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 31 пункта.

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

Такие системы строятся на основе массива сенсоров нескольких видов, высокочувствительных к заданному набору соединений и веществ. Количество и вид сенсоров в этих системах может варьироваться в зависимости от решаемой задачи. В качестве выходных данных, доступных для дальнейшей обработки и анализа, прибор предоставляет матрицу изменений откликов сенсоров, время исследования, рабочие частоты сенсоров и некоторую служебную информацию. Однако, интерпретация результатов требует значительного времени и присутствия эксперта.

Большой интерес в настоящее время представляет разработка интегрированных аналитических систем, образующих единый конвейер, начиная с измерения, через обработку и интеллектуальный анализ данных и

заканчивая принятием решения. При этом, необходимо учитывать специфику подобных исследований: сначала исследователь определяется с набором веществ-маркеров, с которыми он планирует работать, исходя из этого подбирает селективные покрытия датчиков, затем тестирует этот массив, проводя первичные пробы отдельных веществ и их возможных смесей, и только после этого приступает к оценке тех объектов, которым посвящено исследование. Таким образом, несмотря на то, что методы анализа данных могут многое рассказать о каждом веществе из набора, применять их к отдельным матрицам откликов сенсоров нецелесообразно из-за большого числа этих матриц. Этот этап должен быть либо скрыт от конечного пользователя, либо от него следует отказаться в пользу более высокоуровневых подходов. В качестве основного подхода здесь может быть использовано решение задачи бинарной классификации для отдельных веществах и их смесей на этапе обучения и решение задачи классификации на N классов для новых объектов, которые всегда в своем «ароматическом отпечатке» будут содержать следы многих веществ в непредсказуемых концентрациях, часть из которых никогда не будет включена в обучающую выборку по причине ограниченности временных ресурсов исследователя и, как следствие, узости решаемой им задачи.

Новизна и актуальность данного исследования следуют из уникальности источника данных — прибора «МАГ-8». В данный момент группой разработчиков «МАГ-8» используется исключительно графический метод анализа. Данный подход успешно применяется при решении широкого круга задач – от анализа качества колбасных изделий до диагностики некоторых заболеваний. Недостатком данного подхода является трудоемкость и требование к высокой квалификации человека, который интерпретирует результаты. Данная работа позволит шагнуть от этапа измерений до этапа интерпретации результата, минуя рутинный анализ отдельных маркеров и их смесей, а также требующий повышенного внимания этап анализа объектов, на которые нацелено исследование.

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

1. Подбор способа расширить обучающее множество путем размножения наименее представленных объектов ИЛИ Подбор способа предобработки сравнительно небольшого и несбалансированного массива исходных данных
2. Подбор и обучение алгоритма бинарной классификации веществ-маркеров и их смесей
3. Решение задачи обнаружения веществ-маркеров и их смесей в пробах 75 детских игрушек

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

Таким образом, данная работа находится на стыке аналитической химии и машинного обучения. Логика исследования обуславливает структуру работы, состоящую из введения, двух глав, заключения, библиографии и приложения. В 1 главе сделан обзор предметной области [] . Вторая глава посвящена рассмотрению подходов к анализу данных прибора «МАГ-8» и построению прототипа системы анализа данных [] . В заключении подводятся итоги исследования и рассматриваются направления дальнейшего развития. Приложения содержат

1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Архитектура и обзор систем «Электронный нос»

1.1.1 Принципы построения систем «Электронный нос»

Концепция «электронного носа” как инструмента, состоящего из датчиков, используемых для классификации запахов, была впервые введена Персо и Доддом в 1982 году [1]. В ходе своих экспериментов эти два исследователя поставили перед собой цель создать инструмент, способный эмулировать обонятельную систему млекопитающих, распознавая различные запахи и давая воспроизводимые результаты. В частности, разработанный ими электронный нос состоял из: (I) матрицы датчиков для имитации рецепторов обонятельной системы человека; (II) блок обработки данных, который будет выполнять ту же функцию, что и обонятельная лампа; (III) система распознавания образов, которая распознавала бы ароматические отпечатки вещества подобно обонятельной системе человека [2].

Понятие «Электронный нос» было введено в 1988 году Гарднером и Бартлеттом, которые определили его как «инструмент, который содержит множество электронных химических датчиков с частичной специфичностью и соответствующей системой распознавания образов, способных распознавать простые или сложные запахи» [3].

В процессе развития систем «электронный нос» возникла перспектива создания портативных приборов, который позволили бы заменить газовую хроматографию в задачах, критичных к скорости распознавания летучих органических соединений. Одновременно с этим стали активно развиваться математические методы анализа данных, получаемых с этих устройств [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 высокоселективные сенсоры – то есть такие сенсоры, которые способы реагировать на многие классы органических веществ. Этот подход

позволяет представлять большее число веществ с помощью меньшего числа сенсоров; в случае же применения низкоселективных сенсоров, их потребовалось бы ровно столько штук, сколько веществ необходимо обнаружить, что лишает устройство портативности и добавляет сложность при анализе. Управление селективностью осуществляется с помощью подбора пленок-сорбентов, наносимых на поверхность сенсоров. Структурная схема системы представлена на рисунке 1.

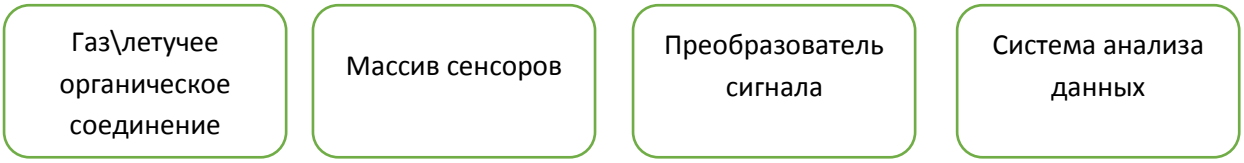

Рисунок 1 – Системы «Электронный нос»

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

пленки увеличивается, что приводит к более медленному перемещению акустических волн. Это изменение может быть обнаружено сенсорной микроэлектроникой, как только акустическая волна преобразуется в электрический сигнал. Существенным ограничением таких сенсоров является необходимость контролировать температуру окружающей среды

Датчик микробаланса кварцевого кристалла (QCM) представляет еще один тип датчиков, в основе которых лежит микровзвешивание. Подобно датчику SAW, преобразователь датчика QCM также чувствителен к массе. Основное различие между SAW и QCM заключается в том, что первый использует датчик поверхностных акустических волн, а второй использует датчик объемных акустических волн. Его чувствительный механизм основан на сдвиге резонансной частоты кварцевого резонатора (QC) за счет адсорбции газовых молекул на поверхности чувствительной пленки [5]. Приборы на основе этих датчиков – одни из самых часто встречающихся на рынке в среднем ценовом сегменте.

Датчики на основе оксидов металлов представляют собой принципиально новое устройство, которые переводит изменение концентрации паров химических веществ в электрические сигналы. На поверхности датчика находится чувствительный металл-оксидный полупроводник с изолирующим слоем под ним, нагреватель и схема для измерения сопротивления. Когда молекулы летучих органических соединений собираются на поверхности оксида металла, начинается процесс их окисление при повышенной температуре, обычно от 250 до 450 ° C. Реакция приводит к переносу электронов от молекул соединения к структуре оксида металла. В результате, регистрируется изменение проводимости. Эти датчики существенно отличаются от кварцевых, поскольку, благодаря такой высокой температуре реакции, индифферентны к изменениям температуры окружающей среды [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 газов и запахов, используя только небольшое количество датчиков, и первым многообещающим результатам, полученным в результате исследований в этой области, к этому предмету возник огромный интерес в научном сообществе и за его пределами.

Зарубежный рынок систем «электронный нос» представлен следующими приборами: FOX с массивом из 6 металл-оксид-электронных сенсоров, и более продвинутая его версия GEMINI (до 18 сенсоров), комбинированная система «электронный нос» с газовым хроматографом HERACLES производства компании Alpha MOS, портативный газовый хроматограф zNose GS/SAW от Electronic Sensor Technology Inc, MOSES II немецкой компании GSG Meß- und Analysengeräte Vertriebsgesellschaft mbH и KAMINA (Германия), Cyranose-320 (Америка). В России подобный прибор разработан в Воронежском государственном университете инженерных технологий группой под руководством доктора химических наук, профессора Татьяны Анатольевны Кучменко и называется «МАГ-8». Далее будут рассмотрены подробнее некоторые наиболее коммерчески успешные продукты. Более полный перечень существующих систем см. в Приложении 1 «Некоторые коммерческие системы «Электронный нос»: модели и технологии» [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 уникальной микросхеме, состоящей из 38 градиентных датчиков окиси олова (SnO2) и

вольфрама (WO3), и помещается в ладони, так как её размеры примерно совпадают с размерами стандартной рации. При более высоких температурах (около 300° C) электропроводность оксида металла зависит от состава окружающего газа и, следовательно, может быть использована для обнаружения газов. Кроме того, каждый датчик нагревается до разных температур, а его толщина мембраны отличается от толщины его соседних датчиков. Как следствие, отдельные датчики имеют различный спектр чувствительности к газам, и все датчики имеют различную реакцию на один газ. Таким образом достигается высокая точность обнаружения газов и их смесей.

Другой коммерческой системой E-nose является система Cyranose 320 E-nose от Smiths Detection [9]. Это портативная система, состоящая из массива из 32 химических датчиков, пробоотборника ЛОС и встроенного системы обработки данных. Она обнаруживает и идентифицирует летучие органические соединения (ЛОС) на основе изменения электрического сопротивления за счет поглощения ЛОС. Сенсорная поверхность представляет собой полимерную сетку с трехмерной непрерывной пористой структурой, заполненной проводящим углеродом. Когда молекулы ЛОС попадают на чувствительную поверхность, реакция между молекулами ЛОС и функциональной группой (группами) в полимерах вызывает расширение объема в полимерной сети. Как следствие, связь между блоками углерода, заполняющими полимерную сетчатую структуру, становится рыхлой, а электропроводность уменьшается. Тип и плотность функциональной группы (групп) в макромолекулах адаптированы для каждого типа датчика, с тем чтобы каждый датчик реагировал на разные ЛОС по-разному.

Система обработки данных Cyranose-320 протестирована в [5], где показано, что точность детектирования резко уменьшается, когда количество компонентов в смеси превышает 3. Это показывает, что данная система не обладает достаточной способностью идентифицировать сложные смеси летучих органических соединений. Однако, идентификация отдельных ЛОС

довольно успешна при условии, что концентрация известных веществ в новых образцах ниже, чем концентрация в объектах тренировочной выборки. Отличные результаты показал данный прибор в задаче классификации бактерий [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), либо 18 (Fox 4000) металлоксидных газовых сенсоров и может использоваться как впрыска газов из внешних баллонов, так и с штатным внутренним насосом и контроллером расхода масс.

Российская разработка - прибор «МАГ-8» - содержит 8 пьезокварцевых сенсоров, разработке прототипа интегрированной системы обработки и анализа данных для этого прибора посвящена данная работа. Существующий подход, прошедший успешную апробацию на ряде типичных для таких систем задач, описан далее. Основные выводы, сделанные в результате ряда исследований, таковы: «электронный нос» «МАГ-8» превосходит физико-химические показатели в задаче оценки органолептических характеристик вина [11]. Сравнительный анализ возможностей интегрального анализатора газа “VOCmeter” (Германия) и дифференциального анализатора «МАГ-8» приведен в [12] на примере задачи количественной и качественной оценки легколетучей фракции ароматических добавок для мясного сырья. Сделан вывод о том, что результат, получаемые с использование отечественной разработки превосходят результаты “VOCmeter” и в большей степени коррелируют с результатами газохроматографии.

Таким образом, первоначальные исследования были направлены на применение систем «электронный нос» в пищевой и косметической промышленности, где они в настоящее время по-прежнему широко используются для оценки качества продуктов питания, контроля вкусовых характеристик и качественного ранжирования сортов вин и пива [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 нормировке данных и приводится несколько формул нормировки, рассматриваются способы отбора признаков. В целом, данная работа скорее обобщает существующие подходы, чем привносит что-то принципиально новое. С момента её написание прошло 11 лет, однако она не перестаёт быть актуальным источником, который позволяет охватить широкий набор методов.

Применение нейронных сетей описал Хоффхайнс в 1989 году в своей работе [4]. Он показал, что, благодаря использованию массивов сенсоров, нейронные сети успешно решают задачу распознавания паров летучих органических соединений, поскольку количество распознаваемых химических веществ в общем случае больше числа сенсоров. Эта фундаментальная работа дала важные результаты как по организации массива сенсоров, так и основу для дальнейших исследований архитектур нейронных сетей и способов представления входных данных. В работе предложен способ представления данных и показано, что худший результат показали сети Хопфилда, вероятно,

и-за малой размерности входных данных, а наилучший – сеть Больцмана, которая не только возвращала лучшую метку, но также показывала следующего подходящего кандидата, что позволяло использовать эту сеть в задаче отображения концентрация газа. Сеть Хэмминга показала наилучший результат в распознавании смесей многих компонентов. Кроме того, показано, что алгоритмы кластеризации способны успешно разделять гексан и этанол, а также высокие и низкие концентрации смесей воды и этанола. Еще один важный результат данной работы состоит в том, что было найдено следующее ограничение: сеть не способна распознавать неизвестные ей смеси веществ, присутствовавших в обучающей выборке.

Большое количество работ [15][16] посвящено подбору массивов сенсоров для решения определенных задач, что говорит о специфичности этих сенсоров и позволяет предположить, что обучение универсального алгоритма классификации невозможно – для решения каждой отдельной задачи он должен обучаться на отдельном наборе.

Такие системы, построенные на основе двух компонентов - «электронного носа» и автоматизированной системы распознавания, нашли применение в медицине, охране окружающей среды и пищевой промышленности. В работе Келлера [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.

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 применения такого устройства.

Недостатком предыдущих работ по анализу данных является то, что они не освещают возможности многослойных нейронных сетей. 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 расстояние Чебышева для выбора сенсоров в задаче классификации сортов сыра, сделан вывод о том, что для устойчивого различения веществ, характеризующих сорта, достаточно выбрать уникальную пару из набора сенсоров.

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., параллельный факторный анализ), сравниваются различные алгоритмы машинного обучения и исследуется зависимость точность классификации от архитектуры нейронной сети, рассматриваются границы применимости методов. Авторы делают вывод о

важности нормировки как таковой и правильности выбора метода, говорят о том, что характеристики переходных процессов несут больше информации, чем стационарные, интегралы обычно дают лучшую производительность, чем максимальные значения, и такой же вывод можно сделать для производных, особенно когда один датчик извлекает несколько производных. Кроме того, DWT обычно показывает лучшие результаты, чем любые другие функции преобразований. Авторы обращают особое внимание на то, что из-за различий в селективности, чувствительности и специфичности датчиков оптимальные характеристики будут строго индивидуальны для прибора и иногда даже для задачи из-за различного протекания процессов сорбции для разных сорбентов, более того, даже 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 образов для повышения селективности массивов сенсоров с использованием интегрированного нейрогенетического алгоритма классификации (NGCA). Предложенная авторами процедура распознавания образов состоит из сбора данных из массива датчиков, предварительной обработки сигналов алгоритмом скользящего среднего (SMMA) и 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 подается на вход ANN, которая обучается с помощью алгоритма обратного распространения ошибки. Наконец, процедура оценки сравнивает изменения в скорости сенсорных данных и новое значение (конечная зондирования), сохраняемое NGCA в базе данных. Эксперименты показывают, что предложенный авторами NGCA лучшую производительность по сравнению с классическими предыдущим генетическими алгоритмами (GA) и искусственными нейронными сетями

(ANN). Так, отдельно ANN работают на собранных авторами данных с точностью 82%, отдельно GA – 91%, только ANN-процедура NCGA – 92%, только AN-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 проблемы, в последние годы стали развиваться подходы к разработке методов генерации массивов данных газовых сенсоров. Совершенно очевидно, что опираться такие методы должны на понимание физико-химических процессов, происходящих на границе сред – газа и твердых сорбентов. Изучением этих процессов занимается коллоидная химия, и, следовательно, исследовательская работа теперь выходит за рамки только обработки данных или выбора массива сорбентов – теперь это построение математической модели на грани химии и информатики. Значительного успеха в этом достигли в университете Каталонии: на основе однокомпонетной модели сорбции Ленгмюра группой исследователей был создан пакет chemosensors для R, который моделирует работу сенсоров, начиная с модели сорбции и заканчивая моделированием шума, обязательно возникающего в реальных электрических системам [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, 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 датасетов

2.1.1 Источники данных

Для начального исследования и предобработки получены 36 матрицы откликов сенсоров «МАГ-8» на набор веществ-маркеров и 4 смеси – эти объекты составляют обучающее множество. Кроме того, предоставлено 75 новых объектов – это матрицы откликов сенсоров на пробы, взятые с детских игрушек. 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:

- 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.;

Метки классов извлекаются автоматически из названий файлов. Правило именования файлов выглядит следующим образом:

«название_вещества [концентрация] мкл на [носитель]» (носитель и концентрация опциональны). Для формирования датасета было решено не делать различий между одним и тем же веществом в разной концентрации или на разных носителях, поэтому алгоритм извлечения меток классов состоит в том, чтобы разрезать название файла по пробелам и сохранять первый (в индексах списков Python - нулевой) элемент.

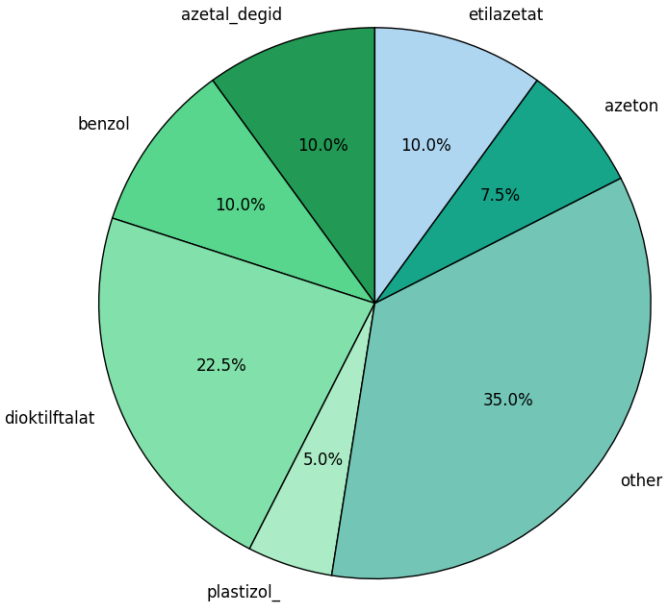

Рисунок 2 – Состав тренировочного множества

В процессе исследования данных выявлены и решены некоторые проблемы. Все проблемы можно разделить на две группы: те, решением которых можно автоматизировать, и те, которые требуют вдумчивого подхода.

Первая среди технически решаемых проблем - опечатки в названия файлов, которые помешают автоматическому извлечению меток классов - всё, что делается вручную подвержено ошибкам, поэтому перед необходимо проверить и при необходимости исправить названия файлов вручную. Вторая проблема с названиями файлов состоит в том, что названия файлов написаны по-русски, а автоматическая бинаризация меток в Sklearn не работает с кириллицей, что лишает нас возможности использовать обучение с учителем – решением этой проблемы является добавление метода, который выполняет примитивную транслитерацию. Третья проблема – это тот факт, что не все файлы содержат матрицу подходящего размера, некоторые – только каждый

двадцатый отсчёт, поэтому невозможно сформировать датасет. Для решения этой проблемы в функцию чтения включена проверка соответствия считанной матрицы заданной размерности.

Более сложная проблема, которая решается неочевидно, состоит в том, что датасет 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 представленные метки класса будут предсказываться чаще. 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 метках, то есть автоматический подбор признаков по методу grid-search также невозможен. Решений этой проблемы было три: поработать с базой данных sniffdb.sdf, пойти по пути расширения набора искусственными данными или попытаться заставить алгоритмы научиться надежно различать классы (возможно, с помощью более сложных методов выделения признаков).

Первый способ решения проблемы рассмотрим ниже, остальные два вынесены в отдельные параграфы. Предполагалось, что из базы данных sniffdb.sdf удастся извлечь дополнительные данные и таким образом расширить множества веществ и смесей. Файл базы данных создан в Microfost SQL Server Compact Edition, что само по себе является серьезным недостатком, так как в команде были ноутбуки не только под ОС Windows, но также под OS X и Linux. Кроме того, это устаревший формат, несовместимый с прочими инструментами Microsoft для работы с базами данных. Для обеспечения кроссплатформенного доступа база с помощью инструмента SDF Viewer была конвертирована в sql-файл, который, в свою очередь был скорректирован для работы с mysql5.5. Для автоматизации работы с БД был написан скрипт на Python parse_sql.py. В процессе анализа извлекаемой информации было обнаружено, что в таблице Data хранятся не изменения частот сенсоров, а

значения частот. Вычисление необходимых матриц ΔFпоказало, что частоты изменяются «ступенькой», что отличается от уже имеющихся данных из XLS, где они изменяются плавно. Возможная причина состоит в том, что система «электронный нос» совершенствовалась, поэтому данные из БД сделаны более старой версией анализатора, а данные в XLS - более новые. Таким образом, было принято решение отказаться от дальнейшей работы с этой базой.

2.1.2 Исследование исходных данных и предобработка

В качестве начального шага было проведено исследование данных методом сингулярного разложения. На рисунке 3 видно, что во всех трёх датасетах основную информацию несёт только 1 компонента. Остальные можно игнорировать, таким образом превратив датасет в множество векторов, а не в множество матриц.

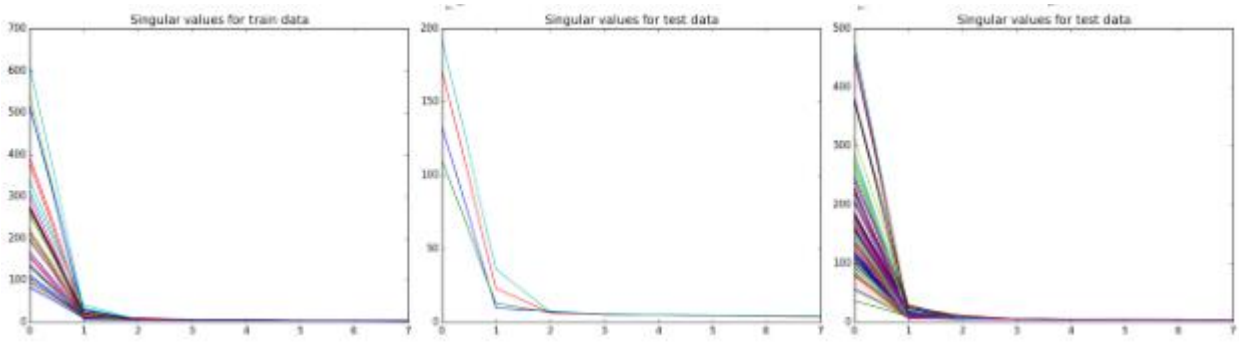

Рисунок 3– Графики сингулярных чисел для веществ, смесей и игрушек

Расширенный текст Дики-Фуллера для выявления стационарности показал, что, вопреки ожиданиям, векторы отклик отдельных сенсоров стационарны. Однако, выявить какие-либо закономерности в этом не удалось. Тем не менее, это ограничивает нас в выборе методов преобразований – преобразование Фурье для нестационарных данных не подходит, остается вейвлет-преобразоване.

Анализ графов кросс-корреляции сенсоров подтверждает гипотезу о том, что сенсоры почти всегда сильно коррелированы (их коэффициент кросс-корреляции – более 0.7), иногда возникает остовное дерево от одной вершины к нескольким. Вероятно, эта информация косвенно подтверждает результаты, описанные в [21] о том, что в случае, если будет избран подход извлечения

численных признаков, необходимо выбирать методы извлечения признаков, опираясь на свойства отдельных сенсоров, а не данных вообще.

Далее был изучен подход, разработанный группой под руководством Т.А.Кучменко – для этого был написан код, который строит радар-диаграммы максимальных изменений частоты для каждого сенсора, остальные же значения заменяет нулями. Визуально, в построенных графиках легко выделяются характерные для каждого вещества элементы. Так, для ацетальдегида явно характерной будет 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 4). Следовательно, преобразованные таким образом матрицы можно использовать в качестве первого метода отбора признаков. Кроме того, в диссертации Дроздовой [25] обращается внимание, что не только dFmax характеризует матрицу откликов, но и dFравн – равновесная частота, то есть момент, когда процесс сорбции вошел в равновесную фазу.

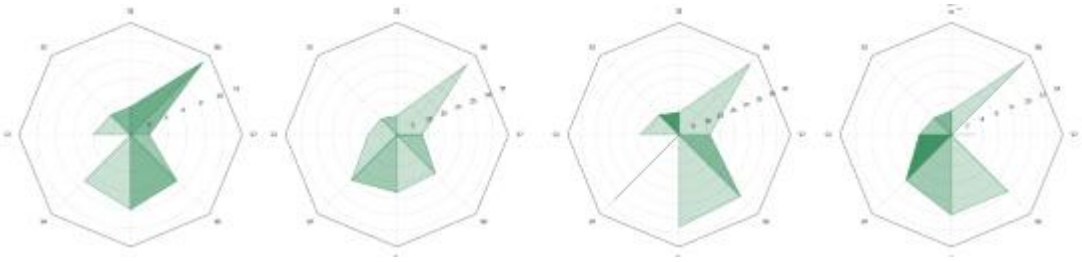

Рисунок 4 – Радар-диаграммы dFmax для ацетальдегида

Проверим предположение с помощью SVM (машина опорных векторов) 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.

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 и не является опасным; бензол, ацетальдегид и этилацетат представлены поровну и занимают второе место – 10% в массиве веществ; ацетон – на третьем месте (7.5%), далее – пластизоль, обнаружение которой на самом деле не означает ничего, так как она является нетоксичным пластификатором, который обычно и применяется в изготовлении детских игрушек. Прочие вещества и смеси представлены 1 объектом – с их верным распознаванием могут возникнуть трудности.

Данные	**Предсказанное**	**Истинное**
Только исходные данные (нормировка + 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 вероятностей

Поскольку оценка точности, рассчитанная для одного объекта, не является точной, рассмотрим влияние компоновки обучающего датасета на такие метрики качества предсказания, как coverage error (CE) и средняя точность ранжирования (label ranking average precision, LRAP).

Coverage error показывает среднее количество меток классов, которое необходимо включить в финальное решение, чтобы были предсказаны все верные метки класса. 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):

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

, 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 основан 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 показателях, назначенных для каждой метки $\hat{f} \in \mathcal{R}^{n_{\text{samples}} \times n_{\text{labels}}}$, метрика LRAP рассчитывается по следующей формуле (2):

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

, 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 l0-норма или мощность множества.

Данные	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Только исходные данные (нормировка + 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 датасет. 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.

Метод	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Только нормировка	15.72	**8.96**	0.26	**0.53**
Только удаление тренда	20.0	10.36	0.05	0.41
Нормировка + 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
Нормировка + удаление тренда + полином	20.0	12.05	0.05	0.38
Нормировка + удаление тренда + масштабирование	19.52	**9.84**	0.07	**0.45**
Нормировка + удаление тренда + полином + масштабирование	20.0	11.85	0.05	0.41

Таблица 3 – Влияние методов предобработки на метрики CE и LRAP

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)
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

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 нормировка. Необходимо учесть, что top-2 предсказанных меток оказались верными для удаления тренда (ДОФ – 33%, ацетон – 30%), для нормировки и удаления тренда (ДОФ – 52%, ацетон – 26%), что ненамного хуже, чем только нормировка. Хорошие результаты дает комплексный подход «нормировка + удаление тренда + масштабирование» (ацетон – 32%, ДОФ – 27%). Кроме того, он ненамного хуже при итоговой классификации, чем нормировка (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 «нормировка + удаление тренда + масштабирование».

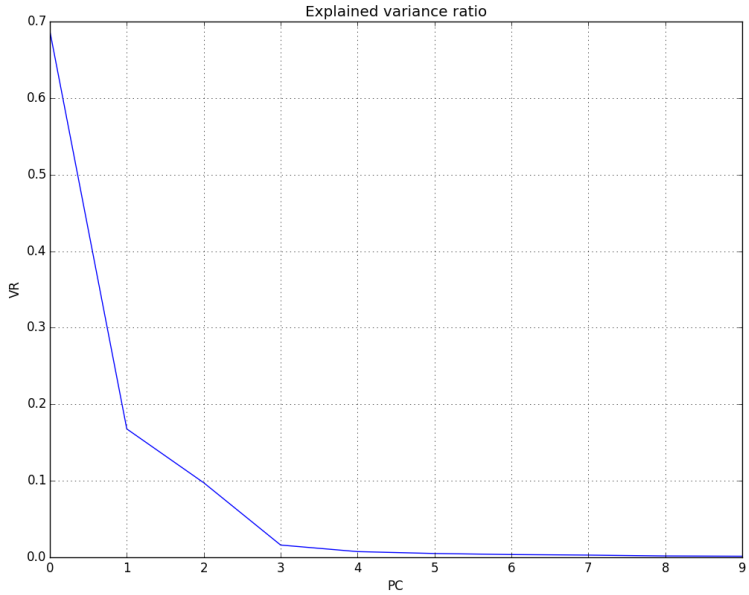

Рисунок 6 – Объяснённая дисперсия остатков для первых 10 главных компонент

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%.

По рисунку 6 видно, что 1 главная компонента объясняет только 14% исходной дисперсии, однако про дальнейшем увеличении числа главных компонент объяснённая дисперсия остатков резко падает. На рисунке показаны только первые 10 компонент, однако необходимости показывать дальнейшее поведение графика нет – с ростом порядкового номера компоненты объясняют менее 1% дисперсии модели. Кроме того, 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 соотношение объяснённой дисперсии начинает стремительно падать (рисунок 6), и уже 3 компонента объясняет лишь 9%. 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 нормировка	0.3062	0.0810	0.0663	0.0554	0.0460
Только удаление тренда	0.0573	0.0438	0.0417	0.0396	0.0381
Нормировка + 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
Нормировка + удаление тренда + полином	0.1657	0.1475	0.1246	0.1163	0.0984
Нормировка + удаление тренда + масштабирование	0.1406	0.1157	0.0994	0.0782	0.0642
Нормировка + 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 нормировка 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 нормировка	0.4071	0.1743	0.1365	0.0824	0.0528
Только удаление тренда	0.1953	0.1323	0.1151	0.1062	0.099
Нормировка + 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
Нормировка + удаление тренда + полином(5)	0.4181	0.2081	0.1193	0.0929	0.051
Нормировка + удаление тренда + масштабирование	0.6846	0.3137	0.0017	0	0
Нормировка + 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: нормировка 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 понижение размерности до 1 главной компоненты LDA.

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 – просто 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 методы (таблица 6).

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 равна 1, то есть метки назначаются наилучшим образом. Однако, для назначения многих меток для игрушек этот подход в корне не верный – копии объектов не несут никакой информации.

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

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: (1) уменьшает смещение, вызванное дисбалансом классов, и (2) адаптивно сдвигает границу принятия решения к трудным примерам.

Объекты каждого класса генерировались в противопоставлении ко всем прочим, чтобы добиться максимального отличия. Таким образом получилось, что экземпляров каждого класса от 35 до 39 штук. Однако, вопреки ожиданиям, возникли серьезные проблемы с обучением на двух наиболее

представленных классах – диоктилфталате и ацетальдегиде. Кроме того, существенного повышения точности распознавания не произошло: диоктилфталат по-прежнему стоит на первом месте среди назначаемых меток для большинства новых объектов.

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% дисперсии. Вероятно

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 модель по кросс-валидации, а в качестве новых данных подавать исходные вещества и смеси. Таблица 8 содержит среднюю точность и стандартное отклонение для логистической регрессии, k ближайших соседей, наивного байесовского классификатора, машины опорных векторов, ExtraTrees и AdaBoost с параметрами по умолчанию. Для выбора наилучшей модели необходимо минимизировать среднее значение по кросс-валидации на сами запусках, а затем, если таковых 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
GaussianNB	0.787	0.038
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

2.2.2 Моделирование объектов всех классов 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 и 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 партии 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. 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 определяются многослойными персептронами, вся система может обучаться с помощью алгоритма обратного распространения ошибки.

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

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:

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

Поочерёдно оптимизируя градиент каждой из сетей на новых партиях 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 также максимизирует взаимную информацию между небольшим 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 волос, 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.

Метод	**G_loss (0)**	**D_loss (0)**	**G_loss (100)**	**D_loss (100)**
Только нормировка	0.693193	0.692456	0.000064	0.000000

Только удаление тренда	0.692932	0.695542	**0.00002**	**0.00003**
Нормировка + удаление тренда	0.693598	0.693190	0.344400	0.000009
Нормировка + 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
Нормировка + удаление тренда + полином	0.693081	0.693099	0.347795	0.002469
Нормировка + удаление тренда + масштабирование	0.692997	0.686141	0.001910	0.000697
Нормировка + удаление тренда + полином(3) + масштабирование	0.693046	0.694560	0.004515	0.000073
Нормировка + удаление тренда + полином(5) + масштабирование	0.693109	0.701059	0.000113	0.000224
Нормировка + 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

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 dFmax, не представляется возможным.

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.

Метод	CE(new)	LRAP (new)	Предсказано
Ацетальдегид			
Ацетон			
Бензин			
Бензол			
Бутилацетат			
Гексан			
Изобутанол			
Изопропанол			
Пластизоль			
Пропанол			
Стирол			
Толуол			
Фенол			
Этилацетат			

Таблица 10 – Проверка сгенерированных данных

2.3 Применение алгоритмов машинного обучения и нейронных сетей

2.3.1 Сравнение основных алгоритмов машинного обучения

Сравнительно небольшое число алгоритмов в фреймворке Sklearn поддерживают многоклассовую классификацию: Naive Bayes, LDA and QDA, Decision Trees, Random Forests, Nearest Neighbors. При этом, более одной метки класса способны назначать только Decision Trees, Random Forests, Nearest Neighbors. [33] Для прочих алгоритмов реализованы подходы «каждый-против-каждого» (one-vs-one) и «один-против-всех» (one-vs-all).

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 обучающей выборки присутствуют в пробах игрушек.

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
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***
GAN				
Support Vector Machine				
Bagging with Random Forest				
Random Forest				

Таблица 11 – Подбор алгоритмов классификации

2.3.2 Применение 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 Keras - фреймворк для работы с нейронными сетями для Python

Алгоритм	CE(train)	CE(new)	LRAP(train)	LRAP(new)
Исходные данные				
Multilayer Perceptron (alpha 97.4)	20.0	10.82	0.05	0.38
Over-sampling (ADASYN)				
Multilayer Perceptron	18.69	13.44	0.11	0.24
GAN				
Multilayer Perceptron				

Таблица 12 – Подбор нейронных сетей

ЗАКЛЮЧЕНИЕ

При написании данной выпускной квалификационной работы было проведено исследование 40 матриц откликов пьезоэлектрических сенсоров системы «Электронный нос», содержащих ароматические отпечатки 36 органических веществ-маркеров и 4 их 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 выстроился конвейер, состоящий из нормировки 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 для балансировки 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 и сделан вывод .

На втором шаге конвейера ставилась задача классификации на 20 классов, в результате ожидалось получить список веществ-маркеров и их смесей, обнаруженных в пробе каждой из игрушек. Наилучшим образом себя показал алгоритм , который дал процент совпадений с эталонными значениями, полученными на кафедре физической и аналитической химии с

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

Для дальнейшего развития работы существуют следующие пути:

Таким образом, цели и задачи, поставленные во введении, были достигнуты, поэтому выполненная курсовая 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. 1994, 18, 211-220.
4.Hoffheins, B. Using Sensor Arrays and Pattern Recognition to Identify Organic Compounds // M.Sc. Thesis. - University of Tennessee. – Knoxville. - TX, USA. - 1989.
(ЕСТЬ ДУБЛЬ В 1.2)
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. J.A. Dickson, et al., “An Integrated Chemical Sensor Arrays Using Carbon Black Polymers and a Standard CMOS Process”, Proc. Solid-State Sensors and Actuators Workshop, Hilton Head Island, SC, June 2000, pp. 162-165.
16. James, D., Scott, S.M., Zulfiqur, A., O'Hare, W.T., 2005. Chemical sensors for electronic nose systems. Microchimica Acta 149, 1-17.

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. Е.В. Дроздова, Определение органических легколетучих токсикантов массивом пьезосенсоров для оценки безопасности полимерных материалов: диссертация кандидата химических наук: 02.00.02 / Дроздова Евгения Викторовна; [Место защиты: Воронеж. гос. ун-т]. - Воронеж, 2016. - 263 с.: ил
26. Scikit-learn: Model evaluation: quantifying the quality of predictions //Scikit-learn [Электронный ресурс]. — Режим доступа: . — (Дата обращения: 02.04.2017)
27. http://contrib.scikit-learn.org/imbalanced-learn/index.html#
28. H. He, Y. Bai, E.A. Garcia, S. Li, 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. 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	Airsense Analytics	i-Pen, PEN2, PEN3	MOS sensors

(только «электронный нос»)	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