
11	Other – 56% ДОФ — 27%	Other – 62% Бензол — 11%	Other – 71%	ДОФ Бензол	**ДОФ** Толуол **Бензол** Ацетальдегид Ацетон	ДОФ – 19% Бензол – 10% Этилацетат – 10% Ацетальдегид – 10% *Ацетон – 7%*	**ДОФ** Толуол **Бензол** **Ацетальдегид** ***Ацетон***
12	Other – 50% ДОФ – 26%	Other – 61% Этилацетат — 14% Бензол — 11%	Other – 71%	ДОФ Этилацетат Бензол	ДОФ с ацетоном **ДОФ** ДОФ с Ацетальдегидом	ДОФ – 22% Ацетальдегид – 10% Бензол – 10% Этилацетат – 10% Ацетон – 9%	ДОФ с ацетоном **ДОФ** ДОФ с ацетальдегидом
13	Other – 47% ДОФ – 30%	Other – 60%	Other – 52% Изопропанол — 12% Стирол — 10% ДОФ с бензолом — 10%	ДОФ Изопропанол Стирол ДОФ с бензолом	-	ДОФ – 24% Этилацетат – 13% Ацетальдегид – 11% Бензол – 12% Ацетон – 10%	-
14	Other – 75%	Other – 63% Ацетон — 12% Бензол — 12%	Other – 62% *Бутилацетат — 3%* *ДОФ с ацетоном — 2%*	Ацетон Бензол *Бутилацетат* *ДОФ с ацетоном*	***Бутилацетат*** ***ДОФ с ацетоном*** ДОФ с ацетальдегидом Толуол **Бензол** Фенол	ДОФ – 18% Ацетальдегид – 11% Этилацетат – 10% Бензол – 10%	Бутилацетат ДОФ с ацетоном ДОФ с ацетальдегидом Толуол **Бензол** Фенол
15	Other – 68%	Other – 61% Ацетон — 13% Ацетальдегид — 10%	Other – 62%	Ацетон Ацетальдегид	ДОФ с ацетоном	ДОФ – 19% Бензол – 11% Ацетальдегид – 10% Этилацетат – 9%	ДОФ с ацетоном
16	Other – 74%	Other – 62%	Other – 68%	-	ДОФ Ацетон	ДОФ – 19% Ацетальдегид – 10% Этилацетат – 9%	**ДОФ** **Ацетон**

						Бензол – 8% Ацетон – 8%	
17	Other – 42% ДОФ — 34%	Other – 60%	Other – 59% Бутилацетат — 12%	ДОФ Бутилацетат	-	ДОФ – 22% Ацетальдегид – 12% Этилацетат – 12% Бензол – 11% Ацетон – 10%	-
18	Other – 62% ДОФ — 28%	Other – 61% Ацетон – 29% Ацетальдегид — 11% Бензол — 10% Этилацетат — 10%	Other — 63% Бутанол — 10%	ДОФ Ацетон Ацетальдегид Бензол Этилацетат Бутанол	ДОФ с этилацетатом ДОФ с ацетальдегидом Гексан **Ацетон**	ДОФ – 20% Ацетальдегид – 13% Этилацетат – 10% Бензол – 10% Ацетон – 8%	ДОФ с этилацетатом ДОФ с ацетальдегидом Гексан **Ацетон**
19	Other – 55%	Other – 62% Этилацетат – 34%	Other – 67%	Этилацетат	-	ДОФ – 21% Ацетальдегид – 13% Бензол – 11% Этилацетат – 10% Ацетон – 9%	-
20	Other – 45% ДОФ — 40%	Other – 59% Этилацетат – 34% Бензол — 14% Ацетальдегид — 12%	Other – 65%	ДОФ Этилацетат Бензол Ацетальдегид	**ДОФ**	ДОФ – 21% Этилацетат – 11% Ацетальдегид – 11% Бензол – 10% Ацетон – 9%	**ДОФ**
21	Other – 53% ДОФ — 49%	Other – 63% Ацетон — 37 Бензол — 25% Этилацетат — 16% Ацетальдегид — 16%	Other – 73%	ДОФ Ацетон Бензол Этилацетат Ацетальдегид	**ДОФ**	ДОФ – 23% Ацетальдегид – 13% Бензол – 12% Этилацетат – 11% Ацетон – 9%	**ДОФ**
22	Other – 56% ДОФ — 23%	Other – 59%	Other – 69%	ДОФ	**ДОФ**	ДОФ – 20% Этилацетат – 10%	**ДОФ**

						Ацетальдегид – 10% Ацетон – 9% Бензол – 9%	
23	Other – 64%	Other – 59% Ацетон — 19% Этилацетат — 18%	Other – 55%	Ацетон Этилацетат	ДОФ с бутилацетатом этилацетатом ацетальдегидом	ДОФ – 20% Этилацетат – 11% Ацетальдегид – 11% Бензол – 10% Ацетон – 8%	ДОФ с бутилацетатом этилацетатом ацетальдегидом
24	Other – 76%	Other – 63% Бензол — 15%	Other – 64%	Бензол	ДОФ с этилацетатом Ацетон	ДОФ – 20% Ацетальдегид – 11% Бензол – 11% Этилацетат – 10% Ацетон – 8%	ДОФ с этилацетатом **Ацетон**
25	Other – 73%	Other – 61%	Other – 60%	-	ДОФ	ДОФ – 20% Ацетальдегид – 10% Этилацетат – 9% Бензол – 9%	**ДОФ**
26	Other – 54% ДОФ — 24%	Other – 58% Ацетон — 22%	Other – 67%	-	-	ДОФ – 21% Этилацетат – 11% Ацетальдегид – 11% Бензол – 9% Ацетон – 8%	-
27	Other – 42% ДОФ — 35%	Other – 59% Этилацетат — 29%	Other – 62%	-	-	ДОФ – 24% Ацетальдегид – 12% Этилацетат – 11% Бензол – 10% Ацетон – 10%	-
28	Other – 69%	Other – 61% Этилацетат — 13%	Other – 66%	Этилацетат	Гексан Толуол	ДОФ – 21% Бензол – 11%	Гексан Толуол

					Бутилацетат Бензол	Ацетальдегид – 11% Этилацетат – 10% Ацетон – 8%	Бутилацетат **Бензол**
29	Other – 67%	Other – 61%	Other – 68%	-	Гексан Толуол Бутилацетат Бензол	ДОФ – 20% Этилацетат – 10% Ацетальдегид – 10% Бензол – 9%	Гексан Толуол Бутилацетат **Бензол**
30	Other – 45% ДОФ — 29%	Other – 62%	Other – 69%	ДОФ	Толуол **ДОФ** Бензол	ДОФ – 24% Этилацетат – 11% Ацетальдегид – 9% Бензол – 9% Ацетон – 8%	Толуол **ДОФ** **Бензол**
31	Other – 64% ДОФ — 24%	Other – 62% Этилацетат — 15% Бензол — 13% Ацетон — 12% Ацетальдегид — 11%	Other – 65%	ДОФ Этилацетат Бензол Ацетон Ацетальдегид	**ДОФ** **Ацетон**	ДОФ – 20% Ацетальдегид – 12% Бензол – 12% Этилацетат – 11% Ацетон – 8%	**ДОФ** **Ацетон**
32	Other – 77%	Other – 61% Ацетон — 10%	Other – 69%	Ацетон	ДОФ **Ацетон**	ДОФ – 19% Этилацетат – 11% Бензол – 9% Ацетальдегид – 9%	**ДОФ** Ацетон
33	Other – 82%	Other – 61% Бензол — 14% Ацетон — 13%	Other – 62%	Бензол Ацетон	ДОФ с этилацетатом ДОФ с ацетальдегидом толуол **бензол** фенол	ДОФ – 18% Бензол – 12% Этилацетат – 12% Ацетальдегид – 9%	ДОФ с этилацетатом ДОФ с ацетальдегидом толуол **бензол** фенол

34	Other – 67%	Other – 60% Ацетон — 18% Ацетальдегид — 10%	Other – 63%	Ацетон Ацетальдегид	Этилацетат ДОФ **Ацетон** Изобутиловый спирт	ДОФ – 20% Этилацетат – 11% Бензол – 10% Ацетальдегид – 9%	**Этилацетат** **ДОФ** **Ацетон** Изобутиловый спирт
35	Other – 62%	Other – 62%	Other – 70% *Фенол – 5%*	Фенол	Этилацетат Ацетальдегид Ацетон Толуол Бензол ДОФ **Фенол**	ДОФ – 19% Этилацетат – 11% Ацетальдегид – 10% Бензол – 10% Ацетон – 8%	**Этилацетат** **Ацетальдегид** **Ацетон** Толуол **Бензол** **ДОФ** Фенол
36	Other – 39%	Other – 60% Ацетон — 11%	Other – 64%	Ацетон	**Ацетон** Бензол	ДОФ – 22% Этилацетат – 12% Бензол – 9% Ацетальдегид – 9% Ацетон – 8%	**Ацетон** **Бензол**
37	Other – 65%	Other – 62%	Other – 63%	-	Ацетон ДОФ	ДОФ – 19% Этилацетат – 11% Бензол – 11% Ацетальдегид – 10% Ацетон – 8%	**Ацетон** **ДОФ**
38	Other – 60%	Other – 60% Ацетальдегид — 13%	Other – 63%	Ацетальдегид	Ацетон ДОФ	ДОФ – 22% Этилацетат – 13% Бензол – 11% Ацетальдегид – 10% Ацетон – 8%	**Ацетон** **ДОФ**
39	Other – 66%	Other – 64% Ацетальдегид — 18%	Other – 68%	Ацетальдегид	-	ДОФ – 20% Бензол – 11% Этилацетат – 10% Ацетальдегид – 9%	-

40	Other – 69%	Other – 61%	Other – 69%	-	Керосин ДОФ Ацетальдегид Бензол Изопропанол	ДОФ – 19% Бензол – 10%	*Керосин* **ДОФ** Ацетальдегид **Бензол** Изопропанол
41	Other – 51% ДОФ — 38%	Other – 60% Ацетон — 27%	Other – 58% Стирол — 12%	ДОФ Ацетон Стирол	-	ДОФ – 23% Бензол – 12% Этилацетат – 10% Ацетальдегид – 9%	-
42	Other – 62%	Other – 61%	Other – 64%	-	-	ДОФ – 21% Бензол – 10% Этилацетат – 8% Ацетальдегид – 8%	-
43	Other – 69% ДОФ — 20%	Other – 62% Этилацетат — 12%	Other – 70%	ДОФ Этилацетат	-	ДОФ – 20% Бензол – 10% Ацетальдегид – 9%	-
44	Other – 63%	Other – 60% Ацетальдегид — 14% Ацетон — 12%	Other – 66%	Ацетальдегид Ацетон	-	ДОФ – 20% Бензол – 9% Ацетальдегид – 9% Этилацетат – 8%	-
45	Other – 58%	Other – 61%	Other – 62%	ДОФ Ацетон	-	ДОФ – 20% Бензол – 10% Ацетальдегид – 9% Этилацетат – 8%	-
46	Other – 71% ДОФ — 20%	Other – 62% Ацетон — 13%	Other – 64%	ДОФ Ацетон	-	ДОФ – 19% Бензол – 9% Ацетальдегид – 9% Этилацетат – 8%	-

47	Other – 57%	Other – 63% Ацетон — 20% Ацетальдегид — 14%	Other – 70%	Ацетон Ацетальдегид	**Ацетон** ДОФ **Ацетальдегид**	ДОФ – 18% Бензол – 9% Ацетальдегид – 9% Этилацетат – 9%	Ацетон **ДОФ** **Ацетальдегид**
48	Other – 71% ДОФ — 21%	Other – 60% Ацетон — 31%	Other – 58%	ДОФ Ацетон	**Ацетон** **ДОФ** Ацетальдегид	ДОФ – 20% Бензол – 10% Этилацетат – 9% Ацетальдегид – 8%	Ацетон **ДОФ** **Ацетальдегид**
49	Other – 59% ДОФ — 20%	Other – 61%	Other – 59%	ДОФ	Гексан ДОФ с бензолом ДОФ с ацетальдегидом Толуол Бензол Стирол Ацетальдегид	ДОФ – 19% Бензол – 10% Этилацетат – 9% Ацетальдегид – 8%	Гексан ДОФ с бензолом ДОФ с ацетальдегидом Толуол **Бензол** Стирол **Ацетальдегид**
50	Other – 69%	Other – 61% Этилацетат — 13% Ацетон — 10%	Other – 64%	Этилацетат ДОФ	Гексан **ДОФ** Бутилацетат Бензол	ДОФ – 19% Бензол – 9% Этилацетат – 9% Ацетальдегид – 8%	Гексан **ДОФ** Бутилацетат **Бензол**
51	Other – 67% ДОФ — 28%	Other – 61% Бензол — 14% Этилацетат — 13% Ацетон — 11%	Other – 69%	ДОФ Бензол Этилацетат Ацетон	Гексан **ДОФ** Бутилацетат **Бензол**	ДОФ – 20% Бензол – 12% Этилацетат – 9% Ацетальдегид – 8%	Гексан **ДОФ** Бутилацетат **Бензол**
52	Other – 75%	Other – 63%	Other – 60%	-	-	ДОФ – 22% Бензол – 11% Этилацетат – 10% Ацетальдегид – 10%	-
53	Other – 57% ДОФ — 30%	Other – 62% Этилацетат — 11%	Other – 60% Толуол — 11%	ДОФ Этилацетат	-	ДОФ – 20% Бензол – 11%	-

				Толуол		Этилацетат – 9% Ацетальдегид – 8%	
54	Other – 63%	Other – 63%	Other – 64%	-	-	ДОФ – 22% Бензол – 11% Ацетальдегид – 9% Этилацетат – 9%	-
55	Other – 70%	Other – 64% Бензол — 10% Ацетальдегид — 10%	Other – 63%	Бензол Ацетальдегид	Этилацетат **Ацетальдегид** Бутилацетат Изопропанол **Бензол** Толуол Диоктилфталат	ДОФ – 19% Бензол – 10% Ацетальдегид – 9% Этилацетат – 8%	**Этилацетат** **Ацетальдегид** Бутилацетат Изопропанол **Бензол** Толуол **Диоктилфталат**
56	Other – 85%	Other – 62% Ацетон — 42% Ацетальдегид — 12%	Other – 60%	Ацетон Ацетальдегид	-	ДОФ – 19% Бензол – 13% Этилацетат – 10% Ацетальдегид – 8%	-
57	Other – 50% ДОФ — 26%	Other – 60% Этилацетат — 11%	Other – 61%	ДОФ Этилацетат	-	ДОФ – 21% Бензол – 11% Ацетальдегид – 10% Этилацетат – 9%	-
58	Other – 83%	Other – 61% Ацетон — 24% Этилацетат — 16% Ацетальдегид — 12%	Other – 56%	Ацетон Этилацетат Ацетальдегид	**Этилацетат** Изобутиловый спирт	ДОФ – 18% Бензол – 10% Ацетальдегид – 8%	Этилацетат Изобутиловый спирт
59	Other – 81%	Other – 63% Бензол — 13% Ацетальдегид — 10%	Other – 59%	Бензол Ацетальдегид	-	ДОФ – 22% Бензол – 12% Этилацетат – 10% Ацетальдегид – 9%	-

60	Other – 60%	Other – 61% Ацетон — 15% Ацетальдегид — 10%	Other – 60%	Ацетон Ацетальдегид	Бутилацетат ДОФ	ДОФ – 21% Бензол – 10% Этилацетат – 9% Ацетальдегид – 8%	Бутилацетат **ДОФ**
61	Other – 66%	Other – 62% Ацетальдегид — 16% Ацетон — 13%	Other – 66%	Ацетальдегид Ацетон	-	ДОФ – 21% Бензол – 11% Этилацетат – 9% Ацетальдегид – 9%	-
62	Other — 59% ДОФ — 21%	Other – 61% Этилацетат — 31% Бензол — 11%	Other – 58%	ДОФ Этилацетат Бензол	-	ДОФ – 21% Бензол – 11% Ацетальдегид – 9%	-
63	Other – 59%	Other – 60% Этилацетат - 20%	Other – 50%	Этилацетат	ДОФ Ацетон Ацетальдегид **Этилацетат** Бутилацетат	ДОФ – 18% Бензол – 8%	**ДОФ** Ацетон Ацетальдегид Этилацетат Бутилацетат
64	Other – 57%	Other — 60% Этилацетат - 20% Бензол — 18% Ацетальдегид — 14%	Other – 64%	Этилацетат Бензол Ацетальдегид	**Этилацетат** **Ацетальдегид**	ДОФ – 20% Бензол – 10% Ацетальдегид– 8% Этилацетат – 8%	**Этилацетат** **Ацетальдегид**
65	ДОФ – 35% Other – 31%	Other – 61%	Other – 70%	ДОФ	Этилацетат Бензол Толуол **ДОФ**	ДОФ – 23% Бензол – 10% Ацетальдегид– 9% Этилацетат – 9%	**Этилацетат** **Бензол** Толуол **ДОФ**
66	Other – 84%	Other – 63% Бензол — 10%	Other – 69%	ДОФ	Ацетальдегид Бутилацетат Толуол Бензол Изобутанол Керосин	ДОФ – 17% Бензол – 10% Ацетальдегид– 8%	**Ацетальдегид** Бутилацетат Толуол **Бензол** Изобутанол *Керосин*

67	Other – 84%	Other – 62% Ацетон — 14% Бензол — 11% Ацетальдегид — 10%	Other – 65%	Ацетон Бензол Ацетальдегид	**Бензол** **Ацетальдегид** Толуол ДОФ	ДОФ – 16% Бензол – 10% Ацетальдегид– 8% Этилацетат – 8%	**Бензол** **Ацетальдегид** Толуол **ДОФ**
68	Other – 83%	Other – 63% Ацетальдегид — 12%	Other – 69%	Ацетальдегид	Фенол Ацетон **Ацетальдегид** Толуол Бензол ДОФ	ДОФ – 15% Бензол – 11% Ацетальдегид– 8% Этилацетат – 8%	Фенол **Ацетон** **Ацетальдегид** Толуол **Бензол** **ДОФ**
69	Other – 76%	Other – 62% Бензол — 14% Этилацетат — 14% Ацетон — 13% Ацетальдегид — 12%	Other – 73%	Бензол Этилацетат Ацетон Ацетальдегид	Фенол Толуол **Бензол** Ацетон	ДОФ – 15% Бензол – 11% Ацетальдегид– 8%	Фенол Толуол **Бензол** Ацетон
70	Other – 54% ДОФ – 33%	Other – 62% Этилацетат — 36% Ацетальдегид — 13% Бензол — 11%	Other – 69%	ДОФ Этилацетат	Стирол Ацетальдегид **Этилацетат** Бутилацетат Бензол **ДОФ**	ДОФ – 21% Бензол – 10% Ацетальдегид– 9% Этилацетат – 8%	Стирол **Ацетальдегид** **Этилацетат** Бутилацетат **Бензол** **ДОФ**
71	Other – 74%	Other – 61%	Other — 64% *Изопропанол — 3%*	*Изопропанол — 3%*	Изобутанол Ацетальдегид Этилацетат Бутилацетат ***Изопропанол*** ДОФ	ДОФ – 18% Бензол – 10% Ацетальдегид– 8% Этилацетат – 8%	Изобутанол **Ацетальдегид** **Этилацетат** Бутилацетат Изопропанол **ДОФ**
72	Other – 81%	Other – 62%	Other — 65%	-	Толуол Ацетальдегид ДОФ	ДОФ – 16% Бензол – 11% Ацетальдегид– 8% Этилацетат – 8%	Толуол **Ацетальдегид** **ДОФ**
73	Other – 80%	Other – 62%	Other — 59%	-	ДОФ Изопропанол	ДОФ – 18% Бензол – 11%	**ДОФ** Изопропанол

						Ацетальдегид– 8% Этилацетат – 8%	
74	Other – 50%	Other – 60% Этилацетат — 43% Ацетон — 13% Ацетальдегид — 12%	Other — 56%	Этилацетат Ацетон Ацетальдегид	бутанол изобутанол пропанол изопропанол	ДОФ – 21% Бензол – 12% Ацетальдегид– 10% Этилацетат – 9%	бутанол изобутанол пропанол изопропанол
75	Other – 74%	Other – 63% Этилацетат — 29% Бензол — 16% Ацетальдегид — 12%	Other — 55% *Изобутанол — 9%* *Пропанол — 8%* *Изопропанол — 7%*	Этилацетат Бензол Ацетальдегид *Изобутанол* *Пропанол* *Изопропанол*	бутанол ***изобутанол*** ***пропанол*** ***изопропанол***	ДОФ – 19% Бензол – 11% Ацетальдегид–9% Этилацетат – 8%	бутанол *изобутанол* *пропанол* *изопропанол*