

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

На правах рукописи



Назаренко Дмитрий Владимирович

**Химический анализ для идентификации лекарственных растений:
хромато-масс-спектрометрия с интерпретацией данных методами
машинного обучения**

Специальность:

02.00.02 – Аналитическая химия

ДИССЕРТАЦИЯ

на соискание учёной степени
кандидата химических наук

Научный руководитель:

д.х.н., в.н.с. Родин И.А.

Москва–2019

Оглавление

1	Литературный обзор	10
1.1	Профилирование лекарственных растений	10
1.1.1	Хроматографические методы	13
1.1.2	ИК-спектроскопия	20
1.2	Хеометрика и машинное обучение	20
1.2.1	Обучение без учителя	23
1.2.2	Обучение с учителем	33
2	Экспериментальная часть	48
2.1	Оборудование и материалы	48
2.2	Экстракция	49
2.3	Схема эксперимента по ВЭЖХ-МС анализу образцов лекарственных растений	50
2.4	Программные средства, использованные для построения классификационных алгоритмов	51
3	Построение классификационных алгоритмов для выборки из 36 классов .	52
3.1	Задача классификации	56
3.1.1	Предобработка данных	56
3.1.2	Логистическая регрессия	57
3.1.3	Метод опорных векторов	58
3.1.4	Решающие деревья и случайный лес	59
3.1.5	Показатели эффективности	61
3.1.6	Искусственные данные	63
3.2	Результаты и обсуждение	64

4	Построение классификационных алгоритмов на выборке из 76 классов. .	70
4.1	База данных	70
4.2	Кросс-валидация и показатели эффективности алгоритмов	76
4.3	Использованные модели	77
4.3.1	Байесовские сети	77
4.3.2	Автоэнкодер	80
4.3.3	Разложение Таккера с условиями неотрицательности и разреженности	82
4.3.4	Матричное разложение с условиями неотрицательности и разреженности	86
4.4	Результаты и обсуждение	88
	Литература	122

Обозначения

BP-ANN искусственная нейронная сеть с обратным распространением ошибки

FN ложноотрицательный

FP ложноположительный

FP-ANN искусственная нейронная сеть с прямым распространением ошибки

HCA иерархический кластерный анализ

k-NN классификация по k ближайшим соседям

LDA линейный дискриминантный анализ

MC/MC тандемная масс-спектрометрия

PCA метод главных компонент

PLS-DA дискриминантный анализ с помощью регрессии на латентные структуры

PLS-DA проекция на латентные структуры

SIMCA формальное независимое моделирование аналогий классов

SOM самоорганизующиеся карты

SVM метод опорных векторов

TCCCN temperature-constrained cascade correlation network

TCM traditional chinese medicine

TN верно-отрицательный

TP верно-положительный

БАД биологически активная добавка к пище

БИК ближне волновая инфракрасная спектроскопия

БС байесовская сеть

ВЭЖХ высокоэффективная жидкостная хроматография

ГВМ графическая вероятностная модель

ГХ газовая хроматография

ГХ×ГХ комплексная двумерная газовая хроматография

ДМД диодно-матричный детектор

ЖХ×ЖХ комплексная двумерная жидкостная хроматография

ИК инфракрасный

МНК метод независимых компонент

МС масс-спектрометрия

НБК наивный байесовский классификатор

УФ ультрафиолетовый

ЯМР ядерный магнитный резонанс

Введение

Актуальность темы. Лекарственные растения представляют собой практически неисчерпаемый источник для поиска новых физиологически активных веществ, пригодных для медицинского использования [1, 2]. Доля новых лекарственных препаратов, в той или иной мере основанных на соединениях растительного происхождения составляет около 30%. Тем не менее, большинство современных фармакологических препаратов содержат в своём составе только 1-2 индивидуальных действующих вещества, по которым были проведены регламентированные клинические испытания. Контроль качества подобных препаратов почти всегда представляет собой определение содержания действующего вещества в лекарственной форме методом внешнего стандарта, а также контроль малого числа известных побочных продуктов или примесей. В противовес этому, использование в качестве лекарственных препаратов частей растений или растительных экстрактов, содержащих от десятков до сотен соединений, многие из которых обладают той или иной физиологической активностью, представляет собой весьма затруднительную задачу [3, 4, 5, 6]. Во-первых, одновременный аналитический контроль большого числа соединений требует использования большого числа дорогостоящих стандартных соединений, во-вторых нет чётких критериев того, какие уровни содержаний тех или иных соединений в растениях устанавливать как допустимые [7]. Дополнительную сложность вносит трудность проведения и однозначной интерпретации результатов клинических исследований с участием сложных по составу растительных препаратов [8].

Вышеперечисленные сложности в том числе ограничивают и нормативное регулирование лекарственных препаратов на растительной основе. В большинстве стран мира, в том числе США, Европейском Союзе и СНГ, вплоть до сегодняшнего дня отсутствует строгая регламентация подобных препаратов [9, 10]. При этом объём данного рынка составляет десятки миллиардов долларов только для развитых стран [11, 12]. А для большей части населения планеты лекарственные растения вообще являются одним из немногих или даже единственным доступным источником медицинского вмешательства [13, 14]. Весьма выгодно в этом направлении выделяются работы китайских учёных, так как фармакология Китая стремится интегрировать традиционные препараты на основе лекарственных рас-

тений в русло современной медицины [15]. Поэтому, несмотря на препятствия и трудности процесса интеграции растительных препаратов в систему научной фармакологии, на сегодняшний день ведется много исследований, посвящённых как клиническим испытаниям, так и разработке методик и подходов по контролю качества лекарственного растительного сырья и препаратов на его основе [16, 17, 18].

Цель работы состояла в разработке подхода к видовой идентификации сырья лекарственных растений с использованием высокоэффективной жидкостной хроматографии масс-спектрометрии низкого разрешения и методов машинного обучения. Достижение поставленной задачи предусматривало решение следующих задач:

1. Создание базы данных ВЭЖХ-МС анализа препаратов лекарственных растений.
2. Выбор способа извлечения значимой информации из данных ВЭЖХ-МС анализа в как можно более компактной форме в смысле размера признакового пространства.
3. Разработка способа идентификации видовой принадлежности неизвестных растительных образцов.
4. Апробация подхода для образцов различного географического происхождения, а также образцов, экстрагированных способами, отличными от использованного при разработке алгоритма и проанализированных на различном ВЭЖХ-МС оборудовании.

Научная новизна

1. Изучены особенности извлечения характеристических химических профилей при ВЭЖХ-МС анализе лекарственных растений.
2. Изучены варианты предобработки и преобразования данных химических профилей и их влияние на эффективность работы идентификационных алгоритмов.

3. Разработаны подходы к идентификации видовой принадлежности препаратов лекарственных растений на основе ВЭЖХ-МС анализа и машинного обучения.
4. Изучено влияние возмущений в химических данных, вызванных различными изменениями в способе анализа экстрактов, на эффективность работы идентификационных алгоритмов.
5. Изучен вклад соединений с различными молекулярными массами в эффективность идентификационных алгоритмов.

Практическая значимость

1. Создан подход к видовой идентификации образцов 74 видов лекарственных растений на основе высокоэффективной жидкостной хроматографии и масс-спектрометрии низкого разрешения, где в качестве образца используется порошок или водно-спиртовой экстракт из растительного материала.
2. Собрана и впервые опубликована в открытом доступе база данных ВЭЖХ-МС анализа экстрактов образцов 74 видов лекарственных растений.
3. Опубликован в открытом доступе программный код всех разработанных алгоритмов обработки первичных данных и построения моделей идентификации.

На защиту выносятся следующие положения:

1. Результаты исследования подходов к предобработке данных химических профилей лекарственных растений, полученных на основе ВЭЖХ-МС анализа экстрактов.
2. Подход к идентификации видовой принадлежности неизвестных растительных образцов в порошковой форме.
3. Результаты апробации предложенного подхода на образцах 74 видов лекарственных растений.

Апробация работы. Результаты работы докладывались на Пятом Всероссийском симпозиуме с международным участием «Кинетика и динамика обменных процессов». Роль Separation Science в развитии прорывных направлений современной науки (2016, г.Сочи), VII Всероссийской конференции с международным участием "Масс-спектрометрия и ее прикладные проблемы"(2017, г. Московский, Россия), V Всероссийском симпозиуме "Разделение и концентрирование в аналитической химии и радиохимии"(2018, г.Краснодар).

Публикации. По материалам диссертации опубликовано 3 статьи в научных изданиях, индексируемых в базах данных Web of Science, Scopus, RSCI, изданиях из перечня, рекомендованных Минобрнауки РФ, и 3 тезисов докладов.

Личный вклад автора состоял в общей постановке задач, систематизации литературных данных, подготовке и проведении всех экспериментальных этапов исследования, обработке, интерпретации и оформлении полученных экспериментальных данных, подготовке материалов к публикации и представлении полученных результатов на конференциях. Все исследования, описанные в диссертации, выполнены лично автором или в сотрудничестве с коллегами.

Структура и объем работы. Диссертация состоит из введения, обзора литературы, экспериментальной части, 2 глав обсуждения результатов, заключения и списка цитируемой литературы, изложена на 140 страницах машинописного текста и включает 48 рисунков, 8 таблиц и список цитируемой литературы из 208 наименований.

ГЛАВА 1

Литературный обзор

1.1. Профилирование лекарственных растений

Аналитическая методология по анализу лекарственных растений переживает интенсивный подъём в последние десятилетия [16, 19, 20]. Задача анализа сложных по составу объектов, в которых сложно установить требующие контроля индивидуальные соединения, привела к широкому применению многомерных статистических методов анализа данных [21, 22]. Данная тенденция наиболее выражено относится к фармакологии Китая, где целью ставится приведение традиционных медицинских практик к стандартам доказательной медицины [23, 24, 25].

Эта цель имеет на своём пути как минимум две проблемы в случае растительных препаратов - стандартизацию самих препаратов и интерпретацию проводимых на их основе клинических исследований. Проблема стандартизации достаточно очевидна - разработка, производство и рутинный контроль качества любого препарата опираются на стандартизованные критерии его качества [7, 26, 27].

Профилирование (profiling) или метод отпечатков пальцев (fingerprinting) стали действенной альтернативой традиционной методологии [28, 29, 30]. Основная гипотеза в профилировании заключается в том, что аналитические данные (например, спектроскопические данные анализа экстракта растения) уже содержат информацию, необходимую для ответа на имеющийся вопрос контроля качества. Необходимо только некоторым образом отделить полезную информацию от шума. Для этого требуется некоторая выборка данных, содержащая данные анализа образцов, отвечающих разным состояниям, таким как настоящий/поддельный [31], чистый/разбавленный [29, 32], разным видам [33], географическим источникам

происхождения [34, 35] и т.п. [36, 37, 38]. За счёт применения различных методов возможно извлечение из данных тех переменных, которые позволят достоверно различить вышеуказанные состояния [39]. Общая схема подобных исследований может быть выражена следующим образом:

1. Комплексный химический анализ (ЯМР, ВЭЖХ-МС, ВЭЖХ-МС/МС и т.п.) растительного материала или готовых препаратов лекарственных растений, в том числе смесей (например препараты ТСМ, имеющие в составе одновременно до 10 и более видов растений) [40].
2. Использование полученных данных для поисков связи качества препарата с каким-либо конкретным маркерным соединением/группой соединений. В результате получают характеристичный профиль, в котором заданы окна приемлемых концентраций маркеров качества [41].

Одна из ключевых проблем разработки эффективных классификационных это выбор признакового пространства, т.е. списка переменных подаваемых на вход обучаемой математической модели. Так, каждый файл, получаемый после ВЭЖХ-МС анализа содержит в исходной форме миллионы переменных, а после обнаружения и интегрирования пиков число переменных сокращается до сотен (тысяч, в случае сложных проб биологического происхождения). При этом в большинстве случаев аналитику доступны очень ограниченные размеры выборок данных, что приводит к необходимости искать минимально возможное число переменных, которое тем не менее содержит всю необходимую в конкретной задаче аналитическую информацию. Популярными способами, помимо прочих, в данной области являются автоэнкодер и тензорные разложения [42, 43]. В отличие от широко используемого РСА, который направлен на поиск направлений максимальной дисперсии, автоэнкодер должен уметь восстановить входные данные в полном виде. Это происходит за счёт выявления внутренней структуры данных. Выход последнего кодирующего слоя автоэнкодера может далее использоваться как признаковое пространство для классификационной модели. Аналогично, тензорные разложения применяют для проекции данных на пространства низкого разрешения и разделения переменных.

Другая важная проблема в анализе лекарственных растений это сложность выработки критериев для ручной маркировки меток классов. Универсальный алго-

ритм по видовой идентификации лекарственных растений может служить хорошей отправной точкой в исследовании данного вопроса. Несмотря на то, что попытки создания чего-то подобного уже наблюдались [4, 6, 44], законченного алгоритма до сих пор не разработано.

ВЭЖХ в комбинации с масс-спектрометрическим детектором стала на сегодняшний день общепринятым аналитическим методом для подобных приложений [45, 46, 47, 48, 49, 50], с меньшим числом работ затрагивающих ГХ-МС [51, 52], ЯМР-спектроскопию [53, 54], ИК-спектроскопию [55, 56, 57] и другие методы (такие как "электронный нос" [52] и электрофорез [50, 56, 58]).

Распространенный подход в профилировании - проанализировать имеющиеся данные и выработать небольшой (от нескольких до десятков) список маркерных соединений, которые можно использовать для контроля качества препаратов какого-либо вида растения. Один из главных недостатков таких подходов заключается в высоких ценах и малой доступности стандартных соединений и их изотопно меченных аналогов.

Профилирование чаще используется в комбинации со спектральными методами анализа, такими как ИК, УФ, масс-спектрометрия и спектроскопия ЯМР. Основная идея такого подхода – попытаться зафиксировать некую внутреннюю структуру химического состава пробы, которая будет служить «отпечатком пальцев», позволяющим надёжно опознавать подделки или некачественные препараты. Профилирование полезно для применения в качестве идентификационного инструмента, улавливающего качественное отличие химического состава пробы от некоего выбранного эталона, однако в силу своей природы не позволяет одновременно решать задачи по количественному анализу. Данная работа является расширением подхода профилирования на случай анализа большого числа видов лекарственных растений. Она посвящена совмещению подходов профилирования и методов машинного обучения для создания программного инструмента по видовой идентификации препарата лекарственного растения на основе результатов химического анализа и в частности ВЭЖХ-МС анализа.

Исторически, при открытии новых видов растений, их постулирование давалось набором морфологических признаков растения и его частей на разных этапах биологического цикла. С развитием возможностей биологии и химии, видовая при-

надлежность также стала устанавливаться и подтверждаться за счёт секвенирования последовательностей нуклеиновых кислот [59, 60]. В состав растений входит широкий перечень классов соединений: белки, углеводы, жиры, соли металлов, нуклеиновые кислоты, низкомолекулярные органические соединения. Наиболее достоверным способом видовой идентификации растения является морфологическая идентификация, то есть идентификация по внешним признакам [61]. Эффективным инструментом по видовой идентификации биологического материала живых организмов также служит секвенирование белковых последовательностей. Белковый состав растений, однако, далеко не всегда позволяет провести однозначную и достоверную видовую идентификацию [62]. В случае работы с сухим измельченным препаратом растения или экстрактом видовая идентификация неизвестного образца является гораздо более сложной задачей.

1.1.1. Хроматографические методы

Одномерная хроматография

Благодаря простоте использования и способности осуществлять разделение сложных по составу проб на индивидуальные компоненты, что делает анализ таких проб возможным, высокоэффективная жидкостная хроматография является одним самых широко используемых методов разделения в химическом анализе. В том числе ВЭЖХ в комбинации с различными детекторами является обязательным элементом в исследованиях по профилированию лекарственных растений. В [63] авторами была сконструирована комплексная аналитическая система для анализа экидистероидов в растительных экстрактах. Хроматографическая система, использующая в качестве элюента перегретый D_2O была соединена с УФ-, ИК-МС- и ЯМР детекторами в режиме онлайн. Данная система использовалась для идентификации и определения производных 20-гидроксиэкидизона и экидистерона

в экстрактах *Silene otites*, *Silene nutans* and *Silene frivaldiskyana*. Разделение проводили на колонках C8 XTerra с 5 мкм сорбентом. Несмотря на высокие пределы обнаружения для ИК- и ЯМР- детекторов (ок. 100 мкг вещества на колонке), подобная система позволила проводить однозначную идентификацию соединений как в смесях стандартов, так и в сложных экстрактах. Авторы также предположили, что при дальнейшей оптимизации пределы обнаружения могут быть значительно снижены.

В похожей работе использовалась система, состоящая из УФ, МС, МС/МС, ^1H -ЯМР и ^1H - ^1H -ЯМР- каналов детектирования подключенных в онлайн варианте к хроматографу [64]. Данная установка использовалась для анализа экстрактов двух видов семейства *Gentianaceae* - *Swertia calycina* и *Gentiana ottonis*. Совмещение 4 каналов данных позволило напрямую идентифицировать основные компоненты в неочищенных экстрактах. Несмотря на большой потенциал с точки зрения структурной информации, получаемой после одного акта анализа экстракта, существует очень ограниченное количество публикаций с использованием ЯМР-спектроскопии в онлайн варианте подключенной к хроматографической системе. Хотя сам интерфейс подключения к ЯМР-спектрометру значительно проще реализуем, чем соединение с МС-детектором, низкая чувствительность ограничивает практическое применение. К тому же, помимо высокой цены самого ЯМР-спектрометра, необходимо либо использование дейтерированных/апротонных растворителей, либо очень высокие концентрации определяемых соединений в пробах.

Детектирование по поглощению (в большинстве случаев в области 200-800 нм) в приложении к хроматографическим системам позволяет проводить быстрый и недорогой анализ широкого круга органических соединений, обладающих поглощением в области 200-800 нм. Детекторы такого типа универсальны, однако в случае сложных по составу проб их селективность может быть недостаточна. Для интерпретации хроматограмм, полученных с использованием детекторов поглощения, необходимо наличие стандартного соединения для каждого определяемого вещества. Тем не менее, благодаря низкой стоимости и высокой чувствительности, ВЭЖХ-УФ и ВЭЖХ-ДМД (диодно-матричный детектор, совмещающий детектирование в УФ и видимой области) системы повсеместно применяются для рутинных задач контроля качества продукции и сырья. В случае лекарственных Для

многих полифенольных соединений их гликозидных производных, содержащихся в лекарственных и пищевых культурах растений, известно их положительное влияние на здоровье человека [65]. За последние десятилетия было опубликовано множество статей, посвящённых поиску источников различных физиологически активных соединений в растениях. Авторы [66] описывают результаты определения индивидуальных фенольных соединений в более чем 30 видах лекарственных растений родом из Греции. После экстракции 62.5% метанолом экстракты разделяли на 250×4.6 мм колонке C18 колонке с диаметром сорбента 5 мкм. Идентификацию соединений проводили путём сравнения спектров и времён удерживания со стандартными соединениями. За один акт анализа проводили одновременное определение 18 соединений, в том числе кверцетина, апигенина и лютеолина.

В работе [67] ВЭЖХ-ДАД использовали для идентификации и определения содержания изофлавонов в растительных экстрактах. На первом этапе водно-метанольные экстракты сои и трёх лекарственных растений анализировали на ВЭЖХ-МС системе с трёхкврупольным масс-спектрометром. После проведения достоверной идентификации всех соединений в пробах, для анализа оптимизировали систему состоящую из УВЭЖХ колонки Zorbax C18 с диаметром сорбента 1.8 мкм и ДАД-детектором. Общее время хроматографического разделения составило порядка 1 мин, при этом количество определяемых соединений составило 10 с пределами обнаружения в области 0.5 – 5 нг/мл.

Аналогичное исследование было посвящено идентификации и определению 5 кумаринов в экстрактах видов *Ammi majusand* и *Ruta graveolens* с использованием ОФ-ВЭЖХ с диодно-матричным детектором [68]. Идентификацию соединений проводили по сравнению времён удерживания и УФ-спектров в области 250-400 нм с таковыми для стандартных соединений. Время анализа одной пробы составляло 50 мин. Примечательно, что для получения как можно лучшей картины хроматографического разделения авторы пробовали различные программы градиентного элюирования с использованием тетрагидрофурана, метанола и ацетонитрила. В итоге разделения умбеллиферона, скополетина и мешающих компонентов удалось достичь только в варианте градиентного элюирования смесью метанол - тетрагидрофуран.

В некоторых случаях ДАД детектирование может использоваться как вспомогательный метод идентификации. Например, при исследовании состава экстрактов

шафрана *Crocus sativus* L. методом ВЭЖХ-МС, цис и транс гликозидные производные кроцина (шафранал и др.) не получилось надёжно разделить как по временам удерживания, так и по масс-спектрам родительских и дочерних ионов [69]. Однако с использованием диодно-матричного детектора было обнаружено две полосы поглощения, на 256 нм и 400-500 нм, позволяющие однозначно идентифицировать все транс- изомеры определяемых соединений, тогда как полосы поглощения в области 200-300 нм соответствовали цис- изомерам. Более того, интерпретация паттернов спектров соединений в УФ и видимой областях также использовалась для установления сайтов связывания и структур гликозидных остатков, что в комбинации с масс-спектрометрией позволило провести достоверную идентификацию исследуемых веществ.

Альтернативный вариант использования спектроскопических детекторов при анализе растительных экстрактов относится к области профилирования или отпечатков пальцев (profiling или fingerprinting). Авторами были собраны ВЭЖХ-ДАД хроматограммы для 60 видов лекарственных растений, используемых в африканской народной медицине [70]. Для каждого растения получали экстракты в 100% дихлорметане и 100% метаноле, которые затем разделяли на C18 колонке 250×4.6 мм с диаметром сорбента 5 мкм. Регистрацию профиля экстрактов проводили на длине волны 254 нм для метанольного экстракта и 270 нм для дихлорметана. Сравнение профилей для образцов, полученных из различных регионов произрастания, показало, что в большей мере варьировались количественные соотношения между соединениями, тогда как качественный состав экстрактов был более устойчив. В работе было высказано предположение, что подобные профили могут быть использованы для видовой идентификации неизвестных образцов, морфологическая идентификация которых не представлялась возможной. Идеи, высказанные в данном исследовании, среди всей проанализированной литературы наиболее близки концептуально данной диссертационной работе. Однако авторы располагали очень ограниченным набором образцов по рассматриваемому в статье виду *Sharon baccifera* и не представили данных по другим 59 видам. Также не было предложено и реализовано алгоритма автоматической идентификации.

Двумерная хроматография

Любые используемые на практике в газовой и жидкостной хроматографии разделительные колонки имеют ограниченную ёмкость по числу пиков, которые можно проанализировать на отдельно взятой системе за один акт анализа. Более того, в случае анализа сложных образцов экологического и биологического происхождения одного механизма разделения оказывается недостаточно для полного разделения компонентов смесей. Проблема неполного хроматографического разделения может быть во многих случаях решена за счет использования более селективного детектора вплоть до обладающего высокой универсальностью и селективностью масс-спектрометрического. Однако даже такой подход не позволяет устранить подавление ионизации минорных компонентов в случае одновременного элюирования с другими компонентами пробы.

Решением данной проблемы может служить использование различных вариантов двумерной хроматографии, ключевой особенностью которой является использование двух типов сорбентов с разными механизмами удерживания. Фракции, полученные на первой колонке в онлайн или офлайн варианте, затем разделяют на второй колонке, которая непосредственно подключена к детектору (Рисунок 1). В случае полностью ортогональных механизмов удерживания такой подход теоретически может позволить достичь ёмкости по пикам равной произведению ёмкостей колонок первого и второго измерения. Одним из наиболее продвинутых вариантов двумерной хроматографии можно считать предложенный в 1990е годы метод комплексной двумерной хроматографии (comprehensive two-dimensional chromatography). Данный вариант включает использование одной «основной» колонки, с длинной программой разделения и большой шириной пиков, и второй «быстрой» колонки, в которой разделение проводится в скоростном режиме (Рисунок ??). Преимуществом подобного варианта является проведение «истинного» двумерного разделения в варианте онлайн, что должно позволяет сократить суммарное время анализа и устранить погрешности, вызванные ручными операциями при работе с фракциями.

За последние годы было выполнено множество исследовательских работ по теме комплексной хроматографии, реализованных на базе как ГХ, так и ВЭЖХ систем. Так в работе [71] авторы использовали ГХ×ГХ-МС систему для изучения

летучих компонентов артишока *Cynara scolymus L.*. На основании интерпретации спектров электронной ионизации и сравнения с базой данных NIST (Национальный институт стандартов и технологий, США) было предварительно идентифицировано 130 соединений, 109 из которых были впервые обнаружены в данном растении. В дополнение, авторами было проведено сравнение данных одномерного ГХ-МС и двумерного ГХ×ГХ-МС вариантов как инструмента идентификации. Было выяснено, что в одномерном случае множество низкоинтенсивных компонентов с перекрывающимися временами удерживания не позволяли получить качественных масс-спектров, необходимых для уверенной идентификации по базе данных.

В силу ограниченного применения ГХ систем для анализа нелетучих и неустойчивых к нагреванию соединений, были также разработаны подходы по применению двумерных ЖХ×ЖХ систем для анализа сложных проб. Одно из таких исследований было посвящено использованию комбинации октадецил-модифицированной обращённо-фазовой колонки (C18) и колонки с пентафторфенил-модифицированным сорбентом (PFP) в паре с УФ и масс-спектрометрическим детекторами [72]. Авторы работы оптимизировали систему под задачу анализа сточных вод на широкий круг низкомолекулярных загрязнителей, таких как полиароматические углеводороды, пестициды и лекарственные препараты. Наличие двух характеристик удерживания по хроматографическим измерениям и точной молекулярной массой позволило провести предварительную идентификацию множества соединений (тербутрина, изопротурона, диазинона и т.д.) на основе данных анализа родственных стандартных соединений. Таким образом, основными преимуществами комплексной двумерной хроматографии в большинстве приложений является как повышения соотношения сигнал-шум определяемых соединений за счёт более полного разделения, так и появление дополнительной качественной характеристики соединения – параметра удерживания по второму хроматографическому измерению.

Ожидалось, что широкое внедрение подобных систем позволит проводить анализ сложных объектов биологического происхождения за меньшее время и с использованием менее дорогостоящих детекторов. Однако подобные системы так и не нашли широкого применения за пределами исследовательских лабораторий в силу сложности настройки и обслуживания, высокого расхода элюента,

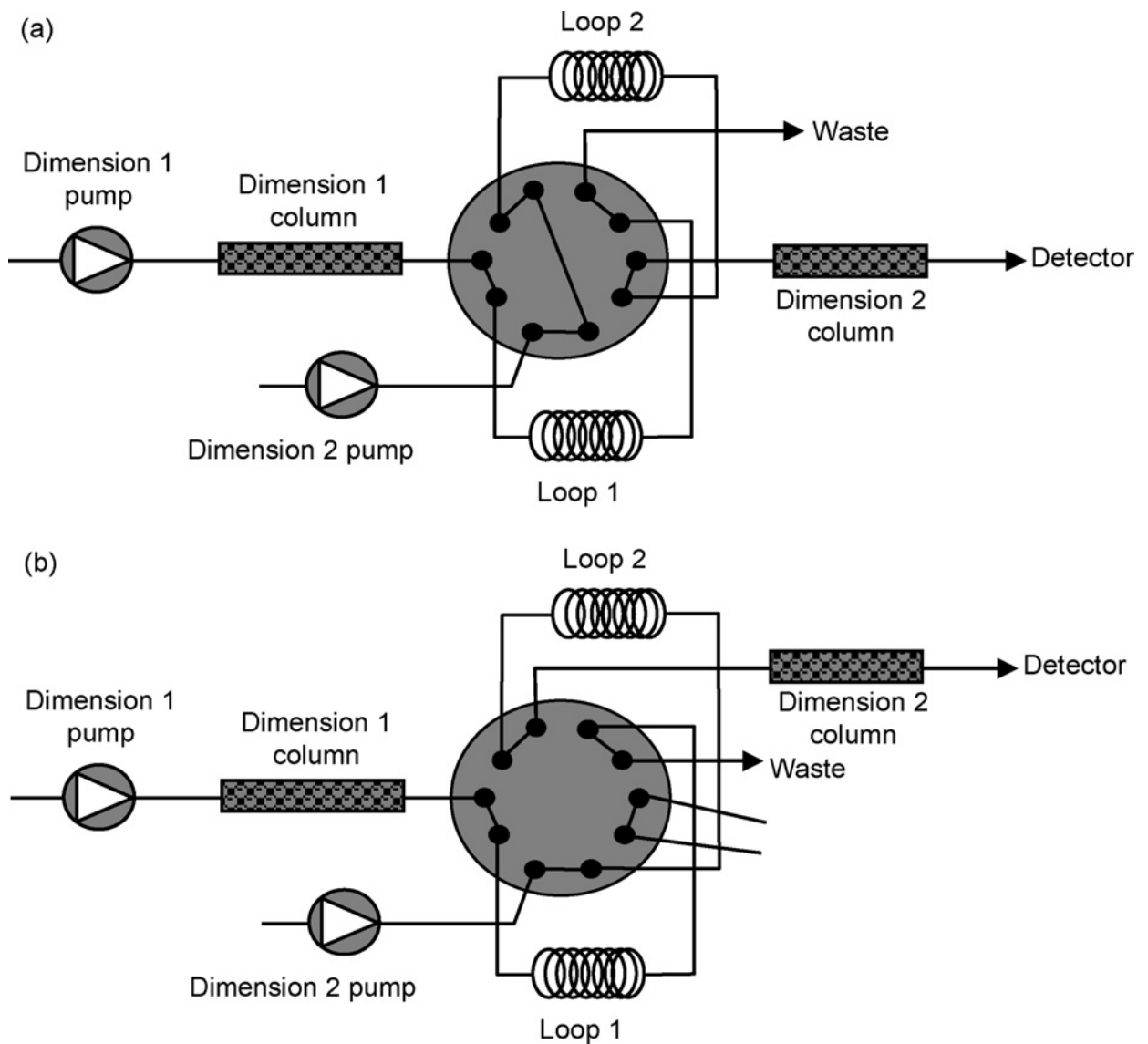


Рис. 1: Распространённая конфигурация комплексной хроматографической системы с использованием 2-позиционных 10-портовых клапанов; симметричная (a) и асимметричная (b) вариации. Взято из [73].

а также гораздо более высоких требований к квалификации персонала. Последующее появление на рынке и широкое распространение масс-спектрометров высокого разрешения, таких как времяпролётные масс-спектрометры и позже масс-спектрометров с орбитальной ионной ловушкой позволило в значительной мере сгладить проблему высокой плотности пиков в пробах сложного состав.

1.1.2. ИК-спектроскопия

Благодаря способности получать интегральную характеристику поглощения соединений пробы и простоте техники исполнения, современная ИК-спектроскопия хорошо подходит для экспрессного химического анализа широкого круга объектов. В комбинации с методами хеометрики, данные ИК-спектроскопии позволяют выявлять различия между видами или местами географического произрастания лекарственных растений с высокой эффективностью. Преимущества ИК-спектроскопии (в том числе низкая стоимость) привели к тому, что её повсеместно применяют в производственном контроле, для анализа почв, еды, напитков и т.д.

Тем не менее, качественный или количественный ИК-анализ почти всегда требует использования математических и статистических методов для экстракции «значащих» данных, а именно спектральных переменных связанных со свойствами аналитов. Также необходимо проводить отбрасывание «незначащих» данных, т.е. мешающих установлению связей спектр-аналит. Подобные методы обработки спектроскопических данных принято относить к области хеометрики [74]. Поэтому более подробно приложение колебательной спектроскопии к анализу лекарственных растений будет рассмотрено в следующей части литературного обзора, касающейся хеометрики и машинного обучения.

1.2. Хеометрика и машинное обучение

Под хеометрикой принято понимать группы математических методов и подходов к обработке результатов химического анализа. Исторически хеометрика

возникла как приложение статистических методов к оценке погрешностей при проведении химического анализа. Оценка достоверности и разброса получаемых в каком-либо методе анализа результатов необходима для их корректного использования при принятии решений, выборе подходящего аналитического метода для конкретной задачи, оптимизации затрат. Статистическая обработка полученных измерений также необходима для возможности сравнения серий анализов, проводимых различными аналитиками, на различном оборудовании, с использованием разных партий реагентов и т.п. Единичный результат измерений при использовании большинства аналитических методов принято интерпретировать как реализацию случайной величины, имеющей распределение Гаусса:

Хемометрика включает в себя набор различных разделов, в том числе планирование и оптимизация эксперимента, стратегии по извлечению информации (моделирование, классификация и проверка статистических гипотез) а также подходы по получению знаний о химических системах [75].

Для изготовления некоторых препаратов на основе лекарственных растений используют разные виды, их морфологические характеристики могут быть весьма похожи, особенно в случае родственных видов. Поэтому проблема быстрой и надежной видовой идентификации одну из решающих ролей в аналитическом контроле данной группы препаратов. Классическая видовая идентификация опирается в основном на визуальные тесты (морфологическая идентификация специалистом ботаников) или тонкослойную хроматографию. Данные тесты могут быть как весьма субъективными, так и требовать наличия соответствующих навыков и опыта, что делает их неэффективными для поточного анализа больших партий растительного материала.

На сегодняшний день существует ряд приложений машинного обучения к задаче морфологического распознавания растений, будь то фотографические изображения листьев [76, 77], коры [78] целого растения [79] или использование спутниковых снимков для распознавания пород в лесных массивах. Во всех вышеперечисленных случаях в качестве входных данных использовали графическое изображение некой части растения, т.е. подобные подходы являются имитацией морфологического распознавания растения квалифицированным специалистом-ботаником. Ботаническая классификация растений как таковая изначально была основана на морфологии растений, то идентификация на основе изображений

является органическим вариантом реализации задачи распознавания. Химический же состав для растений изучен далеко не столь исчерпывающе. Несмотря на большое число исследований на тему изучения состава растений, огромное количество известных человеку видов высших растений ограничивает степень изученности данного направления. Задача видовой идентификации приобретает совсем иной характер, когда входной образец растения был высушен и измельчен до состояния порошка. Ещё более трудной задача становится в случае необходимости идентификации водного или водно-спиртового экстракта растения. Однако в производстве лекарственных средств и биологически активных пищевых добавок лекарственные растения используются именно в этой форме. В классических аналитических подходах, для достоверной идентификации и определения соединений необходимо использование стандартных образцов, представляющих собой высоко чистые индивидуальные соединения. В случае компонентов лекарственных растений, такие соединения чаще всего получают за счёт препаративного хроматографического выделения. Из-за сложности очистки, такие стандартные соединения отличаются высокой ценой, поэтому аналитические лаборатории обычно располагают только ограниченным набором стандартов, используемых в рутинном аналитическом контроле. Таким образом, в общем случае видовой идентификации неизвестного растительного образца представляет собой трудную аналитическую задачу, для которой на сегодняшний день не разработано адекватных методических подходов. Ещё более трудной задача видовой идентификации становится в случае смесей препаратов лекарственных растений. В машинном обучении выделяют три основные группы методов: регрессия, классификация и кластеризация. Под регрессией понимают математическое выражение, отражающее зависимость зависимой переменной y от независимых переменных x при условии, что это выражение будет иметь статистическую значимость. Примером может служить задача предсказания биржевого курса компании. В случае классификации стоит задача определить принадлежность неизвестного объекта к какому либо из заранее заданных классов, например распознавание объектов на фотографии. Регрессия и классификация относятся к категории «обучения с учителем»: алгоритм сначала «обучают» с использованием вручную размеченных данных, а затем используют на неизвестных объектах. Кластеризация относится к методам «обучения без учителя», т.е. поиску закономерностей в неразмеченных данных.

Одним из наиболее популярных направлений данной области является сов-

местное использование методов аналитической химии и машинного обучения для создания эффективных и недорогих методов по контролю качества препаратов сложного состава [80, 81, 82, 83, 84].

1.2.1. Обучение без учителя

Машинное обучение представляет собой раздел науки на стыке математики и программирования, который охватывает создание и применение различных алгоритмов, способных решать различные задачи без чётких заранее прописанных инструкций [85]. Обучение без учителя – раздел машинного обучения, посвященный анализу данных, у которых отсутствует какая-либо входная информация об их внутренней структуре. На вход алгоритму подается выборка данных, в которой он пытается выявить какие-либо закономерности, например, разделить выборку на группы каким-либо образом схожих между собой объектов, провести ранжирование и т.п. Наиболее известными и широко применяемыми алгоритмами обучения без учителя в хемометрике являются иерархический кластерный анализ (HCA, Hierarchical Clustering Analysis) [86] и метод главных компонент (PCA, Principal Component Analysis) [87]. Несмотря на то, что в большинстве реальных приложений использование методов обучения без учителя ограничено предварительными этапами анализа структуры выборки, имеет смысл кратко рассмотреть их применение в исследованиях.

PCA и HCA

Классический метод обучения без учителя — иерархический кластерный анализ. В HCA, имея выборку из N объектов производят следующие шаги:

1. Каждому объекту назначается кластер, таким образом при N объектах в начале имеется N кластеров. Рассчитывают расстояния между кластерами

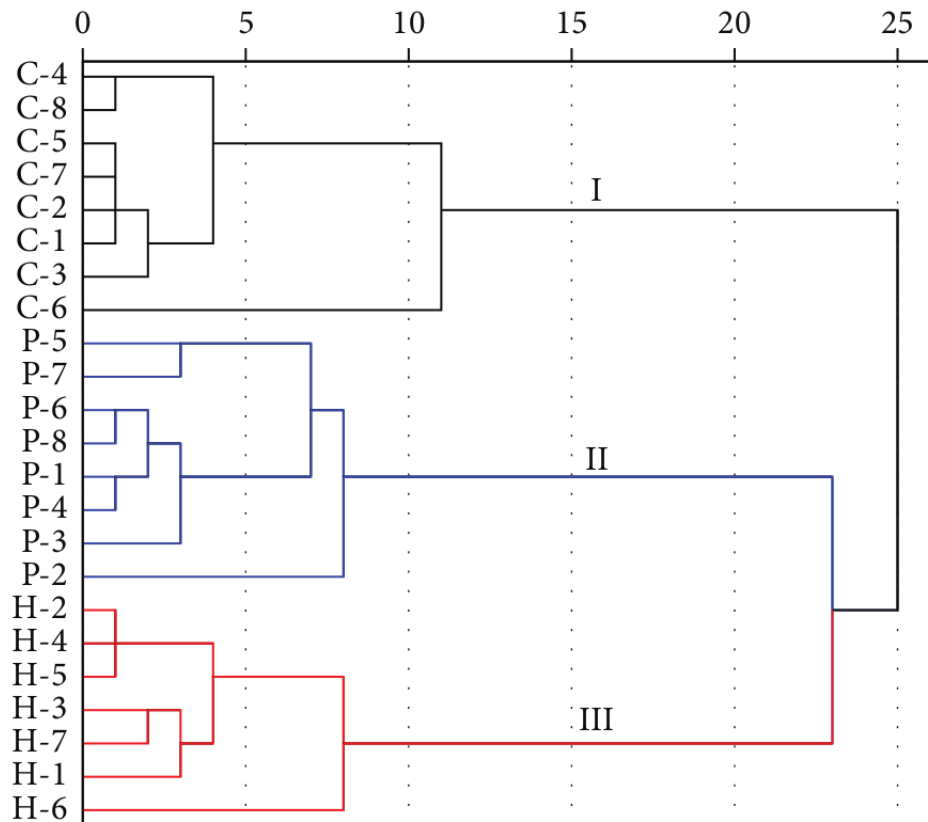


Рис. 2: Пример кластеризации данных УФ-спектроскопии *Wolfiporia extensa*, полученных из 3 регионов произрастания.

как расстояния между объектами, которые в них содержатся.

2. Находят наиболее близкую пару кластеров и объединяют их в один кластер. Таким образом остаётся $N-1$ кластеров.
3. Рассчитывают расстояние между новым кластером и каждым из оставшихся.
4. Повторяют шаги 2 и 3 пока все объекты не окажутся в одном кластере размера N .

По итогу данной процедуры получают иерархическое дерево, отражающее близость объектов выборки друг к другу (Рисунок 2). Анализ данного дерева позволяет выявлять группы похожих между собой и отличных от остальных объектов. Рассмотрение получаемых таким образом группировок позволяет лучше понять объекты исследования. Так как критерием близости кластеров в процессе объединения является расстояние, выбор способа измерения расстояния составляет важный аспект применения кластеризации. Не существует универсальных критериев и рекомендаций по выбору метрик расстояния, однако чаще всего применяют

Евклидово расстояние или расстояние Махаланобиса [88].

Типичным примером использования НСА приведен в работе [29], где иерархический кластерный анализ применяли к данным химического анализа корней женьшеня методами молекулярной спектроскопии. Спектры ближней инфракрасной спектроскопии (БИК) в диапазоне $12000 — 4000 \text{ см}^{-1}$, ИК-спектроскопии диффузного отражения в диапазоне $4000 — 400 \text{ см}^{-1}$ и спектроскопии комбинационного рассеяния в диапазоне $3700 — 100 \text{ см}^{-1}$ использовали для различения трёх типов женьшеня (*Radix Ginseng*, *Radix Ginseng Rubra* и *Radix Panacis Quinquefolii*) и двух типов псевдоженьшеня (*Radix Codonopsis* и *Radix Platycodi*). В результате применения кластеризации к производным спектрам второго порядка было получено прямое разделение данных на 4 кластера, три из которых соответствовали 3 типам женьшеня и один кластер, в который попали образцы псевдоженьшеней. Простота, скорость и неразрушающий (в случае БИК-спектроскопии) анализ позволили авторам сделать выводы о конкурентности их методологии по отношению к традиционным и хроматографическим методам анализа. Аналогичный подход с использованием средневолновой ИК-спектроскопии ($4000 — 400 \text{ см}^{-1}$) применяли уже для различения образцов листьев трёх различных родов растений: *Ranunculaceae*, *Plumbaginaceae*, и *Leguminosae* [5]. Листья растений измельчали с использованием жидкого азота и прессовали с KBr. Благодаря различному липидному метаболизму и содержанию углеводов, а также различным конформациям белков листьев, кластеризация показала надежное разделение для листьев различных видов по трем кластерам рассматриваемых родов. Было также обнаружено устойчивое разделение образцов внутри вида по признаку места сбора.

Интересно также обратиться к случаям, где специально проводили кластеризацию выборок образцов различного географического происхождения. Отличные друг от друга климатические условия регионов произрастания приводят к значительно отличному химическому составу растений и, соответственно, определяют их различную пригодность как сырья для производства лекарственных препаратов. Этим вызвано значительное количество работ, где машинное обучение применяют для установления происхождения сырья. Так в работе [89] проводили неразрушающий анализ плодов рода *Forsythiae* с использованием БИК-спектроскопии. 133 образца плодов собранных в трех провинциях Китая подвергли иерархической кластеризации с использованием вторых производных ИК-спектров и сглаживания

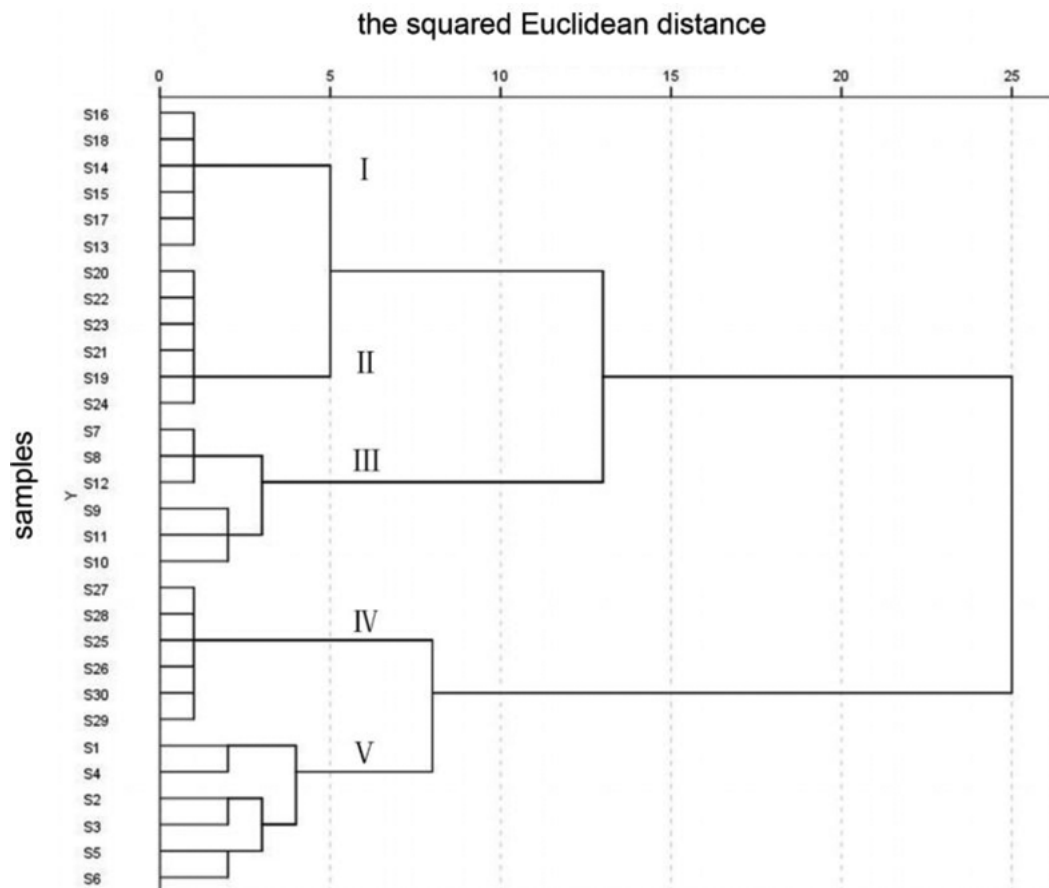


Рис. 3: Дендрограммы, полученные на основе данных анализа 30 образцов методом средневолновой ИК-спектроскопии в диапазоне 1800-600 см⁻¹ [90].

Норриса. Несмотря на то, что большая часть спектров была корректно разнесена по трем кластерам, часть выборки оказалась не в «своих» кластерах. В работе [33] НСА использовали для рассмотрения различий в химическом составе образцов черного перца, а в [90] для сравнения образцов *Gentiana rigescens* подвергнутых различной технологической обработке (Рисунок 3).

По приведённым примерам легко увидеть достоинства кластерного анализа. С его помощью можно быстро и достоверно исследовать насколько тот или иной метод анализа позволяет разделить различные объекты анализа в соответствии с признаками, которые важны аналитику. Другими словами, кластерный анализ выявляет, содержат ли поданные на вход данные аналитическую информацию, необходимую чтобы дать ответ о целевых признаках объекта.

Другим важным методом обучения без учителя является метод главных компонент. Он относится к проекционным методам и классически применяется как метод понижения размерности данных [91]. PCA отчасти схож с кластеризацией, так как позволяет обнаружить закономерности и группировки в данных без какой либо предварительной информации о состояниях объектов в выборке (Рисунок 4). С математической точки зрения PCA представляет собой проекцию пространства

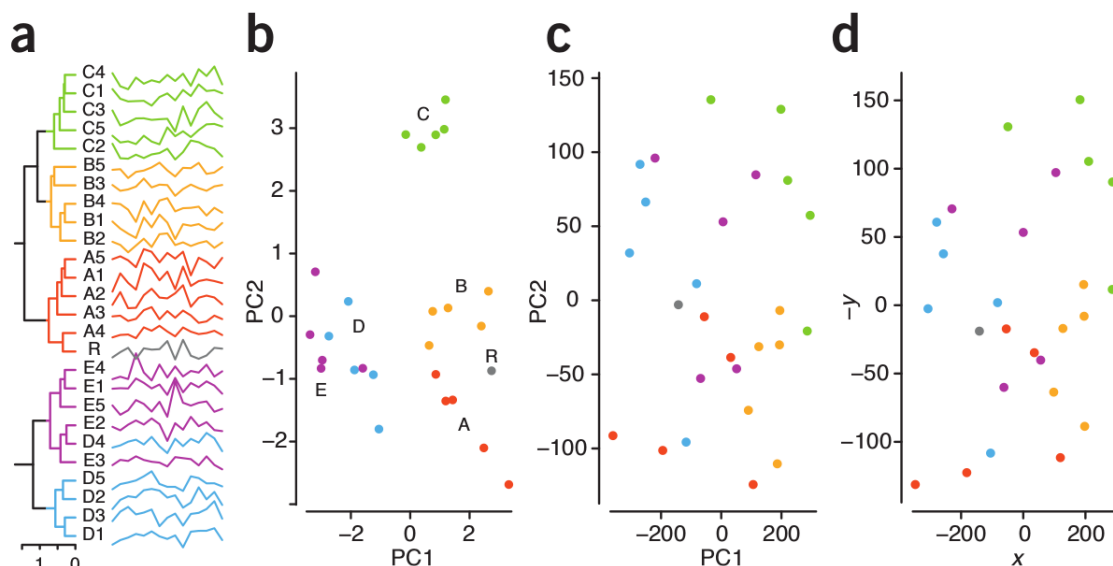


Рис. 4: PCA способен помочь обнаружить группировки объектов в данных. (а) НСА дендрограмма профилей экспрессии 26 генов. (б) При визуализации в координатах двух первых главных компонент можно наблюдать группировку схожую с результатами ИКА. (с) PCA не инвариантна относительно масштаба. График в координатах тех же PC1 и PC2, только с масштабированием двух первых переменных с коэффициентами 200 и 300 соответственно. (д) График двух масштабированных переменных без учета остальных 13. Группировка осталась практически неизменной, т.к. PCA ориентирована более на переменные с большей абсолютной величиной.

данных в новое пространство меньшей размерности, обладающее определёнными свойствами.

В качестве примера использования PCA для видовой дискриминации можно привести работу [92], где с использованием БИК-спектроскопии анализировали кору двух видов рода *Phellodendron*. Оба данных вида используются в традиционной китайской медицине под одним названием. На основании сравнения полученных спектров авторы выбрали диапазон от 4082 до 4545 см⁻¹ для преобразования через PCA. Выбранный диапазон позволил получить 100% корректное разделение образцов по двум классам. В [93] PCA применяли на данных ЯМР-спектроскопии и ВЭЖХ-УФ для видовой дискриминации афродизиаков из Бразилии. ЯМР-спектроскопию под "магическим" углом использовали для анализа различий 4 разновидностей *Hancornia speciosa* методами PCA и НСА [94]. В аналогичном исследовании PCA использовали для различения 26 образцов корней 5 видов растений рода *Angelica* [95]. При схожих названиях, данные растения обладают весьма различным составом и физиологическими эффектами при приёме,

поэтому их корректное различение — важная аналитическая задача. С использованием фильтра Савцкого-Голея получали производные спектры 2 порядка, которые далее использовали в методе PCA. Проекция PCA на двумерную плоскость показала наличие 5 явных кластеров образцов. Наибольший вклад в нагрузки PCA внёс диапазон 1600-900 см⁻¹. В работе [96] PCA использовали для идентификации препаратов "сырого" и обработанного различными способами *Rheum rhabarbarum* на основе данных ВЭЖХ-МС высокого разрешения. PCA применяли также на данные ВЭЖХ-МС определения 17 лимонOIDов использовали для дискриминации различных частей плода *Xylocarpus granatum* [97].

Популярно также использование методов обучения без учителя для хемотаксономических исследований, в том числе для подтверждения данных генетического анализа. Так, в работе [98], с использованием ИК-спектроскопии и ультравысокоэффективной жидкостной хроматографии с масс-спектрометрическим детектированием (UHPLC-MS) проводили хемотаксономический анализ 70 образцов принадлежащих 9 видам рода *Paris*. Применение PCA к данным ИК-спектроскопии позволило увидеть явное разбиение на 3 группы, состоящие из 5, 3 и 1 вида, что было подтверждено после рассмотрения химического состава экстрактов полученного по данным хроматомасс-спектрометрического анализа. PCA также показал наличие явные различия внутри группы из 3 видов, делая их индивидуальное различение на основе только ИК-данных возможным. Хемотаксономическое исследование 9 видов рода *Gentianaceae* с приложением PCA к данным ВЭЖХ-МС (Рисунок 5) и ИК-спектроскопического анализа было осуществлено в [99].

В [100] PCA использовали для поиска корреляций в содержании 8 вторичных метаболитов между образцами 17 видов *Hipericum*, а в [101] для рассмотрения различий между 5 видами широко используемых лекарственных ароматических растений (Рисунок 6). Также была опубликована работа о видовой идентификацию *Actaea racemosa* на основе PCA как способа определения поддельных препаратов [102].

Интересно обратиться также к работе [103], где авторы использовали ИК-спектроскопию нарушенного полного внутреннего отражения и классический вариант ИК-спектроскопии пропускания для анализа 813 образцов цветочной пыльцы принадлежащих 300 видам растений. Для анализа полученного массива данных был применен как иерархический кластерный анализ, так и PCA. Оба метода бы-

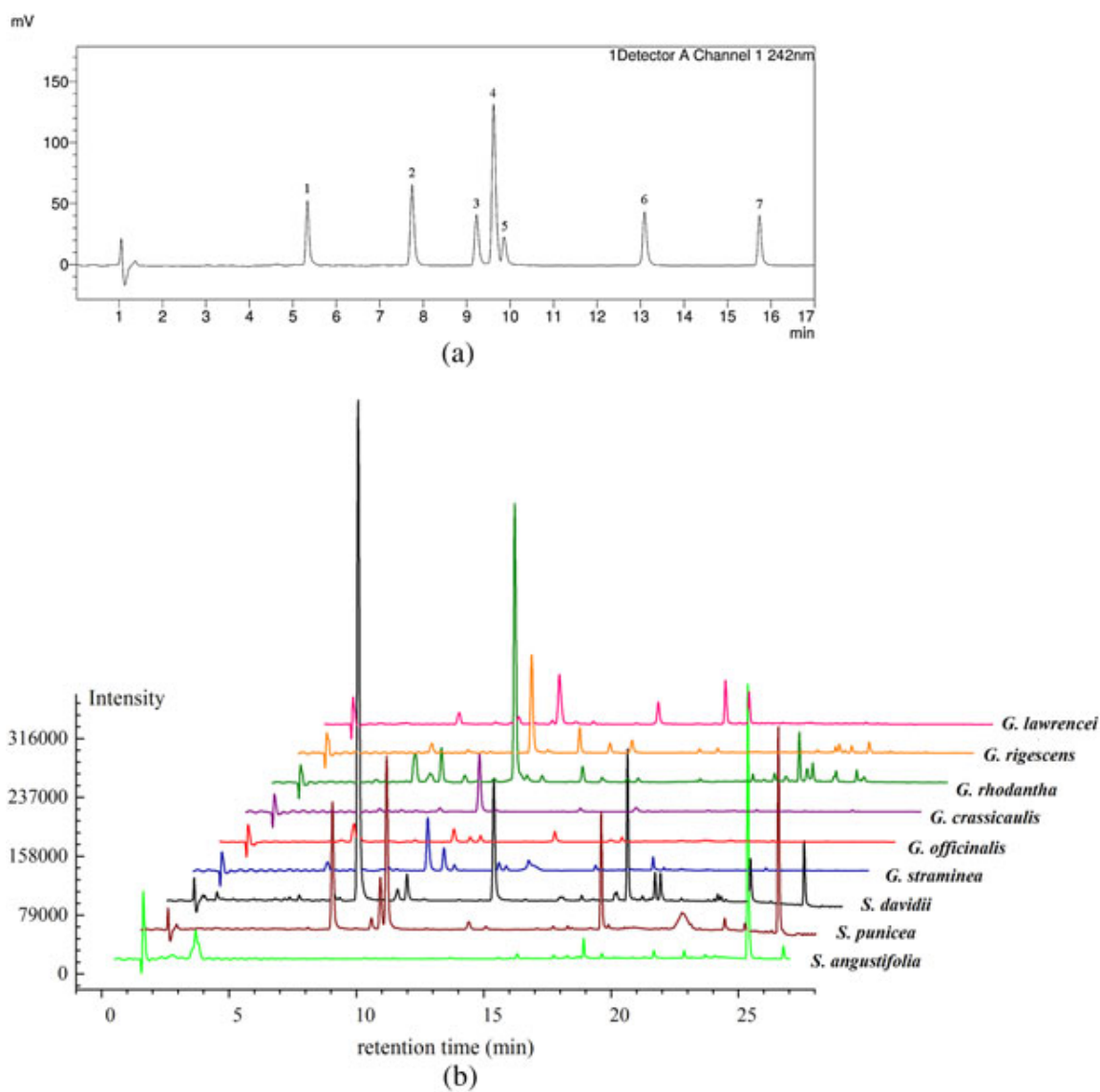


Рис. 5: Хроматограммы смеси стандартов (а) и 9 видов *Gentianaceae* на длине волны 242 нм (b). Пример, когда химические данные для алгоритма получены методом ВЭЖХ [99].

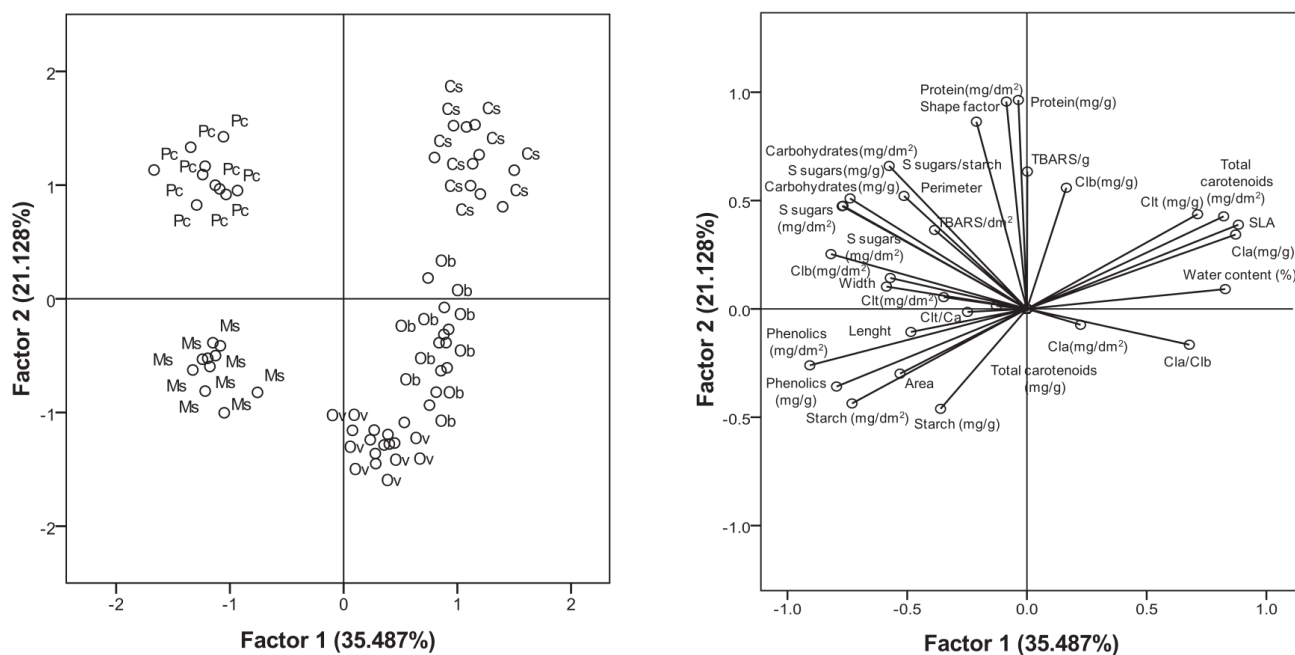


Рис. 6: (слева) PCA для 5 видов лекарственных растений: Cs – *Coriandrum sativum* L.; Ms – *Mentha spicata* L.; Ob – *Ocimum basilicum* L.; Ov – *Origanum vulgare* L.; Pc – *Petroselinum crispum* Mill.. (справа) Композиция двух первых главных компонент [101].

ли применены для анализа вариаций содержания питательных веществ в пыльце, в особенности триглицеридов и белков, рассмотрения различий стратегий опыления, межвидовых и межкладовых различий. По итогам анализа данных с использованием PCA и НСА был сделан вывод о необходимости включать также мужскую пыльцу в состав выборок для экологических и эволюционных исследований. Совместное использование PCA и НСА было также использовано для видовой дискриминации летучих масел различных лекарственных растений произрастающих в Турции [104], видов *Ficus* [105] и *Coptidis* [106], подвидов *Eugenia uniflora* L. [107] (Рисунок 7), а также для дискриминации образцов имбиря по региону произрастания [108].

В [109] PCA и НСА применяли для оценки качества 30 образцов *Curcuma longa* на основе данных УФ-, ИК- и ЯМР-спектроскопии, а в [110] для оценки корреляции антиоксидантной активности и данных ВЭЖХ-УФ анализа образцов *Matricaria chamomilla*. По идентичной схеме PCA также приложили к определению возраста образцов *Gentiana rigescens* [111] и образцов рода *Dendrobium* [112]. Можно отметить также использование PCA в процессе разработки экстракционного метода для получения репрезентативного профиля растений вида *Eurysoma longifolia* [113] (Рисунок 8).

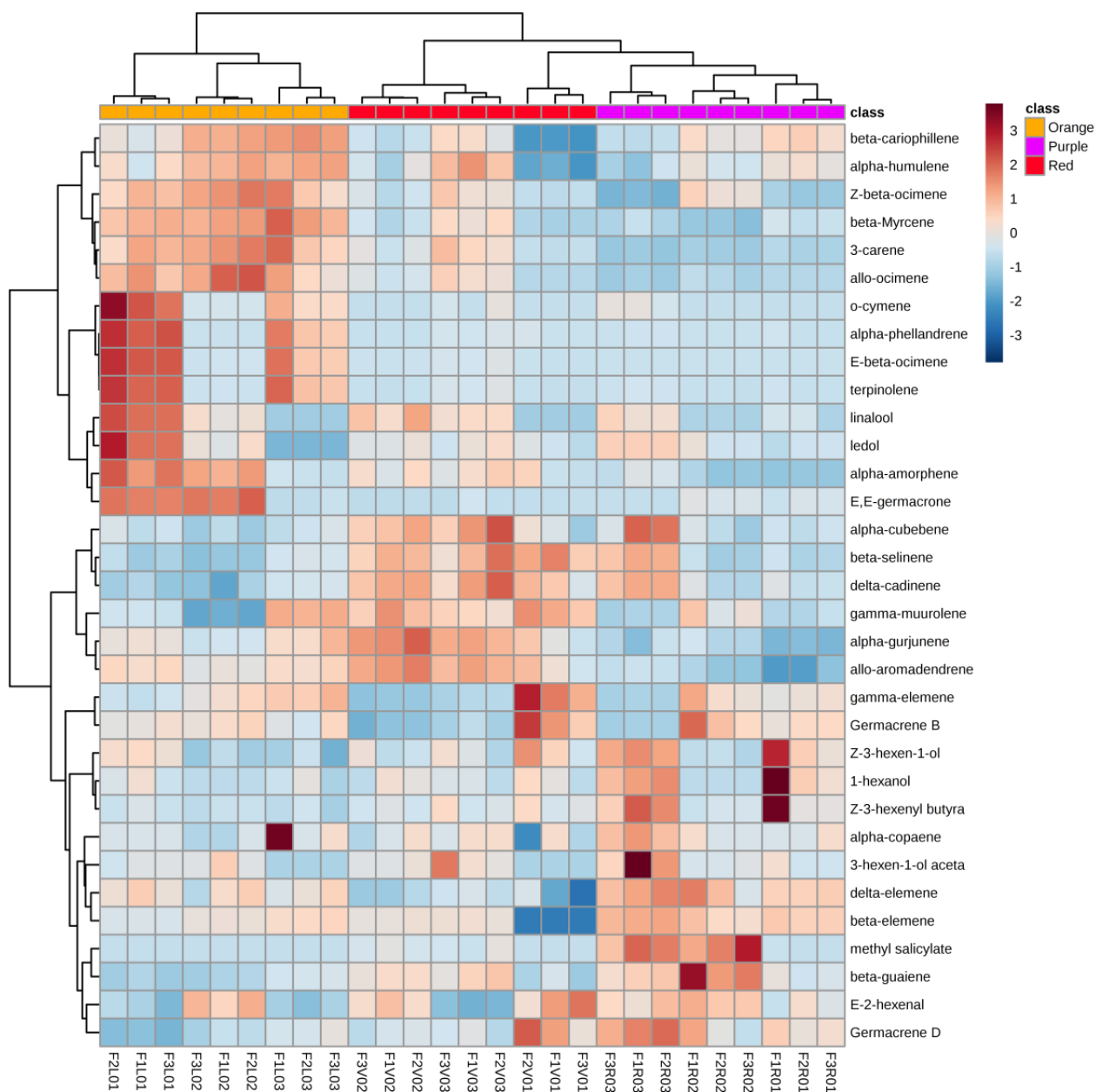


Рис. 7: Иерархический кластерный анализ: дендрограмма и матрица переменных (концентраций летучих органических веществ) для листьев красного, фиолетового и оранжевого сортов *E. uniflora*. Визуализация исходной матрицы переменных иногда позволяет в при небольшом числе переменных наглядно отобразить наиболее важные переменные и соответствующие им параметры объектов рассмотрения.

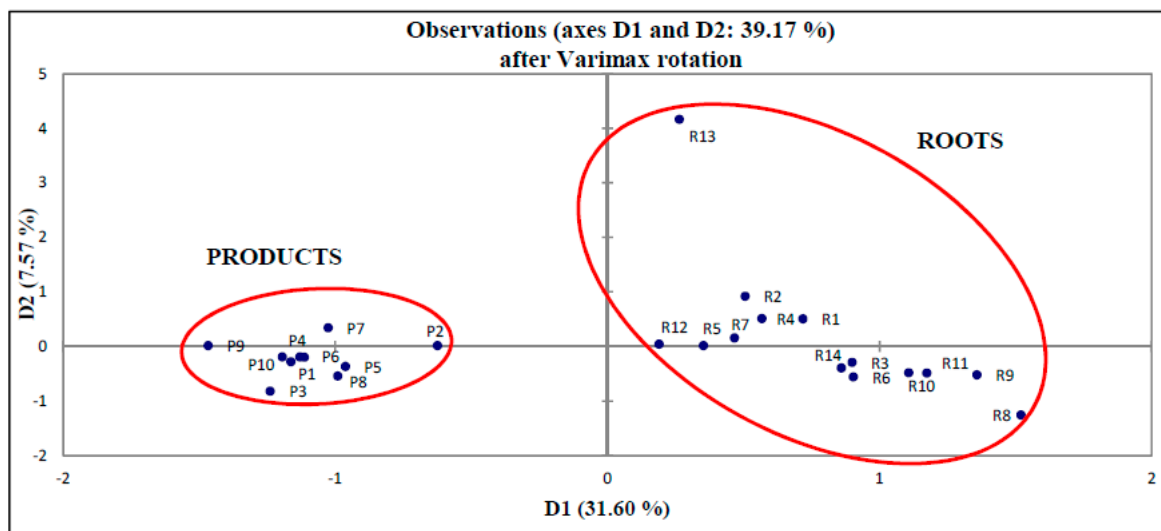


Рис. 8: Визуализация результатов применения PCA к анализу корней и продуктов на основе *E. longifolia roots*.

Таким образом, в хемометрике методы обучения без учителя оказываются наиболее полезны в случае поиска различных закономерностей в массиве данных. Подобные стратегии исследования часто применяются на предварительных стадиях для исследования какого-либо класса объектов и данных. Также, свойство PCA понижать размерность данных часто оказывается удобно для проекции многомерных данных на двумерную или трехмерные поверхности и получения наглядной визуализации объектов внутри исследуемой выборки. Более того, PCA иногда используют как удобную отправную точку для дальнейшего использования методов обучения с учителем на данных меньшей размерности [42, 114, 115].

Безусловно, подобные подходы по анализу выборок и массивов данных не лишены некоторых ограничений (Рисунок 9). Так, например, PCA для матрицы данных рассчитывается таким образом, чтобы отобразить как можно большую часть вариации данных в пространстве меньшей размерности (меньшем числе переменных). Однако важная химическая информация об объектах может содержаться в переменных с малой относительной долей вариации. В таких случаях применение PCA может приводить к потере критической информации, необходимой для корректного различения объектов внутри выборки [116].

Ещё один важный момент заключается в том, что в случае обнаружения явного разделения данных по группам (кластерам) после применения методов обучения без учителя, авторы часто склонны делать выводы о прямой применимости

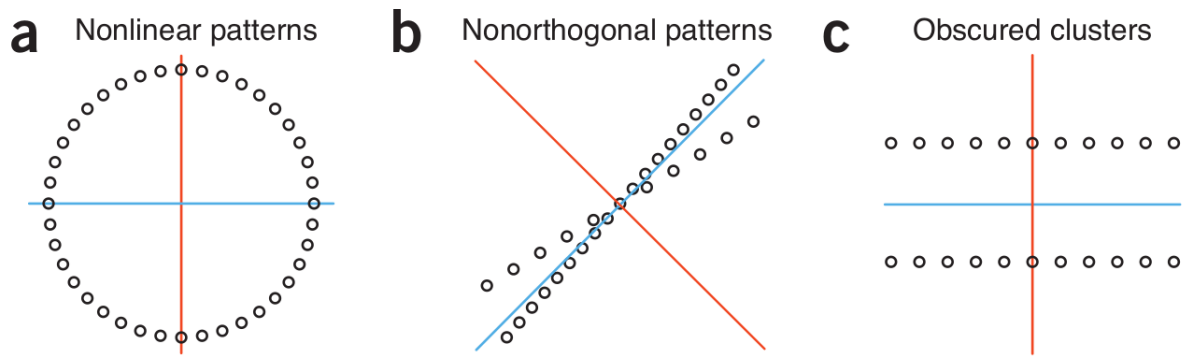


Рис. 9: Принцип расчета PCA налагает значительные ограничения на область его применения. Это ярко проявляется при наличии нелинейных особенностей в структуре данных (a); структуры данных, неортогональные предыдущим главным компонентам (Principal component, PC) могут остаться незамеченными (b); PC1 и PC2 могут быть не в состоянии разделить очевидные группировки объектов (c) [116].

предложенного подхода для дискриминации изучаемых групп объектов. Данный вывод далеко не всегда оправдан, так как многие работы проводятся в условиях очень ограниченных по размеру выборок, вариация внутри которых необязательно отражает реальную ситуацию в генеральной выборке.

1.2.2. Обучение с учителем

Данный раздел машинного обучения включает алгоритмы, которым для решения какой-либо задачи на вход подаётся заранее размеченная выборка данных, для которой разметка данных уже известна. Задачей алгоритма в таком подходе является выработка критериев для подобной разметки. Машинное обучение с учителем можно разделить на два больших класса задач: классификация (идентификация) и регрессия. В случае классификации алгоритм должен идентифицировать принадлежность поданного на вход объекта к одному из заданных классов. Для обучения на вход алгоритма подаются уже размеченные данные, где метка класса была проставлена вручную. Примером в данном случае может служить

выборка образцов растений, видовая принадлежность которых была установлена экспертами-ботаниками. Имеющаяся выборка данных делится на обучающую и тестовую в соотношении 70/30 или 80/20 [117]. Алгоритм далее обучается по предоставленной обучающей выборке за счёт выявления и обобщения внутренней структуры данных и нахождения её связей с тем или иным классом. Затем, на этапе проверки алгоритму подают на вход данные из тестовой выборки для оценки эффективности рассчитанной модели. Существуют различные стратегии для валидации модели и усреднения разбиений данных [117, 118].

В приложении к растительному сырью и препаратам на основе лекарственных растений такие классы могут быть представлены сырьем различного качества, чистоты или географического происхождения, близкородственными видами, и т.п. Практически любая задача аналитического контроля, где требуется принимать решение о соответствии объекта анализа какому-либо критерию по результатам химического или физико-химического анализа, может быть интерпретирована в терминах машинного обучения. Регрессионная задача аналогична по своей сути задаче классификации, однако вместо меток классов используются значения переменных (чаще всего концентрации химических соединений), которые необходимо рассчитать по входным данным. Регрессию в хемометрике классически используют в случае спектральных данных, например данных ИК-спектроскопии и спектроскопии комбинационного рассеяния.

Задача классификации

В качестве простого примера можно привести работу по идентификации 3 видов рода *Ephedra* на основе данных БИК-спектроскопического анализа измельченного порошка цельных растений [119]. После предобработки полученного массива ИК-данные были использованы для построения алгоритма идентификации. Линейный дискриминантный анализ (LDA, Linear discriminant analysis), Самоорганизующиеся Карты (SOM, Self-Organizing Map) и искусственные нейронные сети с обратным распространением ошибки (BP-ANN, Back Propagation-Artificial Neural Network) были проверены на применимость к решению данной проблемы. Для LDA эффективность посчитанных моделей колебалась в диапазоне 84%-92%, с чуть более низкими коэффициентами для двух оставшихся методов. В данном примере можно увидеть, что несмотря на относительную простоту вычислитель-

ной задачи (малое число классов), авторам не удалось получить высоких показателей правильности идентификации ни в одном из примененных методов. Такая ситуация с высокой вероятностью указывает на то, что подаваемые на вход в алгоритмы данные не содержали достаточное количество химической информации, необходимое для однозначного определения видовой принадлежности. При условии, что авторы не «потеряли» важную информацию в процессе предобработки, можно заключить что использованный метод анализа непригоден для эффективного решения поставленной в исследовании задачи. Более успешным применением схожего подхода можно считать работу [120], где проводили классификацию образцов сырья растения *Ganoderma lucidum* по признаку региона произрастания. Методом анализа в данной работе также была БИК-спектроскопия. В качестве данных использовали исходные ИК-спектры и их производные спектры первого и второго порядков. Применив дискриминантный анализ на основе метода проекции на латентные структуры (PLS-DA, Discriminant Partial Least Square — Discriminant Analysis) [121, 122] и LDA были получены значения эффективности на тестовой выборке 100% и 96.6% соответственно. Относительно количества опубликованных работ, дискриминация образцов по признаку географического происхождения является одним из популярных применений PLS-DA в хемометрике. Так, его использовали для классификации образцов проанализированных методами ЯМР-спектроскопии [123], УФ- [124] и ИК-спектроскопии [125], ГХ-МС [126, 127] и ВЭЖХ-МС [128, 129, 130, 131, 132, 133, 134], а также спектроскопии комбинационного рассеяния [135]. Одно из явных преимуществ PLS-DA – ранжирование переменных по их вкладу в разрешающую способность данных (Рисунок 10). Так как в описываемых работах данные в большинстве случаев представляют собой непосредственное отражение химического состава проб, то упорядочение переменных по степени их вклада в различие тех или иных групп позволяет проводить более глубокий анализ химической природы изучаемых явлений.

В другой работе PLS-DA и LDA применяли уже для дискриминации образцов культивированного и дикого *Ganoderma lucidum* [115]. Аналогичным образом PLS-DA также применяли для видовой идентификации растений рода *Chrysanthemum* [136]. Данные БИК-спектроскопии 139 (92 для обучающей выборки и 47 для тестовой) образцов от 3 разных видов использовали для построения классификационного алгоритма. В [137] PLS использовали для различения образцов *Panax ginseng* пяти- и шестилетнего возраста на момент сбора, а в [138] для различения

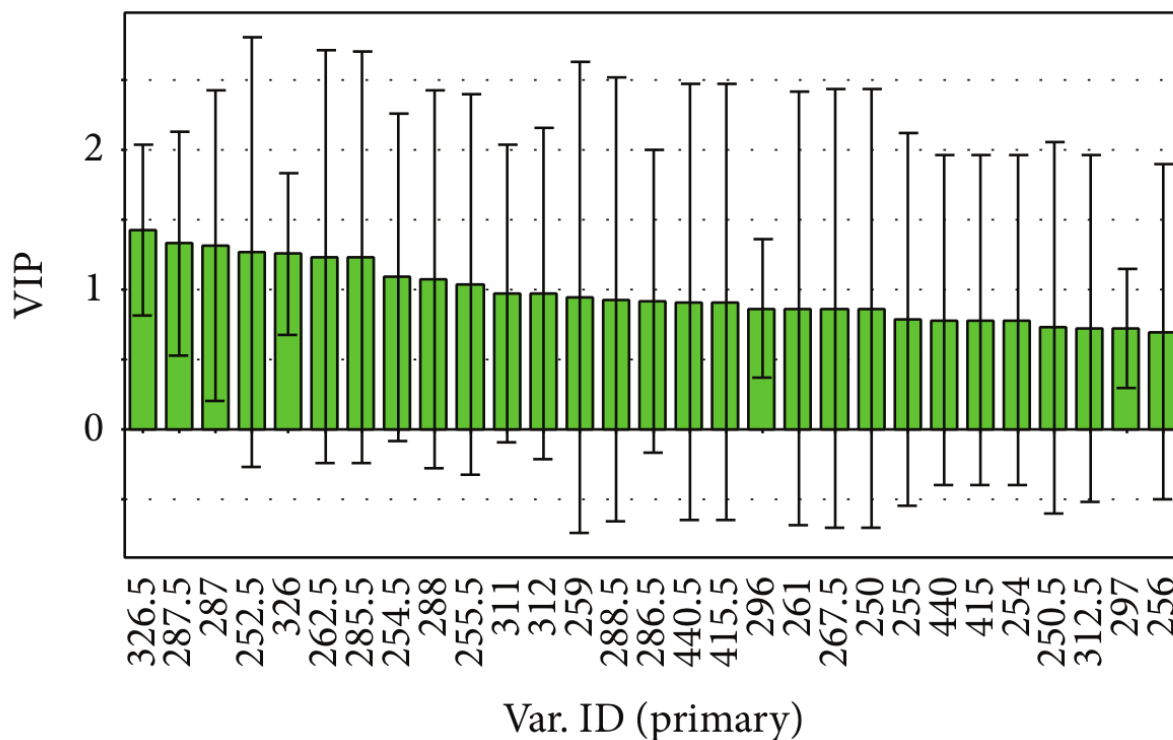


Рис. 10: График вклада (variable importance in projection, VIP) переменных в эффективность классификации методом PLS-DA. [124].

образцов *Areca catechu* подвергавшихся разным способам экстракции.

Идентификационные алгоритмы на основе PLS-DA также использовали для дискриминации видов рода *Chamomile* [139], *Sceletium* [140], а также сортов винограда [141]. Совместные данные ВЭЖХ-МС и ЯМР¹ анализа применяли для определения близкородственных примесных видов в сырье *Harpagophytum procumbens* [142]. Тонкослойную хроматографию в комбинации с PLS-DA применяли для различения видов *Agathosma betulina* и *Agathosma crenulata* (Рисунок 11) [143]. Случайный лес (RF, Random forest) [144] и PLS-DA применили для различения настоящих и поддельных образцов масла амазонского растения *Carapa guianensis* методом ИК-спектроскопии [145].

Интересно привести также работу [4] где различные внешние условия для растений симулировались самими исследователями. В качестве данных использовали масс-спектрометры прямого ввода, собранные для экстрактов листьев *Pharbitis nil*. В процессе роста растения были подвергнуты 6 различным вариантам длительности светового дня. Метод иерархической кластеризации не привел к успешному разделению групп образцов, поэтому авторы перешли к обучению с учителем. С использованием генетического программирования (GP, Genetic Programming) им

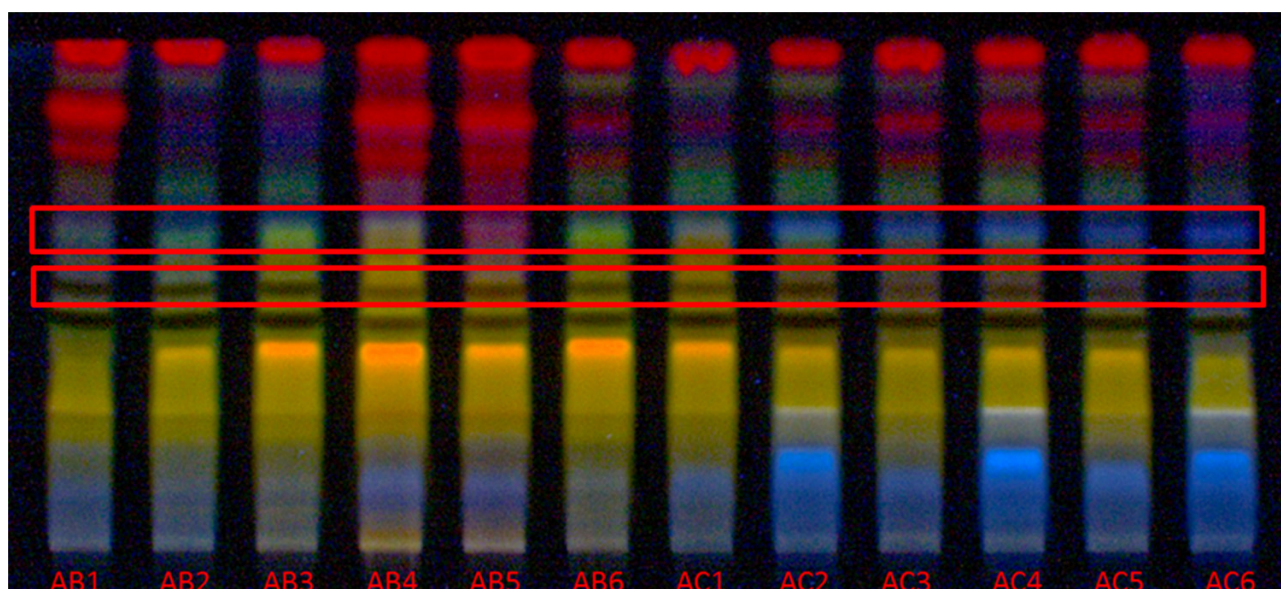


Рис. 11: ВЭТХ профили экстрактов *Agathosma betulina* (AB1-AB6) и *A. crenulata* (AC1-AC6) при освещении с длиной волны 366 нм. Первичные данные, использованные для разработки алгоритма в [143].

удалось успешно построить алгоритм, способный устойчиво различать образцы, подвергнутые наиболее длительному (1 неделя против 2 дней у ближайшей группы) периоду сокращённого светового дня. Благодаря особенностям генетического программирования также были идентифицированы метаболиты, вносящие наибольший вклад в разницу метаболических профилей при различной длительности светового дня. Таким образом, в данной работе была показана возможность различать физиологические состояния растения на основе экспресс анализа, причём в том случае, когда метод главных компонент и кластеризация не позволяли увидеть какой-либо «полезной» структуры данных.

Ещё один популярный метод классификации объектов — метод опорных векторов (SVM, Support Vector Machine) [146]. Принцип SVM состоит в разграничении многомерного признакового пространства (вектор данных объекта рассматривается как точка в n -мерном пространстве) на области, соответствующие отдельным классам. После построения модели по обучающей выборке алгоритм проверяет в какой области пространства оказывается новый неизвестный объект и на основании этого приписывает ему класс.

SVM использовали в работе [147] для идентификации 30 экстрактов 6 сортов чая (по 5 образцов на каждый сорт). Отличительной особенностью данной работы является использование ВЭЖХ-УФ анализа для получения выборки химических

данных. Данными для классификации служила полная хроматограмма на длине волны 280 нм после применения различных алгоритмов выравнивания и сглаживания. Применение PCA позволило авторам увидеть явные различия в химическом составе, позволяющие отличить каждый сорт чая. Тем не менее, величина данных различий оказалась недостаточной для надежного решения данной задачи. После применения SVM правильность идентификации по некоторым сортам составила не более 80%. Лучше всего показал себя алгоритм RF, где удалось достичь наибольшей правильности идентификаций по всем сортам.

Более высокая эффективность при использовании SVM была получена при исследовании хроматографических профилей трёх видов корей из рода *Cirsium* методами одномерной и двумерной газовой хроматографии, а также ВЭЖХ [148]. В случае одномерной хроматографии SVM показал высокую предсказательную эффективность (95% на тестовых выборках), тогда как в случае двумерной газовой хроматографии его эффективность составила менее 80%. При использовании объединённых данных с ГХ и ВЭЖХ SVM показал 100% правильность предсказания на тестовой выборке. Данный пример отражает случай, когда различные методы анализа содержат комплементарную химическую информацию. При объединении данных в подобном случае можно получить высокоэффективный алгоритм, даже если индивидуальные эффективности методов не очень высоки.

Также интересна работа по дискриминации 3 сортов *Panax ginseng* на основе спектрофотометрии и БИК-спектроскопии (376-1025 нм) [114]. На первом этапе PCA применяли для понижения размерности данных, после чего строили классификатор на основе SVM. На выборке размером 78 образцов авторам удалось получить 100% правильность идентификации. С высокой эффективностью SVM и PLS-DA использовали также для видовой и географической дискриминации грибов рода *Boletaceae* комбинируя данные спектрофотометрии и ближней ИК-спектроскопии [149]. С применением неразрушающего анализа (БИК-спектроскопия) и LDA была разработана экспрессная методика дифференциации между плодами подвидов *Euterpe oleracea* [150]. Правильность идентификации составила 93.2% на тестовой выборке. Ещё одним популярным в хеометрическом анализе методом является (SIMCA, Soft Independent Modeling of Class Analogy) [151]. Данный метод является прямым продолжением PCA и использует новое линейное пространство полученное после проекции исходных данных. В этом

линейном пространстве определяются границы, в которых наиболее высока вероятность обнаружить образец какого-либо класса. Особенностью данного метода является его то, что неизвестный образец может быть размечен классификатором на основе SIMCA как принадлежащий одновременно нескольким классам. Метод SIMCA весьма популярен в хемометрике и широко используется для решения различных задач [152].

SIMCA использовали для классификации 140 образцов *Lonicera japonica* собранных в 7 разных провинциях Китая [153]. Для всех образцов были получены БИК-спектры в диапазоне 10000 — 4000 см⁻¹ с шагом 4 см⁻¹ (1500 точек). Построенный на основе SIMCA классификатор обладал 100% точностью предсказаний в пределах тестовой выборки. В [154] SIMCA совместно с УФ-спектроскопией использовали для видовой дискриминации 50 образцов рода *Thyme*. SIMCA и PLS-DA использовали на данных спектроскопии нарушенного полного внутреннего отражения для определения содержания 5 видов лекарственных растений в порошках [155].

Так например в работе [156] описывается работа по идентификации видов рода *Paris*. Данный род насчитывает 27 видов, произрастающих в основном в Европе и Восточной Азии. Однако в китайской фармакопее официально числятся только корневища видов *Paris polyphylla* var. *chinensis* и *P. polyphylla* var. *Yunnanensis*. Видовая идентификация высушенных корневищ данного рода традиционными визуальными морфологическими методами весьма затруднительна, и вообще невозможна в случае порошковых образцов. Комбинация БИК и метода Проекции на Латентные Структуры позволила авторам сделать оценочный обзор различий и схожестей между спектрам видов внутри данного рода. Результаты также показали, что присутствие в образцах диких разновидностей *Paris* оказывает значимый эффект на БИК спектры. Виды *P.cronquistii* var. *xichouensis*, *P. caobangensis*, *P.cronquistii*, *P.polyphylla* var. *alba* и *P.polyphylla* var. *pseudothib* оказались чётко отличимы от своих соседей по роду.

Кудо и др. [157] исследовали приложение БИК спектроскопии к быстрой идентификации *Digitalis purpurea* на фоне четырех близкородственных видов (*Digitalis lanata*, *D.mertonensis*, *D.ambigua*, and *D. orientalis*). Для интерпретации спектральной информации были использованы 5 различных подходов: метод максимального расстояния в пространстве длин волн, корреляция в пространстве

длин волн, метод коэффициентов корреляции, графический метод откладывания двух длин волн и идентификация методом ближайшего соседа. Было обнаружено, что метод максимального расстояния в пространстве длин волн позволяет наиболее эффективно идентифицировать *D. purpurea*, далее по эффективности идёт графический метод откладывания двух длин волн, который также применим для распознавания закономерностей и даёт хорошее графическое представление различий между видами. Использование значений коэффициентов корреляции с другой стороны, оказалось мало полезным при различении видов.

Radix puerariae – важное растение в восточной медицине. Оно широко применялось для лечения диареи, острой дизентерии, глухоты и сердечно-сосудистых заболеваний [158]. В китайскую фармакопею начиная с 2000 официально входили два вида *Radix puerariae* - *Radix puerariaelobata ohwi Yege (YG)* и *Radix thomsonii benth Fenge (FG)*. Фотохимическое сравнение показало, что содержание главных физиологически активных компонентов YG и FG сильно отличалось. Начиная с 2005 года данные виды были разделены на два отдельных раздела. Для разработки подхода по различению данных видов была использована БИК спектроскопия в сочетании с LDA и SIMCA [159]. Было построено две кластерные модели – с использованием полного спектра и с использованием двух выделенных регионов (5556-6250 см⁻¹ и 4082-2878 см⁻¹). Было обнаружено, что модели построенные на использовании полного спектра проигрывали в эффективности классификации моделям с использованием выделенных областей спектров. Описано также использование двумерной корреляционной БИК спектроскопии для дискриминации трёх близкородственных видов: *Dendrobium densiflorum Lindl. exWall. (Mihuashihu)*, *D. aurantiacum Rchb.f. var. denneanum (Kerr.) Z. H. Tsi (Dieqiaoshihu)* и *Dendrobium chrysotoxum Lindl. (Guchuishihiu)* [160]. На первом этапе авторы использовали PCA на первичных спектральных данных для того, чтобы подтвердить возможность различения видов с использованием БИК спектроскопии. Затем были получены температурные корреляционные двумерные БИК спектры. По сравнению с одномерным вариантом, двумерная спектроскопия позволяет повысить разрешение, упростить картину наложения полос и предоставить данные о температурной зависимости спектров. Для различных видов *Dendrobium* были обнаружены заметные различия в области 4750-5600 см⁻¹ при использовании синхронных и асинхронных двумерных корреляционных спектров. Данная спектральная область была в дальнейшем использована для построения дискриминационной модели.

В работе Ву с коллегами [161] было показано, что при использовании комбинации БИК спектроскопии и SIMCA удаётся достичь приемлемой эффективности различения азиатского и американского женьшеней. Причём в сравнении с LDA и PLS-DA, SIMCA лучше справился с обнаружением поддельных образцов. Существует ещё один похожий случай комбинирования БИК спектроскопии и SIMCA для решения задачи дискриминирования между видами *Austragali Radix* и *Smilacis Rhizoma* [162]. В контексте решения задач по контролю качества женьшеня, помимо проблем видовой идентификации, в работе [163] на основе графического PCA отображения данных диффузных БИК спектров был создан подход по различению эпидермиса, флоэмы и ксилемы корней растения. Примером задачи с большим количеством классов может служить использование БИК профилирование совместно с LDA, PLS-DA и SIMCA для быстрой идентификации Элеутерококка Колючего [164]. Элеутерококк совместно с другими 8 видами растений принадлежащих и не принадлежащих семейству Аралиевые. В работе были достигнуты хорошие результаты по обнаружению подделок и некачественного продукта. Вышеупомянутые работы относятся к случаю, когда аналитические данные (БИК спектры) загружали в математические алгоритмы для обнаружения статистически достоверных различий между классами. Помимо ручной интерпретации данных различных математических методов возможно также применения машинного обучения. Под машинным обучением в общем случае понимают совокупность математических методов и программных подходов, применяемых для создания алгоритмов по автоматическому решению задач анализа данных. В такой алгоритм загружаются некоторые предобработанные или сырые данные, а алгоритм на выходе даёт ответ о принадлежности объекта к определенному классу. В настоящее время машинное обучение используется для решения широчайшего круга прикладных и исследовательских задач. Существуют алгоритмы по распознаванию лиц, рекомендательные системы в сфере интернет-торговли, алгоритмы по предсказанию движений фондовых рынков и т.д. Несмотря на то, что машинное обучение является так называемым «чёрным ящиком», т.е. не даёт пользователю ответа о конкретных критериях, используемых для вынесения решений, надёжность подобных методов при корректной реализации привела к их взрывному распространению в последние годы.

Интересно отметить также использование нестандартных математических методов. В [165] применили температурно-ограниченные сети каскадных корре-

ляций (TCCCNs, Temperature-constrained cascade correlation networks) для решения бинарной задачи классификации в анализе образцов растения рода *Rheum* методом БИК-спектроскопии. Из 52 образцов, 25 относились к видам, признаваемым фармакопей Китая как официальные и 27 относились к «неофициальным» видам. Несмотря на сложность задачи различения групп видов, метод TCCCNs позволил добиться 100% правильности идентификации на тестовой выборке, обойдя при этом классический метод искусственных нейронных сетей с обратным распространением ошибки.

Канонический дискриминантный анализ применяли для классификации образцов *Mentha pulegium* по признаку места произрастания на основе средневолновой (4000-400 см⁻¹) ИК-спектроскопии [166]. Правильность идентификации составила 90% на тестовой выборке.

Относительно редко в исследованиях также применяют метод k-ближайших соседей (k-NN, k-Nearest Neighbors) [167]. k-NN относится к одним из наиболее простых методов обучения с учителем. В методе k-NN не проводится никакой предварительной оптимизации модели по обучающей выборке, все расчёты производятся уже на этапе классификации неизвестных объектов. Рассчитывается расстояние неизвестного объекта (вектора данных) до всех объектов выборки и класс объекта определяется голосованием k-ближайших соседей. В работе [32] проводили классификацию 31 неизвестного образца лекарственного препарата на основе корней растений рода *Bupleurum* на основе данных тонкослойной хроматографии (ТСХ) и ВЭЖХ-УФ анализа. Только два вида из данного рода одобрены в фармакопее Китая для изготовления данного препарата. Тогда как существует ещё 5 близкородственных видов *Bupleurum*, которые периодически встречаются в поддельных препаратах. С использованием размеченной выборки из 33 образцов аутентичных препаратов. Из-за значительных дисперсий величин пробега веществ в случае ТСХ, k-NN показал более высокую правильность по сравнению с BP-ANN, 100% против 88%. Тогда как в случае менее изменчивого ВЭЖХ-УФ детектирования оба метода показали 100% правильность идентификации. С высокой эффективностью метод k-NN также применили для дискриминации географического происхождения 128 образцов *Marsdenia tenacissima* на основе данных ИК-спектроскопии [168]. На примере Таблицы 1 можно увидеть, что выбор корректной процедуры предобработки данных также играет критическую роль в

приложениях с использованием машинного обучения. Процедура предобработки первичных данных может как позволить убрать шумовой фон из данных, так и наоборот удалить критически важную химическую информацию.

Таблица 1: Зависимость эффективности предсказаний от способа предобработки данных при использовании метода k-NN для определения региона произрастания образцов *Marsdenia tenacissima* [168].

Предобработка	Выборка		Точность (%)		Время (сек)
	Обучающая	Тестовая	Обучение	Тест	
ВК	77 (60%)	51 (40%)	100%	100%	0.08
МКР	77 (60%)	51 (40%)	76.86%	54.26%	0.02
СНКД	77 (60%)	51 (40%)	76.47%	53.49%	0.02
1я-Производная	77 (60%)	51 (40%)	97.65%	97.67%	0.02
2я- Производная	77 (60%)	51 (40%)	91.18%	79.07%	0.11

* - ВК, вэйвлет коррекция ; МКР, мультипликативная коррекция рассеяния; СНКД, стандартная нормальная коррекция по дисперсии

Остается не до конца ясным вопрос корректной предобработки хеометрических данных, т.к. это может приводить к невысокой эффективности конечных алгоритмов [169]. Интересным шагом в данном направлении можно считать работу [170], где авторы исследовали различные варианты обработки ИК-данных на примере 12 видов лекарственных растений - 60 образцов 6 видов рода *Nurpericum* и 40 образцов 6 видов рода *Epilobium*. В итоге было показано, что качество классификации может колебаться в пределах более чем 20%. Такое различие может определять границу между эффективным и непригодным алгоритмами.

Регрессионные методы

Регрессия относится к категории наиболее часто используемых в химическом анализе методов машинного обучения, т.к. построение градуировочной кривой при использовании метода внешнего стандарта обычно проводят с использованием регрессии по методу наименьших квадратов. Однако использование прямых градуировочных зависимостей мало приложимо к ИК-спектроскопии сложных образцов ввиду отсутствия "чистого" сигнала определяемого соединения. Наиболее часто в данном классе задач применяют множественную регрессию на основе метода PLS.

Так, в [166] проводили ИК-спектроскопическое определение содержания пулегона в эфирном масле *Mentha pulegium* с использованием метода PLS. Результаты определения в исследованном диапазоне 157-860 мг/л оказались статистически неразличимы с результатами полученными по методу газовой хроматографии. Прямое определение валового содержания эфирных масел и основных действующих веществ (α -туйона and β -туйона) в *Salvia officinalis* проводили в работе [171]. Метод основан на PLS и не требовал никакой пробоподготовки кроме высушивания листьев, а длительность анализа составила не более 5 мин., тогда как рутинно используемый для этой цели метод ГХ-МС анализа с дериватизацией занимает несколько часов. Более необычный вариант использования регрессии на основе PLS использовали в работе [123]. Регрессию применяли для установления зависимости между данными ЯМР H^1 анализа и результатами оценки качества препарата экспертами-ботаниками (так называемый сенсорный метод определения качества). ЯМР-спектр препарата в области δ 0.78–4.35 м.д. использовался установления к какой из 5 категорий качества относится каждый образец. Кросс-валидационный коэффициент регрессии Q^2 (аналог широко используемого R^2 для случая кросс-валидации) составил 0.984, что говорит о высокой предсказательной способности данной модели. В [172] метод независимых компонент (ICA, Independent Component Analysis) [173] использовали для определения уровней содержания гентиопикрозида и свертиамарина в лекарственном растении вида *Gentiana scabra* (Рисунок 12). На основе производных БИК-спектров авторам удалось получить коэффициенты корреляции 0.85 и 0.95 соответственно.

Авторам [153] удалось получить устойчивый (переносимый на разные ИК-спектрометры) алгоритм для полуколичественной оценки содержания 6 основных активных компонентов лекарственного растения *Lonicera japonica*. Аналогичным образом PLS регрессию использовали в [92] для определения валового содержания алкалоидов в *Cortex Phellodendri* по методу БИК-спектроскопии. Более масштабное исследование было проведено в [174], где использовали БИК-спектроскопию и ВЭЖХ-УФ для анализа образцов 16 видов лекарственных растений произрастающих в Венгрии. PLS регрессию использовали для проверки возможности определения валового содержания фенольных соединений, аминокислот и углеводов по данным БИК-спектроскопии.

Процесс внедрения препаратов сложного состава в практику современной

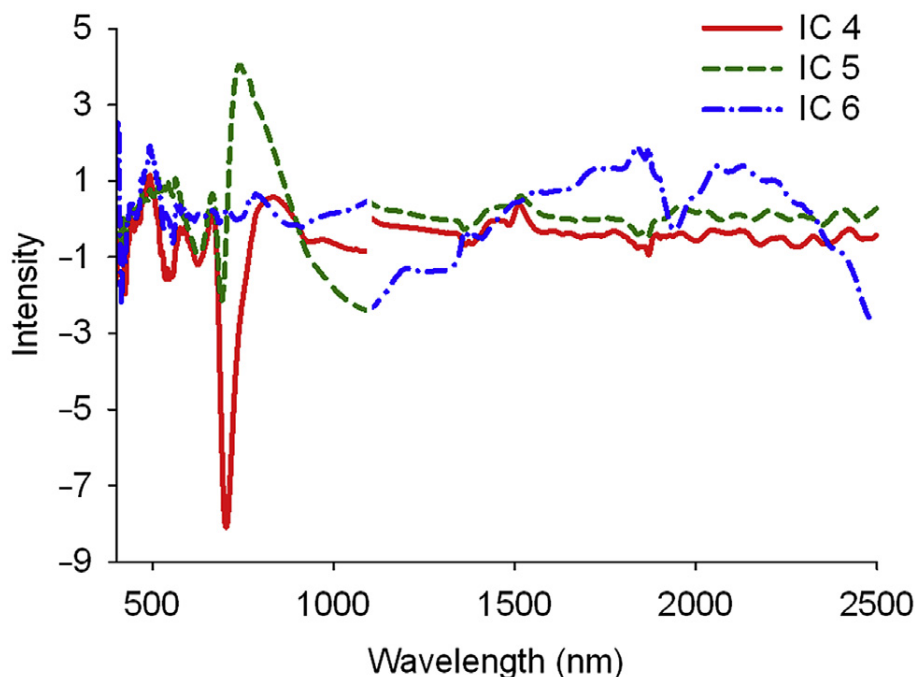


Рис. 12: Три независимые компоненты полученные из исходных спектров порошка *Gentiana scabra Bunge* после мультипликативной коррекции рассеяния. Отображены компоненты, имеющие самую высокую корреляцию с гентиопикрозидом и свертиамарином [172].

фармакологии будет неуклонно вести к всё более широкому применению методов машинного обучения при решении задач контроля качества. Подходы, основанные на строгом количественном анализе каждого целевого соединения будут дополняться новыми гибридными подходами, основанными на обработке «сырых» химических данных методами машинного обучения. Данная тенденция отражает более широкий тренд в аналитической химии, вызванный развитием и повышением доступности аналитических методов, генерирующих большие объёмы данных (масс-спектрометрия высокого разрешения, ЯМР-спектроскопия и т.д.) и их широкого применения для анализа сложных объектов природного и техногенного происхождения (образцы биологических тканей, образцы сточных вод, продукты питания и т.п.). Методы машинного обучения позволяют находить в данных большого объёма информацию, способную помочь ответить на различные вопросы, например различить состояния чистый продукт/продукт с примесью, аутентичный продукт/подделка, качественное/некачественное сырьё и т.п. Такой постепенный переход будет продолжаться ещё многие годы. Дальнейшие исследования в этой области и накапливаемый опыт применения позволят сказать, насколько надёжной альтернативой подобные подходы станут по отношению к классической методологии химического анализа.

Исходя из сделанного обзора можно заключить, что эффективный подход по идентификации видовой принадлежности препаратов растений должен обладать следующими характеристиками:

1. Независимость результатов работы алгоритма от условий сбора ВЭЖХ-МС данных, в том числе: самой ВЭЖХ-МС система, хроматографической колонки, состава подвижной фазы, скорости потока, программы разделения, объёма вкола и т.д.;
2. Простая процедура экстракции с использованием доступных реагентов;
3. Способность работать на масс-спектрометрическом оборудовании низкого разрешения;
4. (опционно) Способность алгоритма распознавать сразу несколько видов в смеси.

Совмещение аналитической химии с машинным обучением требует плотного сотрудничества специалистов химиков со специалистами по машинному обучению, так как эффективные и корректные решения могут быть созданы только с пониманием принципиальной структуры, происхождения и особенностей обрабатываемых данных. Схема исследований за последние 15 лет практически не претерпела значимых изменений. В большинстве случаев авторам удаётся собрать очень ограниченную выборку в пределах нескольких десятков (реже сотен) образцов, которые анализируют методами ЯМР, ИК или ВЭЖХ-МС/УФ. Получаемую выборку данных далее подвергают PCA/НСА, на основе чего делаются какие-либо выводы о сходствах или различиях имеющихся групп, либо делается вывод о перспективности предложенной методологии для осуществления рутинного контроля качества препаратов рассмотренных растений. Во многих случаях авторы напрямую используют какой-либо из методов обучения с учителем (чаще всего какой-либо вариант линейного DA или PLS-DA) для построения классификационного или регрессионного алгоритмов. Несмотря на большой опыт, накопленный научным сообществом в данной области, практическое применение подобных подходов пока не получило сколько-нибудь широкого распространения. Препараты лекарственных растений в большинстве стран мира по-прежнему не подвергаются практически никакому аналитическому контролю. Помимо маленьких размеров

выборок, ещё одной явно подмечаемой проблемой является отсутствие практики у авторов исследований выкладывать в открытый доступ первичные данные, полученными непосредственно после физико-химического метода анализа. Такая ситуация значительно снижает возможности сравнения различных подходов и агрегирование данных для более масштабных исследований. На сегодняшний день явно назрела необходимость создания глобального проекта по стандартизации и сбору аналитических данных в области анализа лекарственных растений.

ГЛАВА 2

Экспериментальная часть

2.1. Оборудование и материалы

В работе использовали следующее аналитическое оборудование: ВЭЖХ-МС/МС систему, состоящую из гибридного tandemного времяпролётного детектора Shimadzu LCMS IT-TOF (Япония) со сферической ионной ловушкой, оснащенного источником электрораспылительной ионизации; и жидкостного хроматографа Shimadzu Nexera X2 (Япония). Регистрацию хроматограмм и конвертацию данных проводили при помощи программного обеспечения LabSolutions (Япония). ВЭЖХ-МС/МС систему, состоящую из tandemного квадрупольного масс-спектрометрического детектора Agilent 6460 Triple Quadrupole (США), оснащенного электрораспылительной ионизацией Jet Stream; и жидкостного хроматографа Agilent 1290 (США). Регистрацию хроматограмм и обработку данных проводили при помощи программного обеспечения Mass Hunter (США).

Хроматографическое разделение проводили на колонке Thermo Hypersil Gold aQ (100 × 2.1 мм), диаметр зерна сорбента 1.9 мкм (Thermo Scientific, США)

Также в работе использовали следующее оборудование для пробоподготовки: Для имитирования условий старения проб использовали термостат Shimadzu СТО-20А (Япония). Для проведения экстракции применяли ультразвуковую ванну «Branson 2210 E-DTC» (Branson», США). Взвешивание точных навесок проводили на весах Explorer Pro (Ohaus Corporation, США). Для центрифугирования образцов использовали центрифугу CM-50 (Elmi, Латвия). Для упаривания растворов использовали ротационный испаритель с автоматизированной вакуумной станцией (Buchi, Швейцария).

Для отбора точной аликвоты использовали автоматические дозаторы 10-100 мкл, 20-200 мкл, 100-1000 мкл с пределом допускаемой погрешности измерения не более $\pm 5\%$ (LABMATE, Польша) и наконечники необходимых объемов. В работе использовали следующие растворы и реагенты: Ацетонитрил (Panreac, HPLC grade, Испания), кислота муравьиная с чистотой $> 98\%$ (Panreac, Испания), деионизированная вода (после очистки системой Milli-Q (Millipore, США)). Рабочие растворы и другие необходимые растворы готовили растворением точных навесок в соответствующих растворителях в день проведения анализа. Растительный материал приобретали через коммерческих производителей.

2.2. Экстракция

1. Для получения водно-этанольного экстракта 3 г. растения измельчали в агатовой ступке и экстрагировали 30 мл 70% этанола в ультразвуковой ванне в течение 1 часа. 2 мл первичного экстракта центрифугировали в течение 10 мин. при 10000g; надосадочную жидкость фильтровали через 0.45 мкм мембранные фильтры. 100 мкл фильтрата разбавляли 1:10 0.1 % раствором муравьиной кислоты. При малом исходном количестве растительного материала массу навески измельченного растительного материала и объём экстракционной смеси уменьшали в 3 раза.
2. Для получения водно-метанольного 1 г. высушенного растительного материала измельчали в агатовой ступке и экстрагировали 10 мл 70% метанола в ультразвуковой ванне в течение 3 часов. 2 мл первичного экстракта центрифугировали в течение 10 мин. при 10000g; надосадочную жидкость фильтровали через 0.45 мкм мембранные фильтры. 100 мкл фильтрата разбавляли 1:10 0.1 % раствором муравьиной кислоты.
3. Для получения водного экстракта 1 г. высушенного растительного материала измельчали в агатовой ступке и экстрагировали 10 мл воды в ультразвуковой

ванне в течение 3 часов. 2 мл первичного экстракта центрифугировали в течение 10 мин. при 10000g; надосадочную жидкость фильтровали через 0.45 мкм мембранные фильтры. 100 мкл фильтрата разбавляли 1:10 0.1 % раствором муравьиной кислоты.

2.3. Схема эксперимента по ВЭЖХ-МС анализу образцов лекарственных растений

Условия ВЭЖХ-МС/МС определения. В работе использовали ВЭЖХ-МС/МС систему, состоящую из жидкостного хроматографа Shimadzu Nexera X2 и масс-спектрометрического детектора Shimadzu LCMS IT-TOF. Разделение проводили на колонке Hypersil Gold aQ (100 × 2.1 мм), диаметр зерна сорбента 1.9 мкм (Thermo, США). В качестве подвижной фазы использовали смесь растворителей: (А) 0.1% муравьиная кислота в воде и (Б) 0.1% муравьиная кислота в ацетонитриле. Разделение проводили в градиентном режиме по следующей программе: от 0% до 95% В (0 – 12 мин.), 95% В (12 – 17 мин.), от 95% до 0% В за 0.01 мин. и 0% В в течение 3 мин. Скорость потока составляла 0.3 мл/мин. Объем вводимой пробы составлял 10 мкл. Масс-спектрометрическое детектирование проводили с использованием электрораспылительной ионизации, в режиме регистрации положительно и отрицательно заряженных ионов при следующих параметрах: температура переходного капилляра 200°C, напряжение на распыляющем капилляре – (+3500В)-4500 В, расход газа для распыления подвижной фазы в источнике ионов 1.5 Л/мин., расход осушающего газа 15 л/мин. Детектирование проводилось в режиме сканирования в диапазоне значений m/z 100–900. Время регистрации каждой полярности составило 400 мс.

Также использовали В работе использовали ВЭЖХ-МС/МС систему, состоящую из жидкостного хроматографа Agilent 1290 и масс-спектрометрического де-

тектора Agilent 6460 Triple Quadrupole. Разделение проводили на колонке Hypersil Gold aQ (100 × 2.1 мм), диаметр зерна сорбента 1.9 мкм (Thermo, США). В качестве подвижной фазы использовали смесь растворителей: (А) 0.1% муравьиная кислота в воде и (Б) 0.1% муравьиная кислота в ацетонитриле. Разделение проводили в градиентном режиме по следующей программе: от 0% до 95% В (0 – 12 мин.), 95% В (12 – 17 мин.), от 95% до 0% В за 0.01 мин. и 0% В в течение 3 мин. Скорость потока составляла 0.3 мл/мин. Объем вводимой пробы составлял 3 мкл.

2.4. Программные средства, использованные для построения классификационных алгоритмов

Расчёты производились с использованием пакета `scikit-learn` [175]. Все алгоритмы были написаны с использованием программного языка Python в среде программного пакета для научных вычислений Anaconda [176], включающего различные встроенные вычислительные библиотеки.

Все написанные в данной работе алгоритмы могут быть найдены по адресу: <https://github.com/dmitro-nazarenko/chemfin-open> и <https://github.com/kharyuk/chemfin-plasp>

В дополнение к вышеуказанным, программные пакеты [176, 177, 178, 179, 180] были использованы для реализации вычислений.

ГЛАВА 3

Построение классификационных алгоритмов для выборки из 36 классов

Задача видовой идентификации растений в терминах машинного обучения является задачей классификации. В данной группе методов применяют размеченную выборку данных, т.е. каждому объекту в обучающей выборке заранее вручную присвоен класс. Задачей алгоритма в таком случае является настройка параметров математической модели для выделения ключевых признаков классов для последующей успешной идентификации неизвестных объектов. Файл, получаемый после ВЭЖХ-МС анализа экстрактов растительных препаратов обычно содержит в себе данные о большом числе низкомолекулярных вторичных метаболитов (малые молекулы в диапазоне 100 - 900 Да). Химический состав экстрактов, т.е. характеристические соединения и соотношения их содержаний в конкретном виде могут быть использованы для построения алгоритма видовой идентификации растений. Широко изучено, что химический состав растений может варьироваться в широких пределах в зависимости от региона произрастания, типа почвы, климата и других факторов [181]. Тем не менее, в данном эксперименте исходили из предположения, что список вторичных метаболитов, синтезируемых растением почти всегда содержит устойчивую группу «ядровых» соединений в определенных соотношениях. Именно эти соединения внутри ВЭЖХ-МС данных могут служить химической информацией, необходимой для построения эффективного алгоритма видовой идентификации.

Результат химического анализа в значительной степени зависит от выбора способа пробоподготовки. Так как профиль растения используется для последующей видовой идентификации, он должен в некоторой мере отображать совокупный химический состав препарата растения. Химический профиль должен быть достаточно сложен (т.е. иметь достаточно большое число параметров, например) для того, чтобы профилирование позволяло одновременно различать сотни или даже тысячи видов растений. При этом целью данной работы являлось создание подхода не только к подтверждению состава растительного препарата при наличии априорной информации о пробе, но и к быстрой видовой идентификации неизвестных

препаратов растений.

Для разработки такого подхода процедура пробоподготовки должна быть простой и при этом достаточно универсальной, т.е. пригодной для извлечения химических соединений разной природы. В традиционной медицине и при производстве современных биологически активных добавок к пище (БАД, dietary supplements) на основе лекарственных растений наиболее популярны 2 метода: экстракция чистой водой при термообработке и водно-спиртовая экстракция. Для последующего сбора данных в итоге выбрали экстракцию 70% этанолом в воде.

После выбора метода анализа и протокола пробоподготовки выбирали список видов для составления выборки и создания базы данных. В машинном обучении для получения эффективных классификационных алгоритмов крайне важны размер и представительность выборки по каждому классу. Поэтому в работе использовались виды произрастающие на территории Российской Федерации. На рынке лекарственных растений в РФ доступно более 400 различных видов, используемых в рецептах народной медицины или в производстве БАД. Среди них было выбрано 36 по следующим критериям:

- Выборка должна включать различные части растений - корни, семена, цветки и т.д.
- Все доступные виды из семейства *Araliaceae* (5 видов). Данное семейство интересно потому, что включает вид *Panax ginseng* и другие адаптогены, такие как *Eleutherococcus senticosus* и *Oplopanax elatus*.
- Все доступные виды семейства *Umbelliferae* (11 видов). *Umbelliferae* - одно из наиболее широко культивируемых семейств растений.
- Остальные 20 видов были выбраны случайным образом среди общего списка доступных видов (Таблица 2).

Полный список использованных в эксперименте видов представлен в Таблице 2. Для большинства выбранных видов были получены образцы двух последовательных урожаев (2014 + 2015 или 2013 + 2015 года) и выращенные в различных климатических условиях: Алтайский край, Пензенская область, черноморское побережье. В базу данных также добавили результаты анализа экстрактов, подвергнутых

различным условиям хранения. В эту категорию попали 36 образцов экстрактов, высушенные в вакуумном испарителе и выдержанные при 60 °С в течение 5 дней для симуляции старения растительного материала, а также 36 образцов, которые хранили на воздухе при комнатной температуре в течение полугода.

Таблица 2: Виды растений, использованные в эксперименте.

Часть растения	Вид
Корни	<i>Acanthopanax sessiliflorum</i> , <i>Eleutherococcus senticosus</i> , <i>Oplopanax elatus</i> , <i>Panax ginseng</i> , <i>Rhodiola rosea</i> , <i>Inula helenium</i> , <i>Helianthus tuberosus</i> , <i>Archangelica officinalis</i> , <i>Acorus calamus</i> , <i>Rosa majalis</i> , <i>Valeriana officinalis</i> , <i>Sambucus nigra</i> , <i>Glycyrrhiza glabra</i> , <i>Levisticum officinale</i> , <i>Ichorium intybus</i> , <i>Arctium lappa</i> , <i>Potenilla erecta</i> , <i>Dioscorea caucasica</i> , <i>Taraxacum officinale</i> , <i>Hedysarum neglectum</i> , <i>Aralia mandshurica</i> , <i>Astragalus membranaceus</i>
Семена/плоды	<i>Coriandrum sativum</i> , <i>Daucus carota</i> , <i>Petroselinum crispum</i> , <i>Foeniculum vulgare</i> , <i>Anethum graveolens</i> , <i>Pimpinella anisum</i> , <i>Silybum marianum</i> , <i>Linum usitatissimum</i>
Листья/наземная часть	<i>Bupleurum aureum</i> , <i>Pimpinella saxifraga</i> , <i>Heracleum sibiricum</i> , <i>Asarum europaeum</i> , <i>Aegopodium podagraria</i>

Условия масс-спектрометрического детектирования и хроматографического разделения подбирали так, чтобы они были просты и воспроизводимы на как можно более широком диапазоне аналитического оборудования. Для регистрации использовали рассмотренный выше диапазон масс малых молекул – 100-900 Да в режимах регистрации положительных и отрицательных ионов. Так как главная цель данной части работы состояла в разработке методологии идентификации, в данном эксперименте не уделяли большого внимания предварительному подтверждению видовой принадлежности используемых образцов. Однако химические профили экстрактов (времена удерживания и значения m/z наиболее интенсивных пиков) полученные после ВЭЖХ-МС анализа использовали для контроля

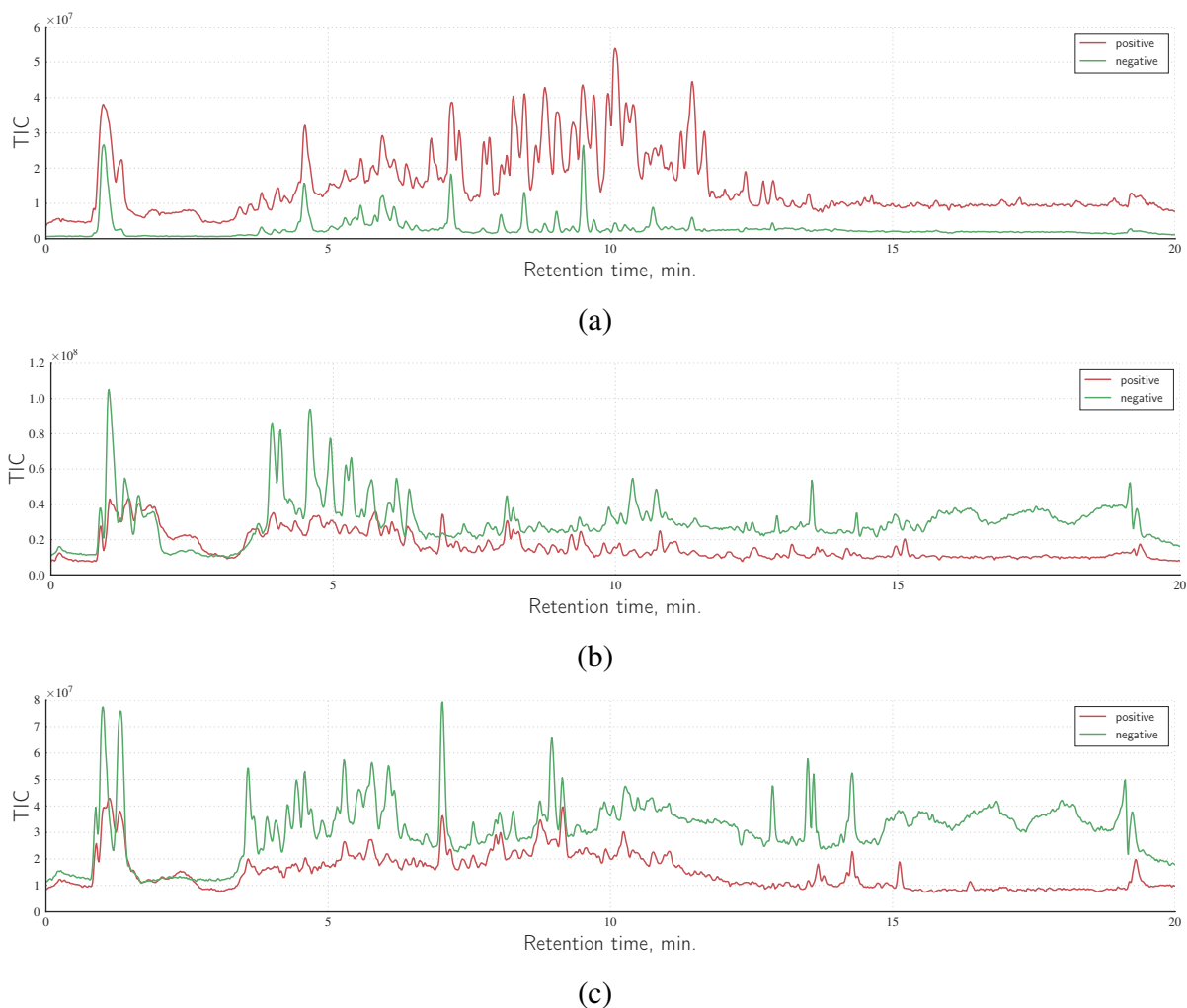


Рис. 13: Хроматограммы растительных экстрактов: *Archangelica officinalis* (корни) (a), *Asarum europaeum* (листья) (b) и *Pimpinella anisum* (семена) (c) в режимах положительной (красный) и отрицательной (зелёный) ионизации. ТИС-полный ионный ток.

правильности видовой разметки образцов от различных производителей за счёт использования матриц попарной корреляции векторов образцов (см.ниже). Примеры типичных хроматограмм экстрактов представлены на Рисунке 13.

3.1. Задача классификации

3.1.1. Предобработка данных

После получения первичных или "сырых"(raw, original) данных проводят их предобработку и выбор признакового пространства (feature selection). После ВЭЖХ-МС эксперимента было получено 870 образцов, проанализированных в одинаковых экспериментальных условиях (ВЭЖХ-МС система Shimadzu IT-TOF). Переноса образцов в контрольных пробах выявлено не было. Файлы хроматограмм были сконвертированы в .mzXML формат. Далее хроматограммы были подвергнуты процедуре автоматического интегрирования хроматографических пиков с использованием программного пакета Waters Progenesis QI. Для каждой хроматограммы была получена таблица, содержащая значения m/z и времен удерживания для каждого обнаруженного пика. Далее значения m/z всех пиков были округлены до целочисленных значений. Для уменьшения размерности данных времена удерживания также были отброшены.

Для каждого значения m/z в положительной и отрицательной полярности были отобраны пики с максимальными значениями площади. Таким образом, для каждой хроматограммы был получен вектор, состоящий из 1600 значений (по 800 для положительной и отрицательной полярностей). Таким образом был получен трехмерный числовой массив $A(i, j, k)$, где $i = \overline{1, m}$, $j = \overline{1, n}$, $k = \{0, 1\}$, - список образцов, значения m/z от 100 до 900 с шагом в 1 при чередовании полярностей.

Для последующих вычислительных экспериментов данные представляли в виде матрицы размера $m \times 2 \cdot n$. Также проводили отбор признаков по следующей

схеме: были оставлены только переменные, столбцы которых в матрице данных имели ненулевые значения арифметического среднего. Из-за этого число признаков (переменных) сократилось с 1600 до 632 на образец.

В литературе есть большое число практических и теоретических пособий по машинному обучению и методам классификации, например [182], [183], [184], и т.д.

В задаче классификации каждому подаваемому на вход алгоритма образцу должен быть приписан какой-либо из нескольких классов или сразу несколько классов (справедливо для SIMCA, например). Для всех образцов в выборке метки классов проставлены вручную. Далее к размеченным данным можно применять различные алгоритмы: так, например, при применении логистической регрессии задача классификации может рассматриваться как поиск апостериорных вероятностей входных данных быть преобразованными в одну из доступных меток классов.

3.1.2. Логистическая регрессия

Логистическая регрессия представляет собой частный случай линейных регрессионных моделей, где зависимые переменные имеют форму бинарных категориальных значений. На зависимые переменные накладывается следующая гипотеза:

$$h(x) = \theta\left(\sum_{i=1}^n \omega_i x_i\right) = \theta(w^T x), \quad (3.1)$$

где $\theta(z) = \frac{1}{1+e^{-z}}$ так называемая "сигмоидная" функция, $w \in \mathbb{R}^n$ - параметры, $x \in \mathbb{R}^n$ - входной образец.

Целью при использовании логистической регрессии является оценить апо-

стериорные вероятности на основе вышеуказанной гипотезы. Параметры w рассчитывают так, чтобы минимизировать кросс-энтропию. Кросс-энтропия широко используется как мера ошибки между предсказанными и реальными значениями переменных.

$$w = \arg \min_w \left[\frac{1}{m} \sum_{k=1}^m \ln(1 + e^{-y_k \cdot w_k^T x_k}) + \lambda \|w\|_2^2 \right], \quad (3.2)$$

(x_k, y_k) - k -й образец с меткой класса, $x_k \in \mathbb{R}^n$, $y_k \in \{-1, +1\}$, $w \in \mathbb{R}^n$ - параметры, которые необходимо оценить. Второй аргумент в сумме есть регуляризация - она штрафует модель за высокие значения w с весом λ .

3.1.3. Метод опорных векторов

В противовес логистической регрессии, метод опорных векторов (МОП) не позволяет естественным образом рассчитывать вероятности классовой принадлежности. В МОП целью ставится оценить параметры w так, чтобы соблюдались следующие условия:

1. $w = [w_0, w_1, \dots, w_n] \in \mathbb{R}^{n+1}$ - вектор с минимальной возможной нормой, $w = \arg \min_{w \in \mathbb{R}^{n+1}} \|w\|_2^2$;
2. w определяет положение гиперплоскости, которая отделяет представителей разных классов

В данном случае мы рассматривали случай классификационной задачи с двумя классами для простоты изложения. Математически данная задача может быть выражена как квадратичная задача оптимизации с наложенными ограничениями:

$$\begin{cases} w = \arg \min_w \in \mathbb{R}^{n+1}, \eta \in \mathbb{R}^n [\frac{1}{2} \|w\|_2^2 + C \sum_{k=1}^m \eta_k] \\ y_k[w_0 + (\hat{w}, x_k)] \geq 1 - \eta_k, \quad \forall k = \overline{1, m} \\ \eta_k \geq 0, \quad k = \overline{1, m} \end{cases}, \quad (3.3)$$

где $w \in \mathbb{R}^{n+1}$ это вектор весов, который необходимо оптимизировать, $C > 0$ регуляризационный параметр, $\eta_k \geq 0$ - ошибка на k -ом образце. Другими словами, МОП естественным образом находит разделяющую гиперплоскость с наибольшим расстоянием (зазором) между классами.

Однако, данный метод необязательно будет эффективно работать на данных, которые линейно неразделимы. В таких случаях используют метод "мягких" границ либо часто применяют так называемые ядровые функции (ядра). Идея ядер состоит в том, чтобы трансформировать исходные переменные в новое пространство за счёт особой функции, которая должна соответствовать теореме Мерсера. Например, функция с радиальным базисом (РБФ) $K(u, v) = \exp(-\gamma \|u - v\|_2^2)$ (Гауссово ядро с параметром γ) широко применяется как мера близости между парами объектов. Таким образом, вместо прямого решения, можно попробовать классифицировать образцы по степени их близости.

3.1.4. Решающие деревья и случайный лес

Решающие деревья это непараметрический метод обучения с учителем, применяемый в классификации. Его суть состоит в поиске простых правил классификации на основе данных за счёт построения иерархического древа разделения признаков пространства.

Однако, решающие деревья большой глубины имеют свойство переобучаться на обучающей выборке, что приводит к плохой обобщающей способности алгоритма. Для предотвращения подобного исхода обычно применяют методы ансамблей.

Одним из них является метод **случайного леса**. Случайный лес (Random Forest, RF) - это ансамблевый метод, который использует несколько простых алгоритмов (решающих деревьев) на различных подмножествах исходной выборки либо на различных подмножествах исходного признакового пространства выборки. Конечное решение о принадлежности образца к какому-либо классу достигается за счёт голосования индивидуальных алгоритмов, что улучшает устойчивость и точность модели. Более подробное описание метода случайного леса может быть найдено в [185].

Во всех вышеописанных моделях обучение проводят путём минимизации функционала ошибки. Функция потерь (ошибки) используемая в алгоритмах, отличается в зависимости от модели:

- кросс-энтропия (логистическая регрессия, решающие деревья);
- кусочно-линейная функция потерь (метод опорных векторов).

Кросс энтропия (логистическая ошибка):

$$J_{\text{CE}} = \frac{1}{m} \sum_{k=1}^m \ln(1 + e^{-y_k \cdot w_k^T x_k}), \quad (3.4)$$

(x_k, y_k) - k -й образец с меткой класса, $x_k \in \mathbb{R}^n$, $y_k \in \{-1, +1\}$, $w \in \mathbb{R}^n$ - оцениваемые параметры.

Кусочно-линейная функция потерь это верхняя граница числа ошибок сделанных классификатором. Она определяется как:

$$J_{\text{hinge}} = \frac{1}{m} \sum_{k=1}^m \max(0, 1 - y_{\text{out}}^k \cdot y_{\text{true}}^k), \quad (3.5)$$

where $y_{\text{out}}^k \in \mathbb{R}$, $y_{\text{true}}^k = \pm 1$ - предсказанная метка класса и реальная метка для k -ого образца.

3.1.5. Показатели эффективности

Главная задача классификационных алгоритмов состоит в выработке набора правил, позволяющих различить к какому из возможных классов принадлежит неизвестный образец. В случае наличия более чем 2 классов также требуется выбрать соответствующую стратегию: "один против всех" где для каждого класса обучается одна модель, должна отличать данный класс от всех остальных, и "каждый против каждого" где необходимо обучение $K(K-1)/2$ моделей для K классов, каждая из которых решает задачу попарного различения классов. Несмотря на то, что стратегия "каждый против каждого" в некоторых случаях может давать более точные предсказания, низкая мощность имеющейся выборки данных нивелирует её возможные преимущества.

Для выравнивания представленности классов в выборке случайным образом отобрали 20 образцов на класс. Для каждой модели расчёты производились отдельно.

В первой части эксперимента проводили поиск оптимальных параметров моделей. Для этого вся выборка была случайным образом поделена на обучающую (70%) и тестовую (30%). Разделение проводили таким образом, чтобы сохранить равные доли каждого класса в выборках. Для каждой из использованных моделей такой процесс повторяли независимо 100 раз. Значения усредненных значений функции потерь были проинспектированы для всех значений настраиваемых параметров моделей, после чего были выбраны те, которые отвечали наименьшей разнице в точности предсказаний на обучающей и тестовой выборках.

Во второй части экспериментов идентичный подход (100 повторений случайного разбиения выборки) был использован для построения кривых зависимости точность предсказаний моделей от размера обучающей и тестовой выборок. Во всех случаях наблюдали значительный зазор между кривыми для обучающей и тестовой выборок. Данный результат означает, что модели имели устойчивую тен-

денцию к переобучению. Данная проблема была частично решена использованием регуляризации, но наиболее корректным способом в данном случае является увеличение размеров выборок и/или изменение процедуры отбора признаков..

Заключительная часть включала расчёт общих показателей эффективности моделей при оптимальных параметрах. В качестве показателей использовали: точность, чувствительность, специфичность, F1 - меру.

В качестве примера можно привести задачу бинарной классификации, где алгоритм принимает решение, принадлежит ли каждый из поданных на вход образцов к классу A или нет. Для такой задачи можно ввести следующие характеристики:

1. True positives (TP) - число образцов принадлежащих классу A и корректно распознанных алгоритмом;
2. True negatives (TN) - число корректно распознанных образцов, которые не принадлежат классу A;
3. False positives (FP) - число образцов не принадлежащих классу A, но распознанных алгоритмом как являющиеся таковыми;
4. False negatives (FN) - число образцов принадлежащих классу A, но не распознанных алгоритмом;

Теперь с использованием введённых характеристик можно дать определение вышеприведённым показателям

$$\text{Специфичность} = \frac{TN}{TN + FP}, \quad \text{Чувствительность} = \frac{TP}{TP + FN}$$

$$\text{Точность} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F_1 = 2 \cdot \frac{\text{Специфичность} \cdot \text{Чувствительность}}{\text{Специфичность} + \text{Чувствительность}}$$

Специфичность определяет способность алгоритма распознавать только образцы класса A как принадлежащие к нему. Чувствительность же описывает спо-

способность алгоритма корректно классифицировать все имеющиеся в выборке образцы класса A. F1- мера является ещё одним показателем точности и определяется как гармоническое среднее между специфичностью и чувствительностью. Данная мера обычно применяется для сравнения относительной эффективности различных подходов.

В случае многоклассовой классификации значения специфичности, чувствительности и F1-меры могут быть определены различными способами. В данной работе использовался подход, когда показатели качества сначала рассчитывались на каждый класс, а затем рассчитывали невзвешенное среднее по всем классам.

3.1.6. Искусственные данные

Для решения проблемы малого размера выборки существует несколько подходов. Наиболее прямой вариант состоит в добавлении новых ВЭЖХ-МС данных в выборку. Из-за ограничений, определяемых условиями экстракции и ВЭЖХ-МС эксперимента, не представлялось возможным использовать данные открытых баз данных.

Другим возможным решением данной проблема является наложение трансформаций, таких как искажения и отражения [186], на имеющиеся данные. Похожий подход может быть применен и в случае ВЭЖХ-МС данных, позволяя сгенерировать искусственные данные.

Для имеющейся выборки данных были введены следующие искажения: площади от 15 до 30% хроматографических пиков в каждом образце были промасштабированы с использованием случайных величин из непрерывного равномерного распределения таким образом, чтобы конечное изменение площади пика находилось в диапазоне от -30% до +30% от оригинальной величины. Размер каждого класса в таком подходе увеличивали в 100 раз.

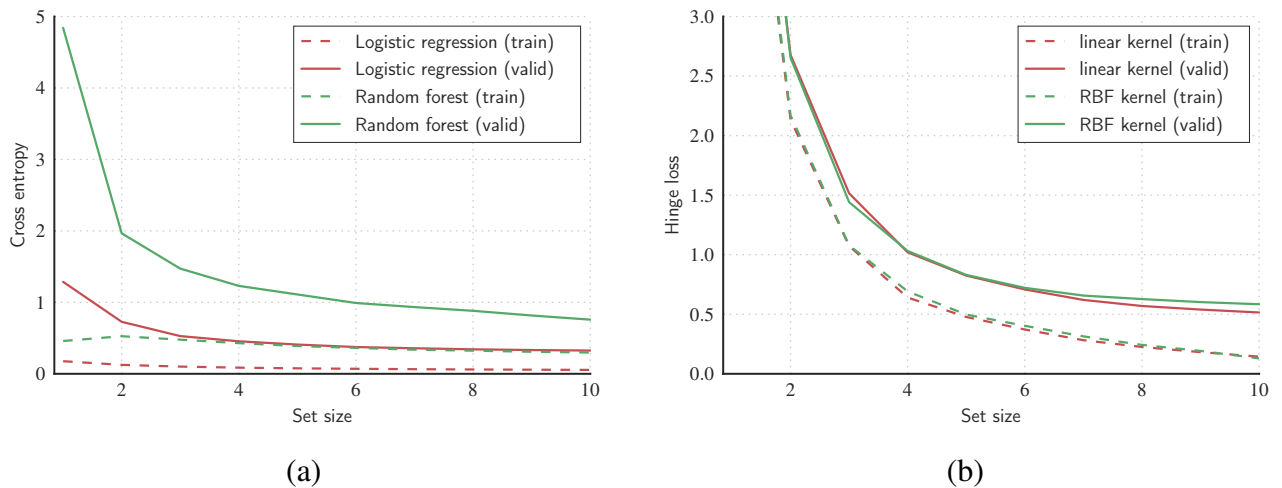


Рис. 14: Кривые обучения для оптимизированных моделей: прерывистые линии для обучающей выборки, сплошные для тестовой: (а) - в единицах кросс-энтропии (красный цвет - логистическая регрессия, зеленый - случайный лес); (б) - кусочно линейная функция потерь (зелёный - МОВ с линейным ядром, красный - МОВ с гауссовым ядром).

3.2. Результаты и обсуждение

Рисунок 14 отражает тенденцию к переобучению для всех применённых алгоритмов: зазор между кривыми обучаемости для сохраняется несмотря на увеличение размеров выборки. Вполне вероятно, что если бы удалось в несколько раз увеличить общий размер выборки данных, это привело бы к сужению зазора.

Для каждого из использованных алгоритмов были протестированы стратегии "каждый против каждого" и "один против всех" (Рисунок 15). Значимой разницы в эффективности обнаружено не было, поэтому более вычислительно затратная стратегия "каждый против каждого" далее не использовалась. Следует отметить, что в процессе оптимизации моделей значения гиперпараметров (коэффициентов регуляризации и т.п.) увеличивали по степенной зависимости без использования подходов строгих оценок.

В качестве примера можно привести оптимизацию для алгоритма случайного

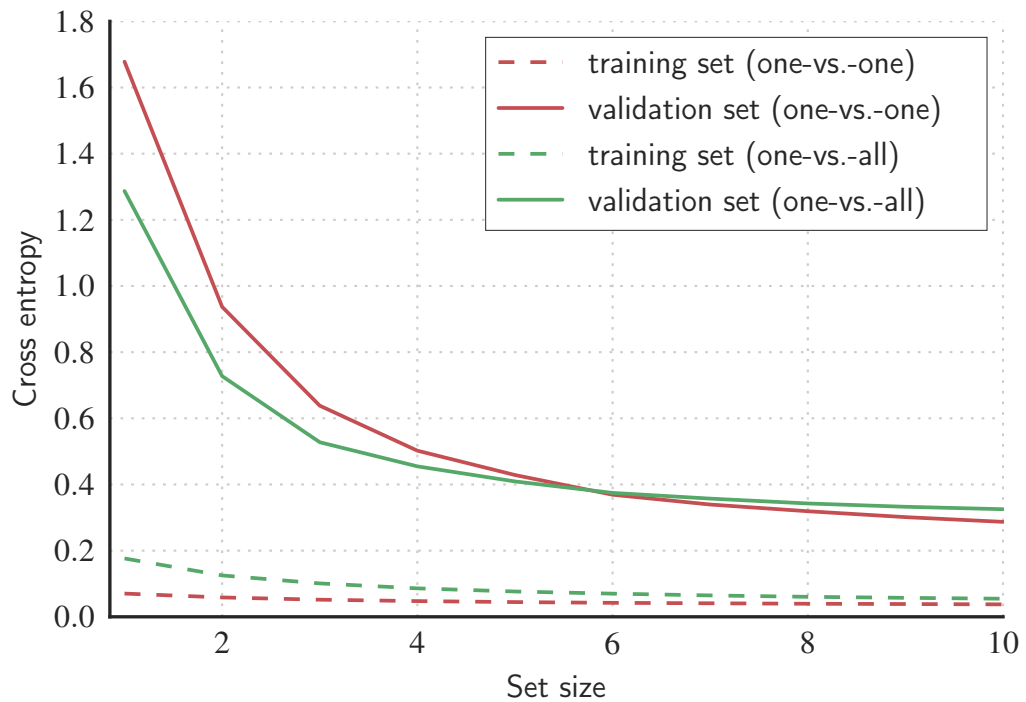


Рис. 15: Сравнение стратегий “один против всех” и “каждый против каждого” для решения классификационной задачи с использованием логистической регрессии.

леса (Рисунок 16). Параметры моделей включали в соответствующие функции потерь для их одновременной минимизации на обучающей и тестовой выборках одновременно. В данном случае оптимальные по итогам перебора параметры были найдены при глубине деревьев от 15 до 20 и числе решающих деревьев не менее 20.

На рисунке 17 представлены кривые обучения для случая использования искусственных данных. Из рисунка можно видеть, что использованный подход для генерации искусственных данных оказался весьма наивным и не привел к сокращению зазора между кривыми обучения обучающей и тестовой выборок. Тем менее, идея использования искусственных данных в задаче идентификации лекарственных растений не может считаться неудачной - требуется более обоснованный и реалистичный механизм внесения возмущений в данные.

Таблица 3 содержит результаты оптимизации моделей всех использованных алгоритмов, выраженные с использованием приведенных выше показателей. Лучшие из использованных подходов показали точность идентификации порядка 95% на тестовой выборке. Однако рассмотренные выше особенности поведения кривых

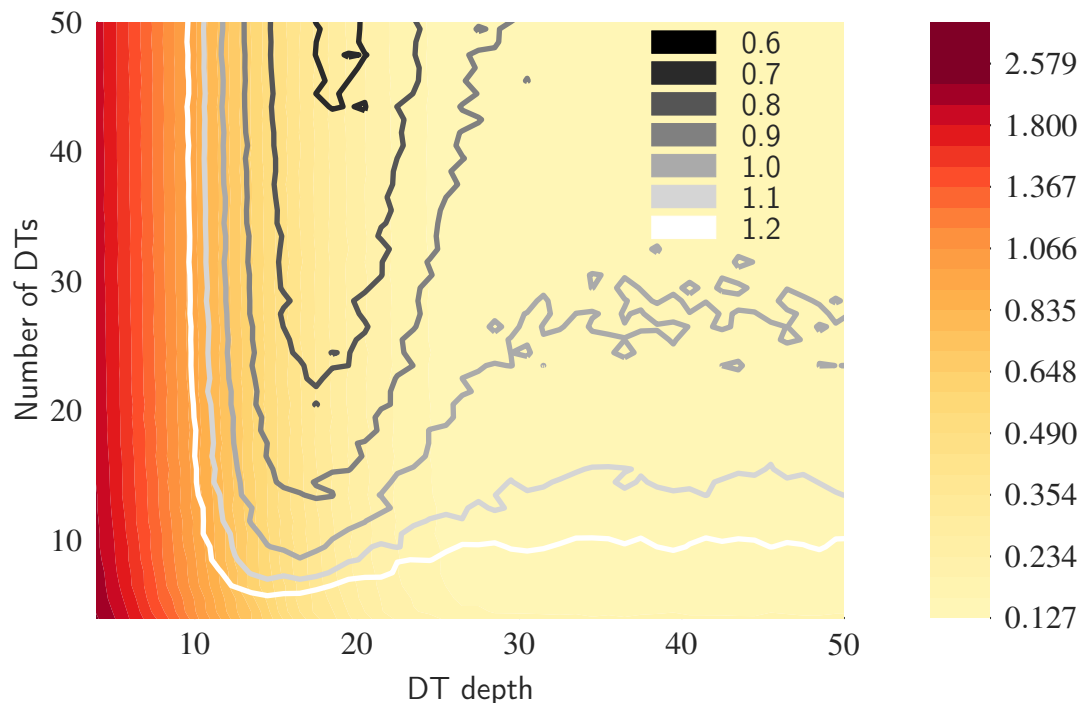


Рис. 16: Значение функции логистической ошибки для случайного леса в зависимости от числа решающих деревьев и глубины каждого дерева. Тепловая карта соответствует ошибке на обучающей выборке, контурные линии отвечают значениям потери на тестовых данных.

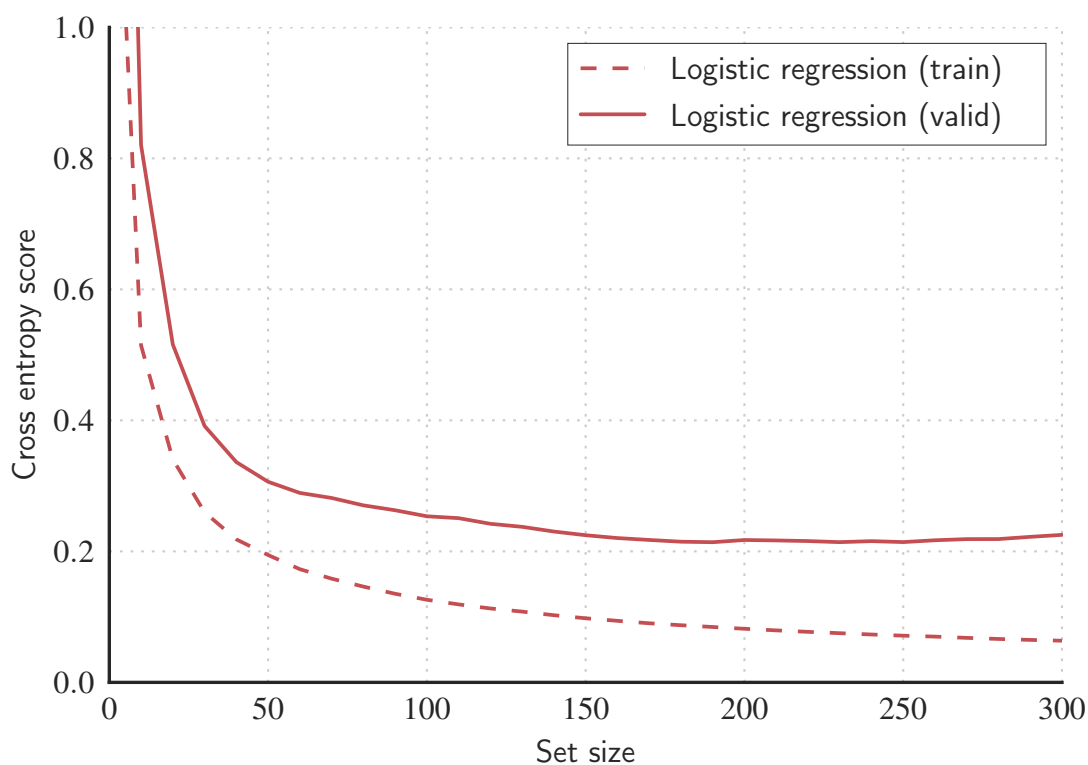


Рис. 17: Кривые обучения в единицах ошибки на кросс-валидации для логистической регрессии на искусственных данных. Прерывистой линией обозначены кривые обучающей выборки, сплошная линия - для тестовой выборки.

обучения демонстрируют проблемы в обобщающей способности алгоритмов.

Таблица 3: Сравнительные характеристики использованных алгоритмов

Метод	Точность, %	Чувст., %	Специф., %	F1-мера, %
Лог. регрессия	94.77	94.56	93.71	93.45
SVM линейный	87.88	87.74	87.05	85.83
SVM гауссово ядро	87.54	87.42	86.66	85.41
Случайный лес	94.36	94.55	94.22	93.53

Использование времен удерживания пиков могло бы позволить значительно увеличить специфичность алгоритмов при распознавании видов, однако ввело бы не менее значимые новые трудности. Значения времен удерживания сильно зависят от типа хроматографической колонки, используемой программы градиентного элюирования, модификаторов подвижной фазы и т.д. Тем не менее, существует вариант категоризировать переменную, отвечающую времени удерживания, и, таким образом, резко сократить её возможный диапазон значений. Учитывая тот факт, что обращённо-фазовые колонки с октадецильными привитыми группами являются наиболее используемыми в данной области, времена удерживания разделяемых соединений сильно коррелируют с их полярностью. Соответственно, могут быть введены категории: сильно полярная, полярная, нейтральная категория, низкополярная и неполярная. В таком варианте значения времен удерживания можно использовать в качестве дополнительной признаковой оси в классификационных задачах.

На сегодняшний день по-прежнему не разработано подходов к видовой идентификации растений на основе ВЭЖХ-МС анализа без использования стандартных соединений. В части работы машинное обучение было использовано для расширения существующей аналитической методологии в приложении к анализу растений и препаратов на их основе. Практически все современные масс-спектрометры могут быть использованы для сбора данных в режиме сканирования в положительной и отрицательной полярностях. Благодаря воспроизводимым спектрам оборудование для ГХ-МС с электронной ионизацией позволяет проводить идентификацию химических соединений на основе сравнения получаемых спектров со спектрами в базах данных. Однако, насколько позволяет судить просмотренный массив научной литературы, алгоритмов идентификации лекар-

ственных растений для ГХ-МС оборудования на сегодняшний день также не разработано.

Логистическая регрессия, метод опорных векторов и случайного леса применили к задаче классификации 36 видов лекарственных растений с высокой точностью (>95%). Тем не менее, остаётся открытым вопрос апробации данного подхода на более широких выборках данных, с увеличением как общего числа видов, так и числа образцов каждого класса в выборке.

В случаях, когда имеющаяся выборка данных относительно мала, алгоритмы машинного обучения имеют тенденцию переобучаться и, таким образом, неэффективно работать на новых поступающих образцах. Данная проблема может быть минимизирована применением регуляризации и сужением используемого признакового пространства. На основании данного рассуждения значения времен удерживания пиков были отброшены, а регуляризация применялась во всех применяемых алгоритмах. Тем не менее, лучшее решение данной проблемы заключается в увеличении доступных выборок данных. В дополнение, более сложно организованная процедура отбора признаков также имеет шансы внести положительный вклад.

Другое практическое решение - генерация искусственных данных. Однако данный подход тоже имеет свои ограничения: для генерирования искусственных данных способных улучшить работу алгоритмов необходимо знание о распределениях концентраций различных классов вторичных метаболитов внутри органов растений для различных видов.

Несмотря на свои очевидные преимущества, у комбинации профилирования и машинного обучения есть свои ограничения и недостатки. Часто метод машинного обучения работает по принципу "черного ящика" т.е. несмотря на эффективное решение той или иной задачи сложно понять какие именно правила и критерии алгоритм использует для принятия решений [187]. Другими словами, тяжело спроецировать значимые (strongly weighted) компоненты алгоритма на свойства исходных объектов. Альтернативой для решения данной проблемы могут служить графические вероятностные модели (ГВМ) [85], такие как байесовские сети (БС), широко применяемые в различных научных и производственных областях [188, 189, 190]. Одним из преимуществ БС является явная визуализация зависи-

мостей и связей переменных модели, что может помочь лучше понять критерии классификации и их связь со свойствами объектов рассмотрения. Более того, БС принадлежат к категории генеративных моделей и могут быть использованы для создания искусственных данных [191].

ГЛАВА 4

Построение классификационных алгоритмов на выборке из 76 классов.**4.1. База данных**

На основе результатов точности алгоритмов, полученных в первой серии экспериментов, стала видна необходимость расширения базы данных и получения образцов с предварительно установленной видовой принадлежностью. Было решено включить в базу виды растений, на которые уже существуют альтернативные стандарты идентификации (в подавляющем большинстве случаев это морфологическая видовая идентификация специалистом ботаником). Таким образом, в данной части работы выбор видов лекарственных растений был основан на Государственной Фармакопее РФ, практически все доступные виды из которой и были использованы (около 50). В дополнение были также использованы все виды из первой части работы. В итоге второй серии ВЭЖХ-МС экспериментов, общий размер сгенерированной базы данных составил более 2200 хроматограмм экстрактов 74 видов растений, составляющих при этом 76 классов, так как для вида *Sambucus nigra* одновременно использовались цветки и корни, а для вида *Betula pendula* - листья и почки. Полный список видовой и морфологической принадлежности использованного материала лекарственных растений представлен в Таблице 4. Видовые имена растений даны в соответствии с проектом классификацией международного проекта The Plant List [192].

Таблица 4: Виды растений, использованные в эксперименте.

Часть растения	Вид
Корни и корневища	<i>Eleutherococcus sessiliflorus</i> (Rupr. & Maxim.) S.Y.Hu, <i>Eleutherococcus senticosus</i> (Rupr. & Maxim.) Maxim., <i>Oplopanax elatus</i> (Nakai) Nakai, <i>Panax ginseng</i> C.A.Mey., <i>Sedum roseum</i> (L.) Scop., <i>Inula helenium</i> L., <i>Helianthus tuberosus</i> L., <i>Angelica archangelica</i> L., <i>Acorus calamus</i> L., <i>Rosa majalis</i> Herrm., <i>Valeriana officinalis</i> L., <i>Sambucus nigra</i> L., <i>Glycyrrhiza glabra</i> L., <i>Levisticum officinale</i> W.D.J.Koch, <i>Cichorium intybus</i> L., <i>Arctium lappa</i> L., <i>Potentilla erecta</i> (L.) Raeusch., <i>Dioscorea caucasica</i> Lipsky, <i>Taraxacum campylodes</i> G.E.Haglund, <i>Hedysarum neglectum</i> Ledeb., <i>Aralia elata</i> var. <i>mandshurica</i> (Rupr. & Maxim.) J.Wen, <i>Astragalus propinquus</i> Schischkin, <i>Bergenia crassifolia</i> (L.) Fritsch, <i>Polemonium caeruleum</i> L.
Семена или плоды	<i>Coriandrum sativum</i> L., <i>Daucus carota</i> L., <i>Petroselinum crispum</i> (Mill.) Fuss, <i>Foeniculum vulgare</i> Mill., <i>Anethum graveolens</i> L., <i>Pimpinella anisum</i> L., <i>Silybum marianum</i> (L.) Gaertn., <i>Linum usitatissimum</i> L., <i>Aronia melanocarpa</i> (Michx.) Elliott, <i>Rhamnus cathartica</i> L., <i>Juniperus communis</i> L., <i>Prunus padus</i> L., <i>Vaccinium myrtillus</i> L.
Листья, цветы или наземная часть растения	<i>Bupleurum aureum</i> Fisch. ex Hoffm., <i>Pimpinella saxifraga</i> L., <i>Heracleum sphondylium</i> subsp. <i>sibiricum</i> (L.) Simonk., <i>Asarum europaeum</i> L., <i>Aegopodium podagraria</i> L., <i>Betula pendula</i> Roth, <i>Sambucus nigra</i> L., <i>Ginkgo biloba</i> L., <i>Melilotus officinalis</i> (L.) Pall., <i>Origanum vulgare</i> L., <i>Fragaria vesca</i> L., <i>Hypericum perforatum</i> L., <i>Viburnum opulus</i> L., <i>Urtica dioica</i> L., <i>Atropa belladonna</i> L., <i>Frangula alnus</i> Mill., <i>Tilia cordata</i> Mill., <i>Tussilago farfara</i> L., <i>Mentha × piperita</i> L., <i>Calendula officinalis</i> L., <i>Tanacetum vulgare</i> L., <i>Plantago major</i> L., <i>Artemisia absinthium</i> L., <i>Leonurus quinquelobatus</i> Gilib., <i>Matricaria chamomilla</i> L., <i>Senna alexandrina</i> Mill., <i>Pinus sylvestris</i> L., <i>Populus balsamifera</i> L., <i>Viola tricolor</i> L., <i>Equisetum arvense</i> L., <i>Thymus serpyllum</i> L., <i>Salvia officinalis</i> L., <i>Aerva lanata</i> (L.) Juss., <i>Echinacea purpurea</i> (L.) Moench, <i>Bidens tripartita</i> L.

18 классов, для которых суммарное число хроматограмм в базе было менее 20, объединили в отдельный "негативный класс" для получения более устойчивых результатов классификации. Такая стратегия используется в различных приложениях машинного обучения [193]. Негативный класс был составлен для стратегии "победитель получает всё" (winner takes all), когда алгоритм выдаёт только одну метку класса, на которую более всего "похож" вектор входных данных. Идея негативного класса в том, чтобы при подаче на вход алгоритма данных анализа нового вида (которого нет в базе данных алгоритма), алгоритм приписывал ему метку негативного класса. Видовая структура полученной выборки данных суммирована на Рисунке 18.

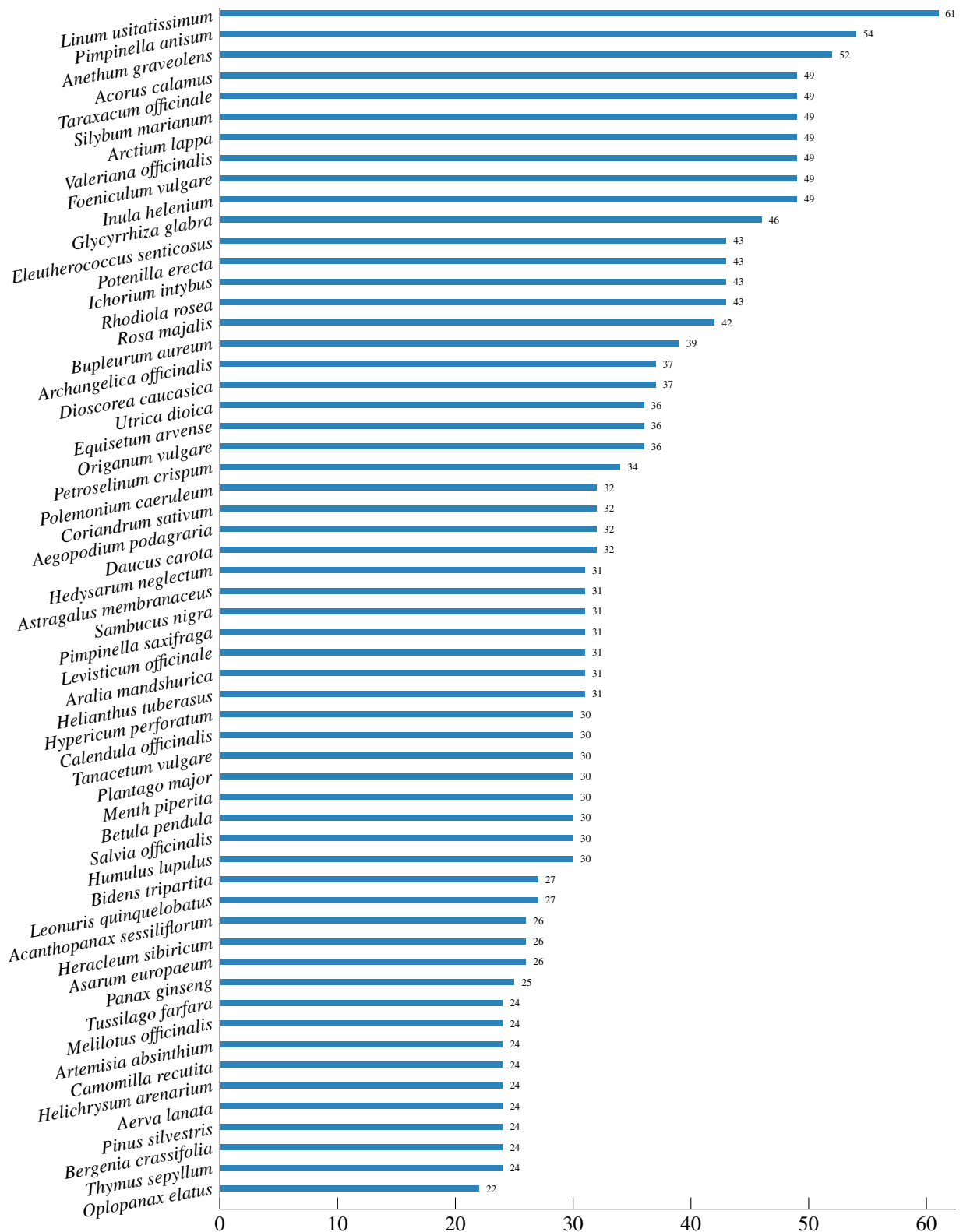


Рис. 18: Повидовая структура базы данных. Отображены только 58 видов с более чем 20-ю хроматограммами.

Так как в расширенной выборке изменилась структура данных (особенно это касается резкого уменьшения числа нулевых столбцов в признаковом про-

странстве), был проведен пересмотр процедуры отбора признаков. Тем не менее, учитывая необходимость сохранить как можно большую применимость подхода в практических условиях два главных критерия - сохранение достаточной информации для однозначной видовой идентификации и устойчивость признаков при проведении анализа на различном оборудовании. Идея фильтрации нулевых столбцов также не представляется устойчивой к изменению видового состава имеющейся для обучения алгоритмов базы данных, т.к. список ненулевых переменных строго коррелирует с химическим составом растительного материала. Ограниченный набор хроматографических пиков индивидуальных соединений с уникальной комбинацией значения m/z и времени удерживания неустойчив к видовому составу выборки по тем же причинам. Более того, значения времен удерживания соединений, хоть они и могут служить критерием идентификации веществ в хроматографии, сильно зависят от большого числа условий - скорости подвижной фазы, температуры, при которой проводится разделение, программы градиентного элюирования и т.д. Таким образом, в качестве отправной точки в данной части работы также был выбран вектор из 1600 переменных, содержащий только значения интенсивности и m/z пиков, обнаруженных на хроматограмме экстракта.

Несмотря на то, что большая часть данных ВЭЖХ-МС анализа лекарственных растений, использованных в данной работе, была получена с использованием масс-спектрометра высокого разрешения, вектор данных состоял только из целочисленных значений m/z . Основная причина такого подхода - стремление к некоторой степени универсальности. Одной из задач данной работы являлось создание как можно более недорого в применении способа видовой идентификации неизвестного растительного материала. Масс-спектрометрическое оборудование низкого разрешения (часто в виде масс-спектрометров с моно- или трёхквadrupольным масс-анализатором) имеют гораздо более низкую цену и более широко распространены в рутинном анализе. Таким образом преобразование данных к низкому разрешению, если даже точнее, то округление значений m/z всех интегрированных пиков до ближайшего целочисленного значения, входило в обязательную ступень предварительной обработки ВЭЖХ-МС данных независимо от использованного оборудования. В процессе описания разработок подходов к предварительной обработке данных читателю может показаться, что использование хроматографического оборудования в таком случае теряет смысловую нагрузку. Многие масс-спектрометры позволяют проводить анализ растворов методом

прямого ввода пробы в источник ионизации - что позволяет получить данные о значениях m/z и интенсивностях пиков компонентов пробы в масс-спектре. Такой тип данных весьма близок к используемым в данной работе признакам - значениям m/z и площадям хроматографических пиков. Безусловно, интенсивность пика соединения в масс-спектре прямого ввода прямо пропорциональна (с некоторыми оговорками) площади хроматографического пика того же соединения при проведении ВЭЖХ-МС анализа. Тем не менее, данные масс-спектрометрии гораздо менее пригодны для использования в качестве источника получения данных:

1. Взаимное подавление ионизации - если в источник электрораспылительной ионизации поступает одновременно несколько соединений, происходит изменение интенсивностей в масс-спектре по сравнению с интенсивностями масс-спектров чистых веществ в тех же условиях [194]. При этом, помимо компонентов анализируемого экстракта, любая дополнительно присутствующая в пробе примесь способна невоспроизводимо подавлять ионизацию компонентов пробы.
2. Использование хроматографии позволяет проводить отсечку компонентов пробы по времени удерживания. Так, проводили фильтрацию всех пиков с временем удерживания менее 1 минуты. Данный параметр можно варьировать при использовании хроматографических систем различной конфигурации. Такая процедура также позволяет увеличить стабильность данных, а значит и эффективность и надежность алгоритмов идентификации.
3. В случае когда идентификационный алгоритм выдаёт в качестве наиболее вероятного результата 3-5 близкородственных вида, данные ВЭЖХ-МС анализа способны помочь провести конечное установление видовой принадлежности.

На Рисунке 19 представлена синтетическая хроматограмма, восстановленная из вектора данных для одного из образцов вида *Anethum graveolens*. Значения интенсивностей всех 1600 переменных (диапазон 100-900 m/z для положительной и отрицательной полярности) были смоделированы как гауссовские пики, площадь которых пропорциональна площади пика на исходной хроматограмме. Значения времен удерживания пиков были взяты из исходных данных. Данная хроматограмма является своего рода отпечатком пальца вида. Одним из главных предпо-

ложений, положенных в основу данной работы являлось то, что данный отпечаток уникален для каждого вида и может использоваться для его однозначной идентификации.

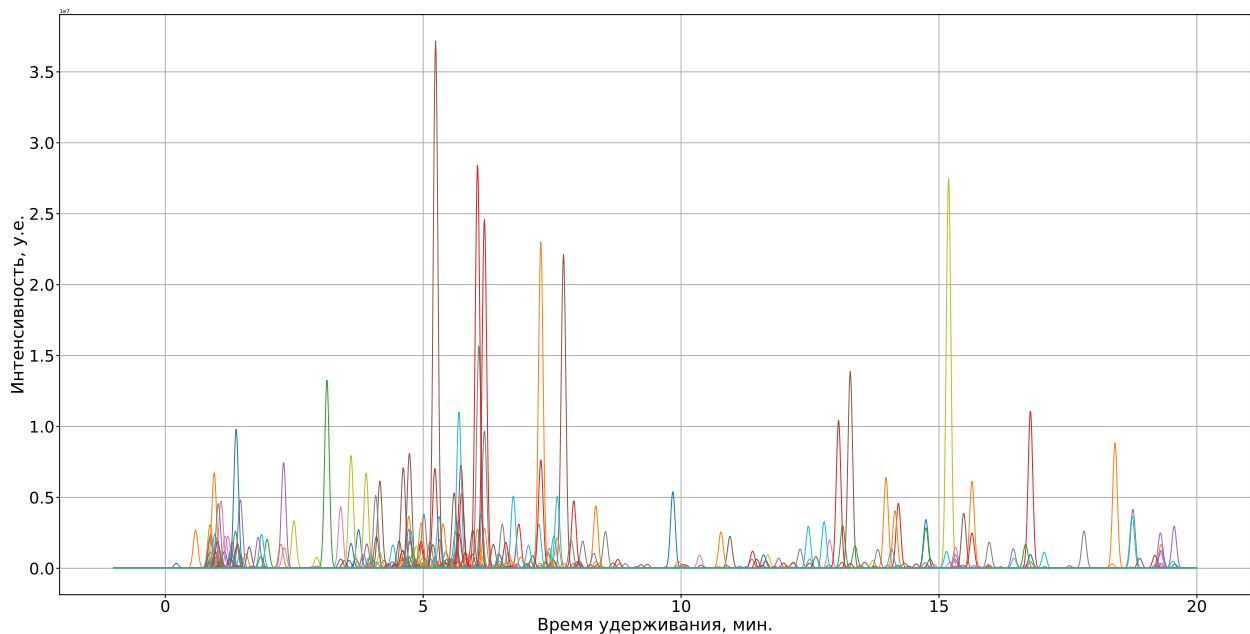


Рис. 19: Искусственная хроматограмма для одного из образцов *Anethum graveolens*, восстановленная на основе одного из векторов выборки. Высоты пиков прямо пропорциональны значениям переменных в векторе. Времена удерживания взяты из первичных данных. Цвета пиков выбраны случайным образом.

4.2. Кросс-валидация и показатели эффективности алгоритмов

Все образцы из выборки были 5 раз случайным образом разделены на 5 равных частей. Каждая из полученных частей (20% выборки) последовательно использовалась как тестовая выборка, тогда как оставшиеся 4 части (80% выборки) использовались как обучающая выборка (всего 25 повторений). Для единообразия индексы разбиения выборки были сохранены и использовались для каждого

использованного алгоритма. Разбиения осуществлялись таким образом, чтобы сохранить постоянную долю образцов каждого класса во всех частях выборки на каждом этапе.

В данной части работы для оценки результатов работы алгоритмов использовали точность и F1-меру. Принимая во внимание несбалансированность размеров классов в выборке использовали взвешенные средние показателей.

4.3. Используемые модели

4.3.1. Байесовские сети

Признаковое пространство может быть рассмотрено как случайный вектор. В таком случае, каждый образец можно считать реализацией данного случайного вектора. Для такого вектора возможно рассчитать совместную функцию распределения его компонент. Однако, большая размерность вектора в данной задаче не позволяет провести прямую оценку данной функции. Наложение гипотез на случайные величины - компоненты вектора может сделать задачу выполнимой. Наиболее простой вариант - наложение условия статистической независимости, позволяющее факторизовать совместную функцию распределения: $p(x_1, \dots, x_m) = p(x_1) \cdot \dots \cdot p(x_m)$. В предположении, что $p(x_i)$ совместная функция распределения существуют. Если к случайному вектору $x = (x_1, \dots, x_m)$ добавить одну категориальную переменную y , они становятся условно-независимыми: $p(x_1, \dots, x_m|y) = p(x_1|y) \cdot \dots \cdot p(x_m|y)$.

Получаемая в таком случае вероятностная модель известна как наивный байесовский классификатор (НБК) [85] и может применяться для решения классификационных задач. Тем не менее, предположение о статистической независимости всех компонент вектора x слишком строгое, более естественным было бы предположить наличие взаимных статистических зависимостей между некоторыми из компонент. Любая совместная функция распределения может быть разложена в произведение 4.1 путем многократного применения правила произведения вероятностей [85]:

$$p(x_1, \dots, x_m) = \prod_{i=1}^m p(x_i | pa[x_i]), \quad (4.1)$$

где $pa[x_i]$ - обозначение набора родительских переменных для i ой переменной. Данное разложение можно легко визуализировать графом, где каждая переменная представлена вершиной, а условие зависимости - ребром графа. Такое представление совместных распределений в форме графов известно как графовые вероятностные модели (ГВМ). Байесовским сетям соответствуют направленные ациклические графы (НАГ), структура которых отражает условные зависимости между случайными величинами. Если все случайные величины сети имеют дискретные функции распределения, такая сеть называется дискретной байесовской сетью (ДБС). Если же в сети присутствуют как дискретные, так и непрерывные случайные величины, то такая сеть называется гибридной байесовской сетью. Если одна из переменных является категориальной и соответствует меткам классов, то её значения могут быть предсказаны при наличии обученной БС. Однако, для обучения необходимо на первом этапе установить структуру самой сети.

Существуют различные подходы к установлению структуры графа из данных [195]; большинство из них требуют высоких затрат вычислительных мощностей. Ввиду большой размерности имеющихся векторов, использовали метод Чо-Лиу [196] для оценки суб-оптимальной структуры сети. Все расчёты моделей, основанные на дискретных БС осуществлялись на основе использования библиотеки Pomegranate [197] для языка Python.

Другие подходы с применением БС включали использование данных с резко пониженной размерностью, преобразованных автоэнкодером. Scikit-learn [175] использовали для реализации НБК, а пакет bnlearn [195] языка R использовали для

гибридных Байесовских сетей. Библиотеку Rpy2 применяли как интерфейс для использования библиотек R в среде Python. Визуализацию сетей осуществляли использованием библиотеки NetworkX [198].

Одним из ограничений применений Байесовских сетей к задаче классификации являются их высокие требования на вычислительную мощность процессоров. Байесовские сети более чем со 100 переменными, имеющими нормальное распределение (распределение Гаусса) уже считаются мало применимыми, ввиду необходимости проводить вычисления на суперкомпьютерах. Исходя из этого было принято решение превратить имеющиеся 1600 переменных из непрерывных числовых в категориальные.

Входные данные были подвергнуты предварительной обработке - только 30% самых интенсивных значений площадей хроматографических пиков было оставлено в каждом векторе данных, затем все оставшиеся ненулевые значения были заменены на единицы. Таким образом, преобразованные векторы данных включали только данные о значениях m/z 30% самых интенсивных пиков. Фактически, такая бинарная (состоящая из нулей и единиц) форма данных отражает только список значений m/z самых интенсивных пиков, но не их площади. Такой качественный подход к рассмотрению масс-спектрометрической информации позволил достичь 90% точности на выборке Тест 1.

Безусловно, возможно дальнейшее усовершенствование данного алгоритма, заключающееся например в более сложной процедуре преобразования переменных. Так, вместо обнуления всего, кроме 30% самых интенсивных пиков возможно приравнять те же 30% самых больших пиков 2, а остальные ненулевые пики приравнять 1. Таким образом, диапазон значений, принимаемых переменными составил бы 0, 1 и 2. Однако даже на при имеющейся упрощенной конфигурации метод требует очень больших вычислительных мощностей и процессорного времени.

Тем не менее, можно уверенно заключить, что 1600 переменных это слишком много, как с точки зрения затрат вычислительных мощностей, так и отсутствия желаемой эффективности алгоритма идентификации. Было решено применить к массиву данных процедуру понижения размерности для ускорения расчётов. Для данной цели был выбран автоэнкодер.

4.3.2. Автоэнкодер

Автоэнкодер это разновидность искусственных нейронных сетей, которая предназначена для обучения прямой и обратной трансформации данных. Выходной слой автоэнкодера должен как можно более точно соответствовать входным данным. Такая нейронная сеть может быть использована для адаптивного извлечения признаков из данных большой размерности [43].

В данной работе применяли нейронную сеть с прямым распространением ошибки с $N = 2n$ слоями, где структура последних n слоев отражала размеры и форму первичных n слоёв. Нелинейные преобразования были выбраны в соответствии с условием о неотрицательности ВЭЖХ-МС данных (ReLU, rectified linear unit, для последнего слоя и сигмоидную функцию для остальных).

$$\hat{x} = f_N(W_N[\dots f_2(W_2[f_1(W_1[x])]) \dots]); \quad N = 2n, \quad f_i(t) = \begin{cases} \frac{1}{1+e^{-t}}, & \text{if } i = \overline{1, N-1} \\ \max(0, t), & \text{if } i = N \end{cases} \quad (4.2)$$

Параметры нейронной сети оценивали через оптимизацию функционала специального вида. Благодаря своей большей устойчивости к выбросам по сравнению с квадратичной функцией, была выбрана сглаженная l1 функция потерь 4.3 (известная также как функция потерь Хубера, Huber loss). Данный функционал оптимизировали методом Адама [199]. Все расчёты осуществляли с использованием пакета pyTorch [200] для языка Python.

$$l(x, \hat{x}) = \sum_i z_i, \quad z_i = \begin{cases} 1/2(x_i - \hat{x}_i)^2, & \text{if } |x_i - \hat{x}_i| < 1 \\ |x_i - \hat{x}_i| - 1/2, & \text{otherwise} \end{cases} \quad (4.3)$$

Для предотвращения переобучения сети и увеличения её обобщающей способности, число переменных на каждом следующем слое кодирующего участка

было в 4 раза меньше. Наилучшие результаты были достигнуты при использовании следующей процедуры предобучения: сначала обучают модель с $N_1 = 2$, затем все обученные слои используют в модели с $N_2 = 4$ как параметры инициализации. Подобным образом доходят до слоя k с $N_k = 2k = N$. Было обнаружено, что 3 кодирующих слоя достаточно для извлечения признаков, которые могут быть использованы в различных моделях для классификации - логистической регрессии, НБК и гибридной байесовской сети. Дополнительно также проверяли насколько малый размер может иметь последний кодирующий слой без значимых потерь в эффективности алгоритмов. Проверку проводили на логистической регрессии с l_1 регуляризацией. Другие классификаторы были построены на основе вышеописанных программных библиотек.

Закодированные векторы длиной в 25 переменных использовали для обучения моделей логистической регрессии и байесовских сетей с непрерывно распределенными случайными величинами. Наивный байесовский классификатор и гибридная байесовская сеть показали точность классификации на Тесте 1 85% и 87% соответственно, тогда как для логистической регрессии была получена точность 96%. В случае Теста 2 все вышеописанные модели показали точность 68-77%.

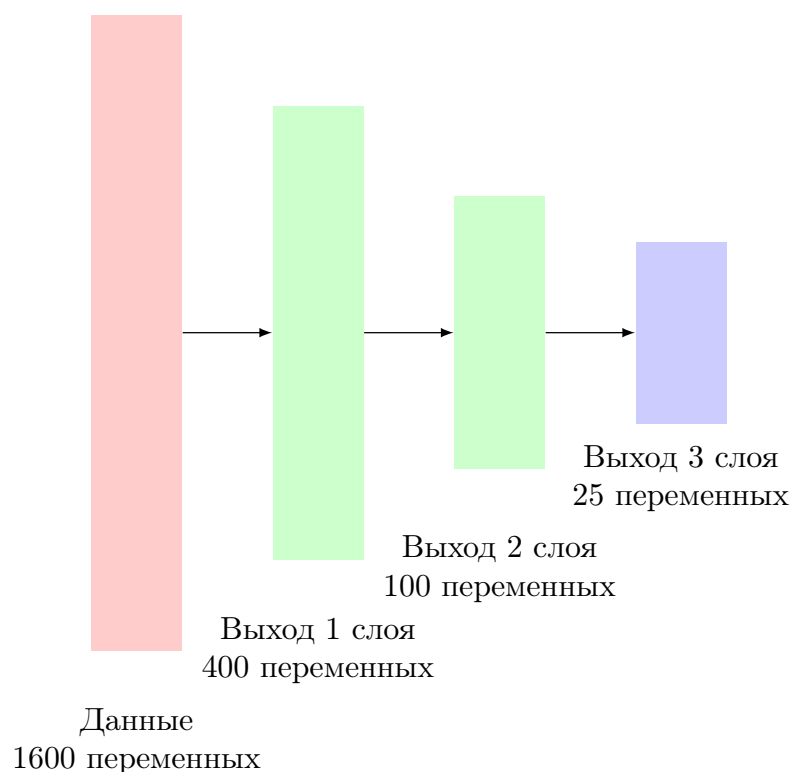


Рис. 21: Финальная структура кодирующей части автоэнкодера.

В качестве альтернативного подхода можно также рассмотреть разделение

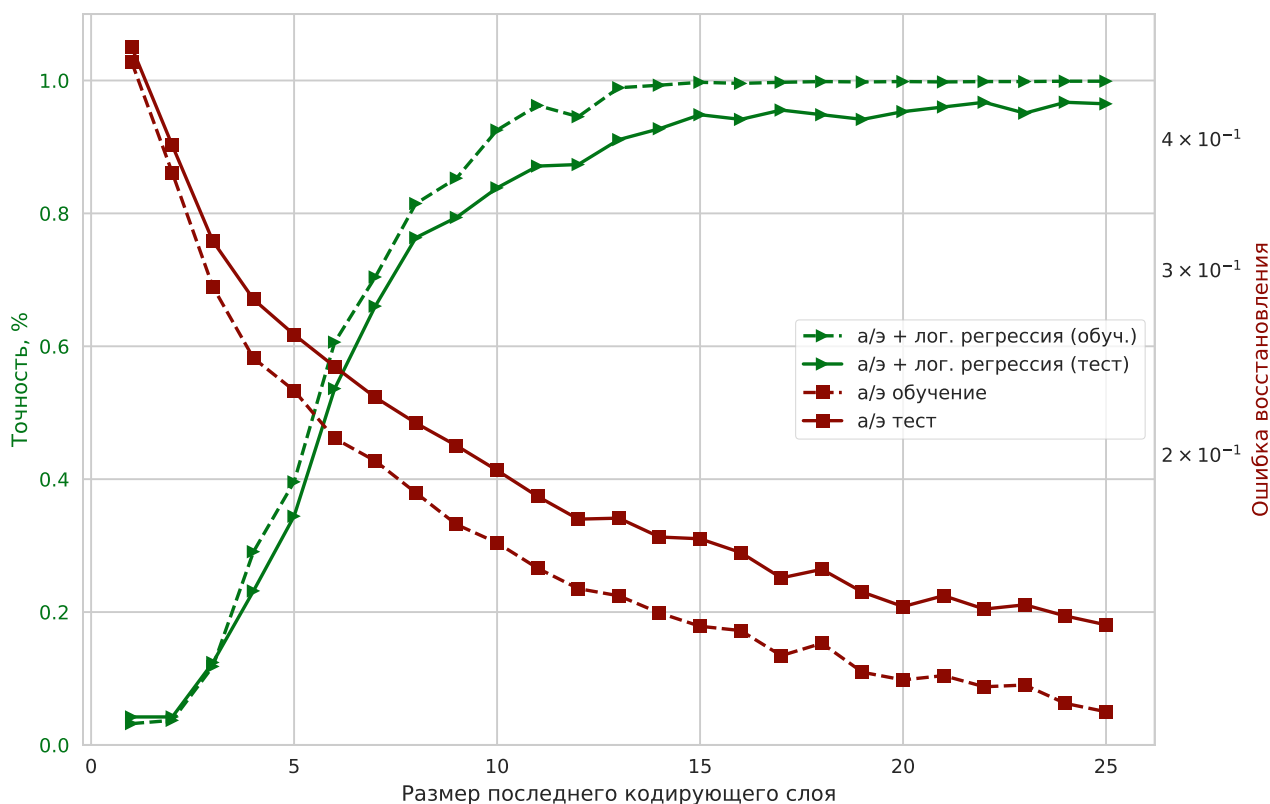


Рис. 20: Точность классификации логистической регрессии на обучающей/тестовой выборках (зеленые линии) и ошибка восстановления данных декодером (красные линии) в зависимости от размера последнего слоя кодирующей части автоэнкодера.

измерений данных разложением Таккера, которое позволяет выявить в них мультилинейные зависимости.

4.3.3. Разложение Таккера с условиями неотрицательности и разреженности

Исходные ВЭЖХ-МС данные могут быть рассмотрены как неотрицательная функция интенсивности, зависящая от 3 параметров: отношения массы к заряду m/z , полярности и времени удерживания. Первичная обработка данных (поиск и

интегрирование пиков) не изменяет координатную сетку, однако делает данные более разреженными. На последующих этапах предобработки значения времен удерживания отбрасываются и данные становятся функцией 2 переменных. Объединение данных всех образцов даёт ещё одну ось. После квантования пространства m/z , данные представляют собой 3-мерный тензор, $T \in \mathbb{R}^{N_{\text{sample}} \times N_{m/z} \times N_{\text{polarity}}}$.

Для подобных данных для выявления их внутренней структуры возможно применение малоранговых тензорных приближений.

В разложении Таккера, данные параметризуются фактор-матрицами A , B , C и тензорным ядром G новой формы:

$$T_{ijk} \approx \sum_{\alpha=1}^{r_1} \sum_{\beta=1}^{r_2} \sum_{\gamma=3}^{r_3} g_{\alpha\beta\gamma} a_{i\alpha} b_{j\beta} c_{k\gamma};$$

$$T \approx |[G; A, B, C]|, G \in \mathbb{R}^{r_1 \times r_2 \times r_3}, A \in \mathbb{R}^{N_{\text{sample}} \times r_1}, B \in \mathbb{R}^{N_{m/z} \times r_2}, C \in \mathbb{R}^{N_{\text{polarity}} \times r_3}$$
(4.4)

В уравнении 4.4 даны определения разложения 3-мерного массива. Гиперпараметры (r_1, r_2, r_3) известные как ранги Таккера влияют на размер фактор-матриц и тензорного ядра. В общем случае разложение Таккера неотрицательно [201], однако было показано, что в случае достаточно разреженного и неотрицательного массива разложение стремится к уникальности [202]. Так как получаемый после предобработки массив данных соответствует вышеописанным условиям, разреженное неотрицательное разложение Таккера (РНРТ) было применено по аналогии с [203].

Оцененные фактор-матрицы далее использовали для классификации. Так как любой новый образец представляет собой матрицу размера $N_{m/z} \times N_{\text{polarity}}$, фактор-матрица для образцов была отброшена, что эквивалентно приравниванию её к единичной матрице I . Получаемая в таком варианте интерпретации оптимизационная задача, которую нужно решить для оценки параметров разложения выглядит следующим образом:

$$\min_{G, B, C} ||[G; I, B, C] - T||_F^2 + \lambda_G \|\text{vec}(G)\|_1 + \lambda_B \|\text{vec}(B)\|_1 + \lambda_C \|\text{vec}(C)\|_1$$

$$\text{s.t. } G \in \mathbb{R}_{\geq 0}^{N_{\text{sample}} \times r_2 \times r_3}, \quad B \in \mathbb{R}_{\geq 0}^{N_{m/z} \times r_2}, \quad C \in \mathbb{R}_{\geq 0}^{N_{\text{polarity}} \times r_3},$$
(4.5)

где $\lambda_G, \lambda_B, \lambda_C$ это штрафы за недостаточную разреженность, $\text{vec}(\cdot)$ векторизация входных данных, $\mathbb{R}_{\geq 0}$ - неотрицательное пространство действительных чисел.

Для разложения Таккера был выбран ранг оси полярностей равный двум, а ранг оси m/z выбирался одинаковым для всех классов. Выбор конкретного значения ранга оси m/z осуществляли сеточным поиском (grid search).

Процедура классификации была организована следующим образом: (1) Расчёт фактор матриц пространств m/z и полярности для каждого класса по отдельности; (2) Умножение каждого образца из тестовой выборки на обратную фактор-матрицу полярности; (3) Расчёт вектора, содержащего значения дистанций между преобразованным вектором образца и фактор-матрицами каждого класса; (4) Выбор класса, расстояние до которого имеет наименьшую величину как предсказание классификатора.

Для расчёта расстояний использовали две метрики - главный угол [204] (подход, близкий к использованному в [205]) и корреляцию дистанций [206]. Корреляцию дистанций рассчитывали с использованием библиотеки `docr` для языка Python.

Фактор матрицы двух осей (m/z и полярность) использовали для классификации. В случае использования РНРТ важно отметить два важных параметра: выбор ранга и выбор метрики расстояния между пространствами столбцов фактор-матриц (линейными оболочками натянутыми на столбцы матриц) (Рисунок 22). Сравнение метрик показало, что метрика главного угла оказалась гораздо более эффективной чем другие варианты на высоких рангах разложения Таккера, таким образом дальнейшие вычислительные эксперименты проводились только с использованием этой метрики. Однако, более высокие ранги разложения в первую очередь означают более длительные вычисления (Рисунок 23).

Одно из главных достоинств разложения Таккера в приложении к задаче классификации заключается в том, что добавление новых классов в обучающую выборку не требует заново рассчитывать уже вычисленные фактор-матрицы, только разложения для массивов добавляемых классов.

Основываясь на достигаемой при ранге разложения Таккера 25 для оси m/z высокой точности классификации и достаточно узким зазором между кривыми

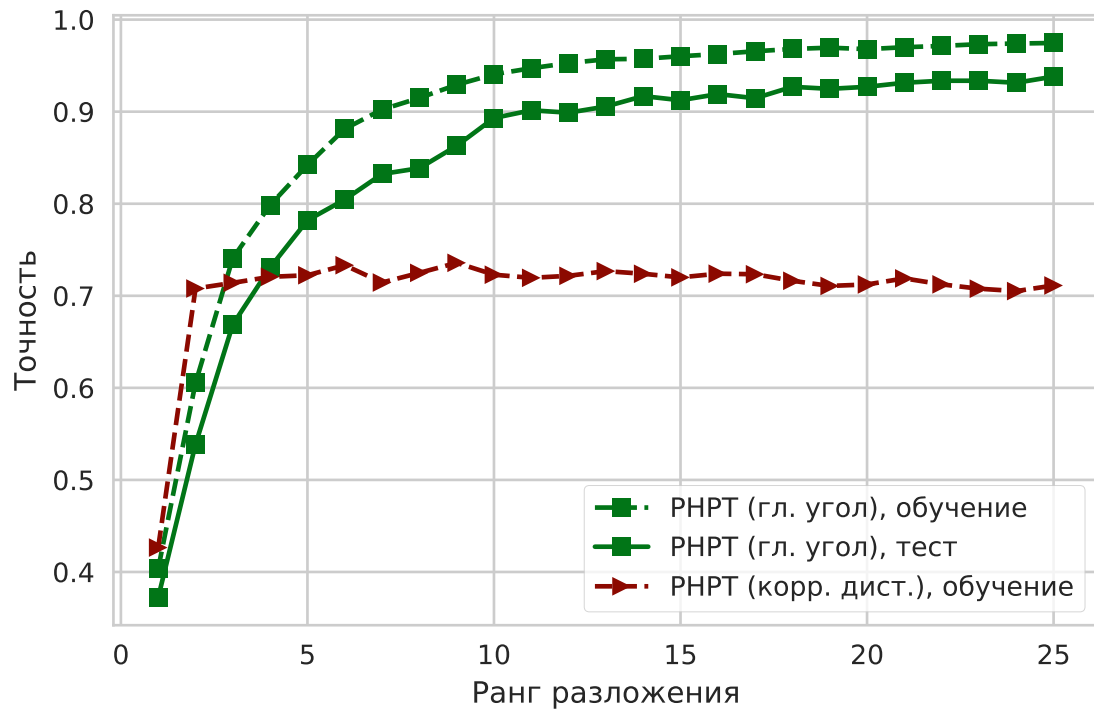


Рис. 22: Выбор ранга для разложения Таккера и сравнение двух метрик, главного угла и корреляции расстояний как основы для классификации.

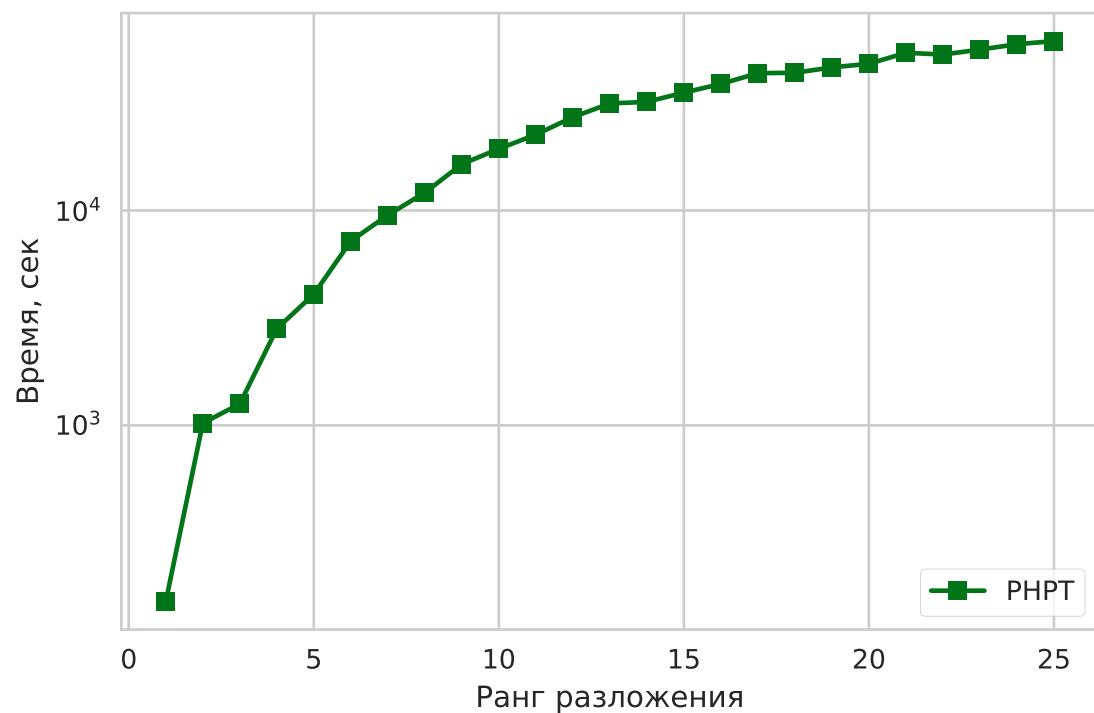


Рис. 23: Время, необходимое на расчёт различных рангов разложения Таккера. Даны медианные значения, посчитанные в кросс-валидации.

обучения на обучающей и тестовой выборках привели к тому, что последующая сравнительная кросс-валидация рассчитывалась на разложении при данном ранге.

Согласно таблице 5, метод РНРТ совместно с метрикой главного угла позволили достичь точности 93% на Тесте 1 и 86% на Тесте 2. Несмотря на то, что зазор между эффективностью на обучающей и Тест1 выборках у РНРТ выше, чем у логистической регрессии на кодированных данных, РНРТ показал наилучшие результаты по точности классификации на независимой выборке Тест2.

Разложение Таккера основано на 3-мерном представлении данных. Однако для любого тензорно массива возможно сделать развертку вида

$$\text{unfold}_k(X) : X^{n_1 \times \dots \times n_d} \rightarrow X^{n_k \times \prod_{l \neq k} n_l}$$

4.3.4. Матричное разложение с условиями неотрицательности и разреженности

В задаче матричного разложения требуется найти две такие матрицы S и M , чтобы их произведение приближало исходную матрицу X как можно более точно, т.е. $X \approx SM^T$. Эта задача связана с поиском базиса в линейном пространстве. Столбцы матрицы M содержат в себе информацию об образцах каждого из классов в выборке. Аналогично с разложением Таккера, в качестве метрики используется расстояние между входным вектором данных и линейными оболочками компонент, извлеченных разложением.

Одно из базовых предположений заключается в том, что разложение содержит малое число параметров, т.е. является низкоранговым. Ещё одно предположение касается свойств параметров. Так же как и в разложении Таккера, на матрицы S и M можно наложить ограничения неотрицательности и разреженности, что в итоге даёт разреженное неотрицательное матричное разложение (РНМР)

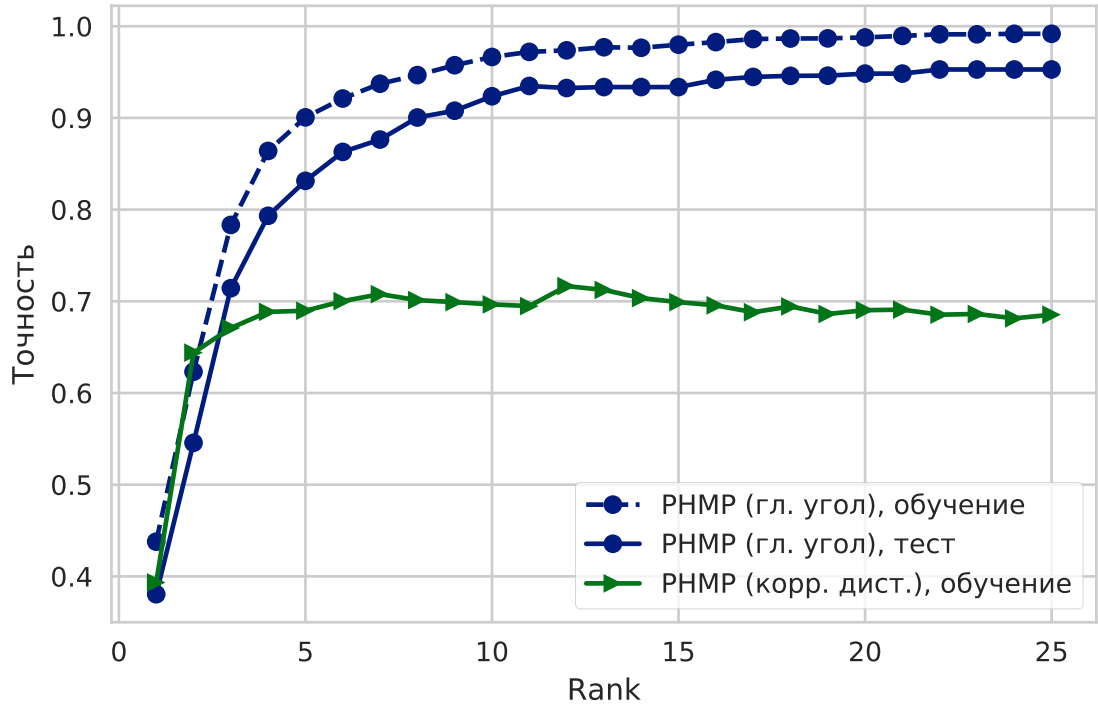


Рис. 24: Выбор ранга для матричного разложения и сравнение двух метрик, главного угла и корреляции расстояний как основы для классификации. Даны медианные значения, посчитанные в кросс-валидации.

Можно отметить, что рассмотренный выше подход РНРТ может рассматриваться как особый случай РНМР, где были разделены моды полярности и m/z .

$$T_{(1)} \approx \underbrace{G_{(1)}}_S \underbrace{(C \otimes B)^T}_{M^T}, \quad T_{(1)} \in \mathbb{R}^{N_{\text{sample}} \times N_{m/z} \cdot N_{\text{polarity}}}, \quad G_{(1)} \in \mathbb{R}^{N_{\text{sample}} \times \hat{r}}, \quad \hat{r} = r_2 \cdot r_3 \quad (4.6)$$

где символом $\otimes$ обозначено Кронекерово произведение между матрицами, а $T_{(1)}, G_{(1)}$ - развёртки по моде номера образцов тензоров T и G . Ранг матричного разложения выбирали на основе точности моделей в 5-кратной кросс-валидации (Рисунок 24).

Матричное разложение показало значительно более быстрые времена оценки фактор матриц (Рисунок 25) по сравнению с разложением Таккера (Рисунок 23).

Таким образом, матричное разложение с теми же наложенными ограничениями, что и РНРТ также применили для задачи идентификации видовой принадлежности растений.

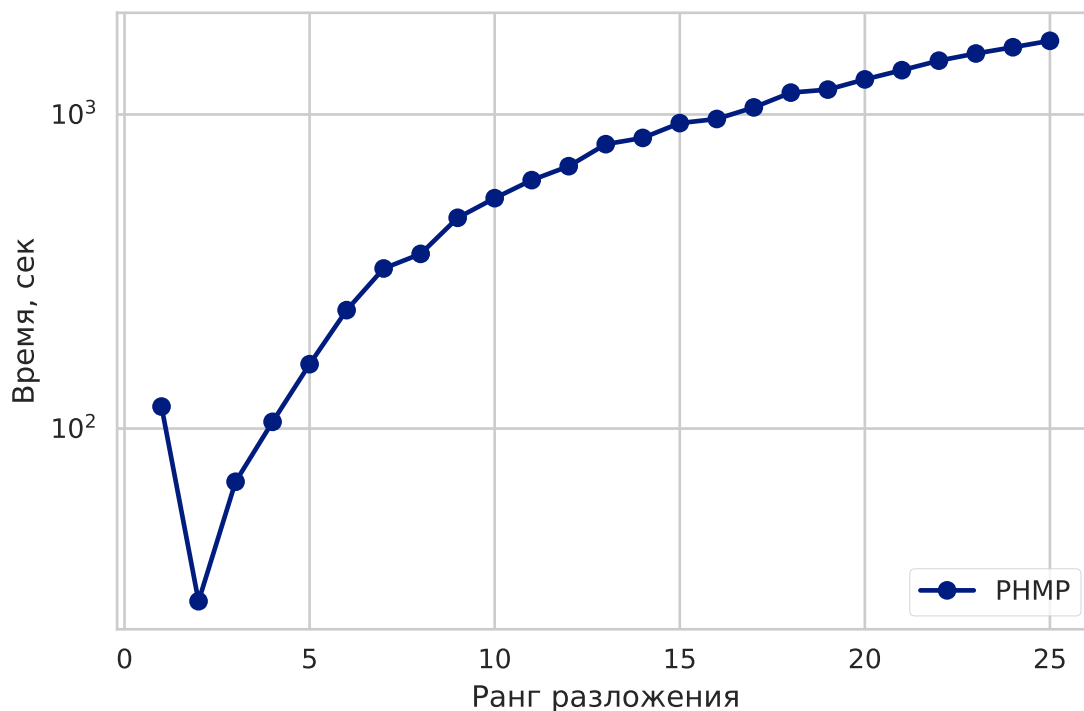


Рис. 25: Время, необходимое на расчёт различных рангов матричного разложения. Даны медианные значения, посчитанные в кросс-валидации.

4.4. Результаты и обсуждение

Общая схема вычислительных экспериментов данной части работы представлена на Рисунке 26, а показатели эффективности отображены в таблице 5

Для расчёта показателей эффективности моделей были сформированы тестовые выборки Тест1 и Тест2. Выборка Тест1 была сформирована на основе стратифицированного случайного разделения выборки данных из 2263 хроматограмм, полученных на масс-спектрометре высокого разрешения со сферической ионной ловушкой в идентичных экспериментальных условиях (пробоподготовка и ВЭЖХ-МС анализ). Разделение проводили в широко применяемом в машинном обучении соотношении 70:30 для обучающей и тестовой выборок. Выборка Тест2 была сформирована из 44 образцов, полученных при альтернативных условиях пробоподготовки (водная и водно-метанольная способы экстракции), анализ

которых проводили как на исходном приборе высокого разрешения, так и на трехквадрупольном масс-спектрометре низкого разрешения. Выборка Тест2 была сформирована для оценки устойчивости получаемых алгоритмов к изменению условий анализа.

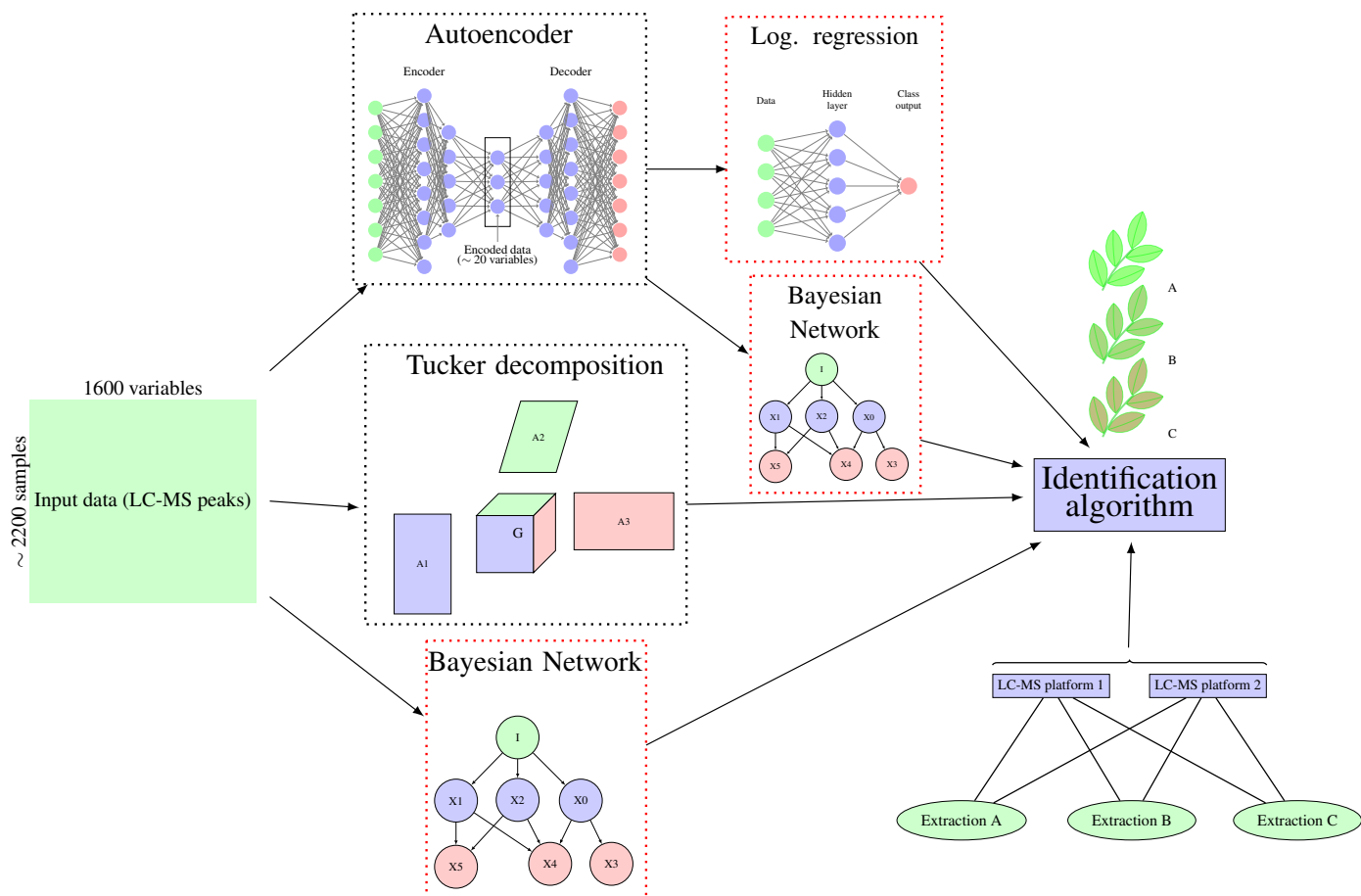


Рис. 26: Схематическое представление всех расчетных экспериментов части 2. Выборка данных была использована 3 различными способами: резкое понижение размерности за счёт автоэнкодера с дальнейшим применением логистической регрессии или Байесовских классификаторов, разложение Таккера с наложенными ограничениями, и прямое применение дискретной БС на первичные данные. Все классификационные алгоритмы были кросс-валидированы.

Таблица 5: Сравнительные характеристики использованных подходов. В таблице представлены медианные значения, рассчитанные в процессе кросс-валидации. Выборка Тест 2 независима от обучающей и Тест 1. Результаты для стратегии “победитель получает всё”

Метод	Accuracy, %			F1-мера, %		
	Обучающая выборка	Тест 1	Тест 2	Обучающая выборка	Тест 1	Тест 2
Лог. регрессия (а/э)	99.8	85.0	72.7	99.8	83.2	79.6
НБК (а/э)	90.3	68.4	75.0	90.6	66.7	79.8
Гибридная БС (а/э)	92.1	69.2	68.2	92.3	67.5	72.4
Дискретная БС	-	78.5	72.7	-	76.5	81.6
РНРТ (гл. угол)	98.0	77.9	84.1	98.0	76.3	89.8
РНМР (гл. угол)	99.2	75.4	81.8	99.2	74.4	85.2

В процессе разработки алгоритмов также важно учитывать временные затраты на обучение моделей и предсказание меток классов неизвестных образцов. Особенно значим параметр времени, когда алгоритм работает в режиме онлайн-сервера и дообучается с получением новых данных (запросов). В Таблице 6 показаны времена обучения всех моделей и времена предсказания в расчёте на один образец. Наиболее долго расчёт модели происходит для автоэнкодера, однако благодаря устойчивости структуры данных большой базы необходимости часто проводить дообучение автоэнкодера нет. Более критичен вопрос времени в случае большой дискретной БС (1600 переменных), для которой время предсказания одного образца даже дольше времени обучения всей модели (поэтому не даны времена расчёта). Данная проблема определяет ограниченную применимость столь огромных БС, даже при учёте бинаризации всех данных.

Таблица 6: Время предсказания метки неизвестного образца для разных алгоритмов в расчёте на один образец. Для классификаторов, обученных на выходных переменных автоэнкодера, в скобках даны дополнительные данные по времени работы автоэнкодера.

Метод	Время	
	Обучение модели	Предсказание
Лог. регрессия (а/э)	1мин. 16сек. (+1час 30мин.)	0.06мсек.
НБК (а/э)	8мин. (+1час 30мин.)	0.02мсек.
Гибридная БС (а/э)	50мин. 47сек. (+1час 30мин.)	1.8мсек.
Дискретная БС	3мин. 14сек.	9мин.
РНРТ (гл. угол)	18час 19мин	1.1сек.
РНМР (гл. угол)	28мин. 46сек.	1.1сек.

Эксперименты по обучению классификаторов показали, что при вдумчивом выборе признакового пространства и тщательной настройке применяемых моделей возможно достичь точности идентификации вплоть до 85% даже при условии наличия большого и разнородного негативного класса. Для более наглядного анализа работы использованных алгоритмов были построены матрицы ошибок [207] для каждой использованной модели (Рисунки 27, 28, 29, 30, 31, 32) по результатам предсказания на выборках Тест 1. Матрицы ошибок показывают, какие метки алгоритмы назначают для каждого класса и с какой частотой, что помогает оценивать возможные проблемы в структуре данных или организации вычислений. Для рассмотренных моделей, большинство случаев ошибок алгоритмов относилось к ситуации, когда образцы ошибочно помещались в негативный класс и наоборот, образцы негативного класса в некоторых случаях получали метки других классов. Таким образом, в сомнительных случаях алгоритмы имели тенденцию назначать образцу метку негативного класса значительно чаще, чем какую-либо другую. В качестве явных исключений в этом правиле можно указать пары видов *Bidens Tripartita* – *Anethum graveolens* и *Aerva lanata* – *Salvia officinalis*, которые устойчиво (вплоть до 30%) ошибочно принимались друг за друга большинством моделей. Данная ситуация может быть объяснена похожестью векторов данных для данных пар.

Для более подробного анализа особенностей функционирования моделей строили матрицы ошибок.

[illegible]

Рис. 27: Усредненная по результатам кросс-валидации матрица ошибок для большой дискретной БС, построенной на бинаризованных данных.

[illegible]

Рис. 28: Усреднённая по результатам кросс-валидации матрица ошибок для логистической регрессии, построенной на кодированных данных.

Negative class	0.37	0.0	0.0	0.1	0.0	0.6	3.5	0.5	0.0	2.6	0.5	1.1	0.0	1.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.8	0.5	1.9	0.0	0.0	0.0	1.6	0.0	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.6	0.0
Bergenia crassifolia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Betula pendula (buds)	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Helichrysum arenarium	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Valeriana officinalis	4.9	0.0	0.0	0.0	0.0	0.0	85.3	0.4	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.4	0.0	4.1	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	
Melilotus officinalis	4.2	0.0	0.0	0.0	0.0	0.0	17.7	7.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Origanum vulgare	10.6	0.0	0.0	0.0	0.0	0.0	80.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Panax ginseng	1.6	0.0	0.0	0.0	0.0	0.0	86.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	1.6	0.0	2.4	0.8	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	2.4	0.0	0.0	
Hypericum perforatum	49.3	0.0	0.0	0.0	0.0	0.0	80.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Coriandrum sativum	47.5	0.0	0.0	0.0	0.0	0.0	59.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0</																																	

Рис. 29: Усреднённая по результатам кросс-валидации матрица ошибок для НБК, построенного на кодированных данных.

[illegible]

Рис. 30: Усреднённая по результатам кросс-валидации матрица ошибок для разложения Таккера с метрикой главного угла.

Negative class		0.0	0.2	0.0	0.1	1.9	0.2	0.0	0.8	2.6	0.2	0.0	0.0	0.0	0.2	0.0	0.5	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.4	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.4	0.0	0.6	0.0	0.0	
Bergenia crassifolia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Betula pendula (buds)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Helichrysum arenarium	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Valeriana officinalis	0.8	0.0	0.0	2.0	90.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.8	0.0	0.0	0.0	
Melilotus officinalis	0.0	0.0	0.0	0.0	0.0	95.8	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Origanum vulgare	0.0	0.0	0.0	0.0	0.0	96.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Panax ginseng	0.0	0.0	0.0	0.0	0.0	0.0	89.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	3.2	0.0	0.0	0.0
Hypericum perforatum	5.3	0.0	0.0	0.0	0.0	0.0	0.0	94.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Coriandrum sativum	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	91.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	1.2																				

Рис. 31: Усреднённая по результатам кросс-валидации матрица ошибок для матричного разложения с метрикой главного угла.

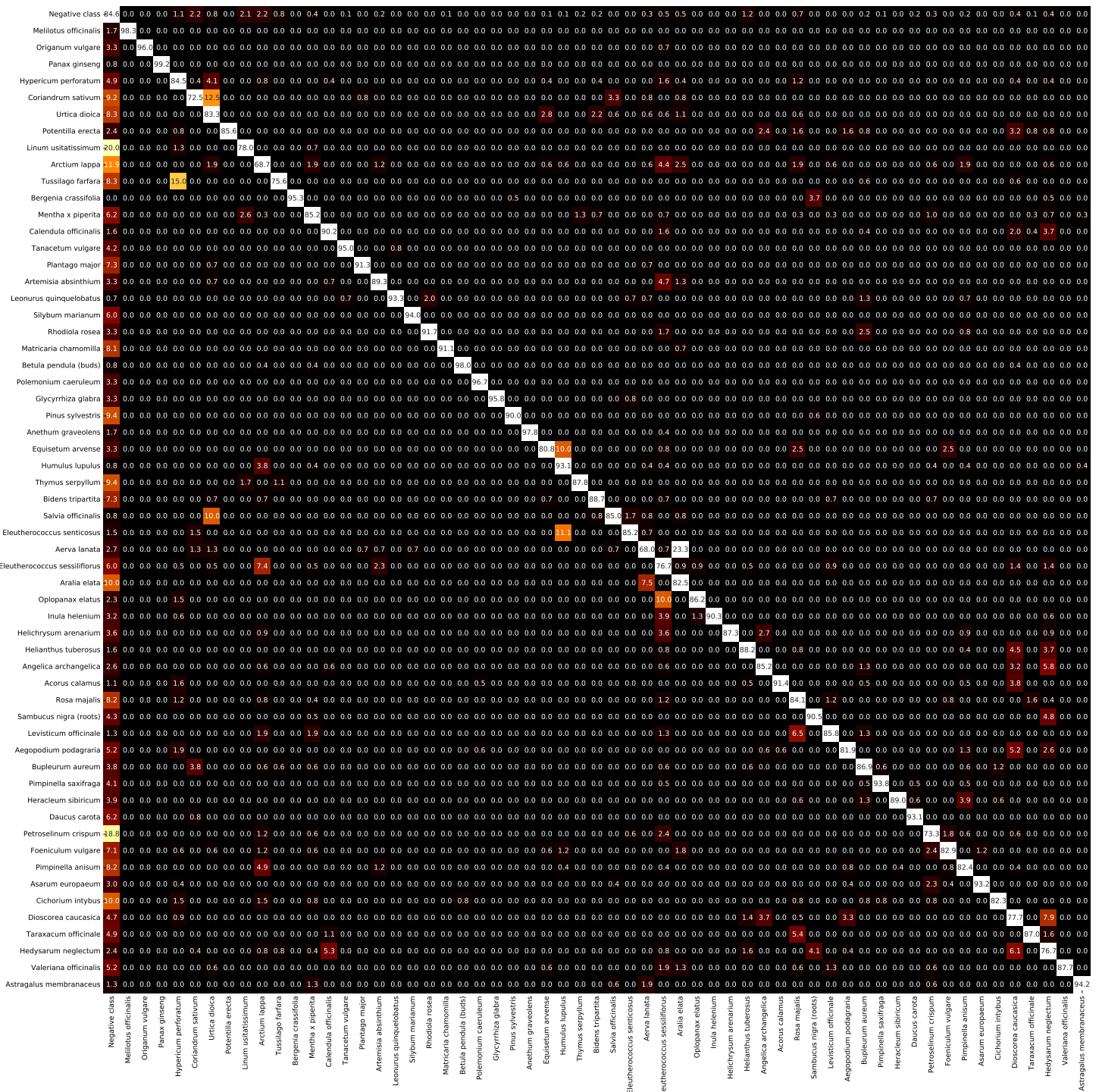


Рис. 32: Усреднённая по результатам кросс-валидации матрица ошибок для гибридной БС, построенной на кодированных данных.

Сравнительно высокие показатели эффективности моделей на выборке Тест2 показали, что модели, обученные на выборке данных сгенерированных в одном наборе условий (одна и та же ВЭЖХ-МС система и методика экстракции), могут

быть использованы для идентификации образцов, проанализированных в других условиях, включая процедуру экстракции и/или аналитическое оборудование. С этой точки зрения наилучшие результаты были показаны классификатором на основе разложения Таккера и метрики главного угла, тогда как Логистическая Регрессия показала себя значительно хуже, указывая на её переобучение.

Для отображения относительного родства представленных в работе видов было построено филогенетическое дерево (Рисунок 33) с использованием платформы PhyloT [208]. Филогенетическое дерево позволяет увидеть преимущество гетерогенной видовой структуры собранной базы данных по сравнению с более традиционным использованием в подобных подходах малого числа видов: среди видов имеются группы с различными степенями родства. Это позволяет попробовать оценить насколько используемая в работе схема предобработки данных позволяет соотнести межвидовые филогенетические дистанции и химическую информацию для образцов этих видов. Для реализации данной задачи применяли широко используемый НСА. Как уже обсуждалось в литературном обзоре, несмотря на широкую популярность данного метода обучения без учителя при исследовании структуры данных, нет единых критериев выбора метрик связывания и расчёта дистанций. Полезным методом в такой ситуации может служить так называемый grid search ("сеточный поиск" или метод перебора), который и был адаптирован к данной проблеме. Также была выбрана цветовая схема для более адекватной интерпретации получаемых дендрограмм (Рисунок 34).

Иерархический кластерный анализ проводили как для исходного признакового пространства (1600 переменных), так и для кодированных данных (25 переменных). Результаты представленные на Рисунке 35 позволяют увидеть, что кодирование данных приводит к сужению дистанций между образцами одного класса.

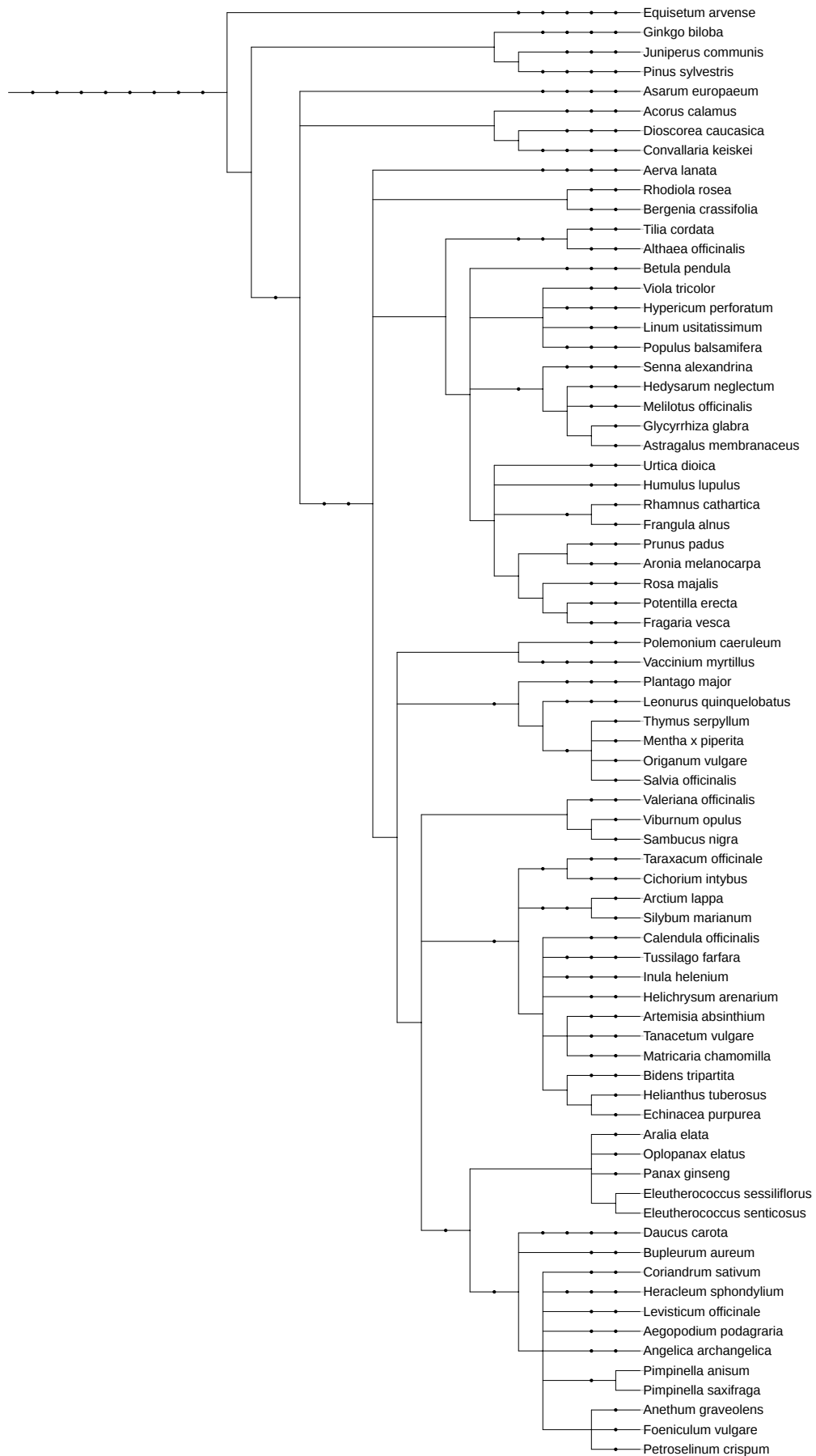


Рис. 33: Филогенетическое дерево 74 видов растений. Точки представляют классификационные единицы: рода, семейства etc. Составлено на основе таксономии NCBI с использованием платформы PhyloT [208].

Althaea officinalis	Coriandrum sativum	Leonurus quinquelobatus	Prunus padus	Levisticum officinale
Aronia melanocarpa	Urtica dioica	Silybum marianum	Vaccinium myrtillus	Aegopodium podagraria
Bergenia crassifolia	Frangula alnus	Rhodiola rosea	Salvia officinalis	Bupleurum aureum
Betula pendula (buds)	Convallaria transcaucasica	Matricaria chamomilla	Eleutherococcus senticosus	Pimpinella saxifraga
Betula pendula (leaves)	Potentilla erecta	Senna alexandrina	Aerva lanata	Heracleum sibiricum
Helichrysum arenarium	Tilia cordata	Polemonium caeruleum	Echinacea purpurea	Daucus carota
Sambucus nigra (flowers)	Linum usitatissimum	Glycyrrhiza glabra	Eleutherococcus sessiliflorus	Petroselinum crispum
Valeriana officinalis	Arctium lappa	Pinus sylvestris	Aralia elata	Foeniculum vulgare
Ginkgo biloba	Tussilago farfara	Populus balsamifera	Oplopanax elatus	Pimpinella anisum
Melilotus officinalis	Juniperus communis	Anethum graveolens	Inula helenium	Asarum europaeum
Origanum vulgare	Mentha x piperita	Viola tricolor	Helianthus tuberosus	Cichorium intybus
Panax ginseng	Calendula officinalis	Equisetum arvense	Angelica archangelica	Dioscorea caucasica
Rhamnus cathartica	Tanacetum vulgare	Humulus lupulus	Acorus calamus	Taraxacum officinale
Fragaria vesca	Plantago major	Thymus serpyllum	Rosa majalis	Hedysarum neglectum
Hypericum perforatum	Artemisia absinthium	Bidens tripartita	Sambucus nigra (roots)	Astragalus membranaceus
Viburnum opulus				

Рис. 34: Схема цветового кодирования для 76 классов.

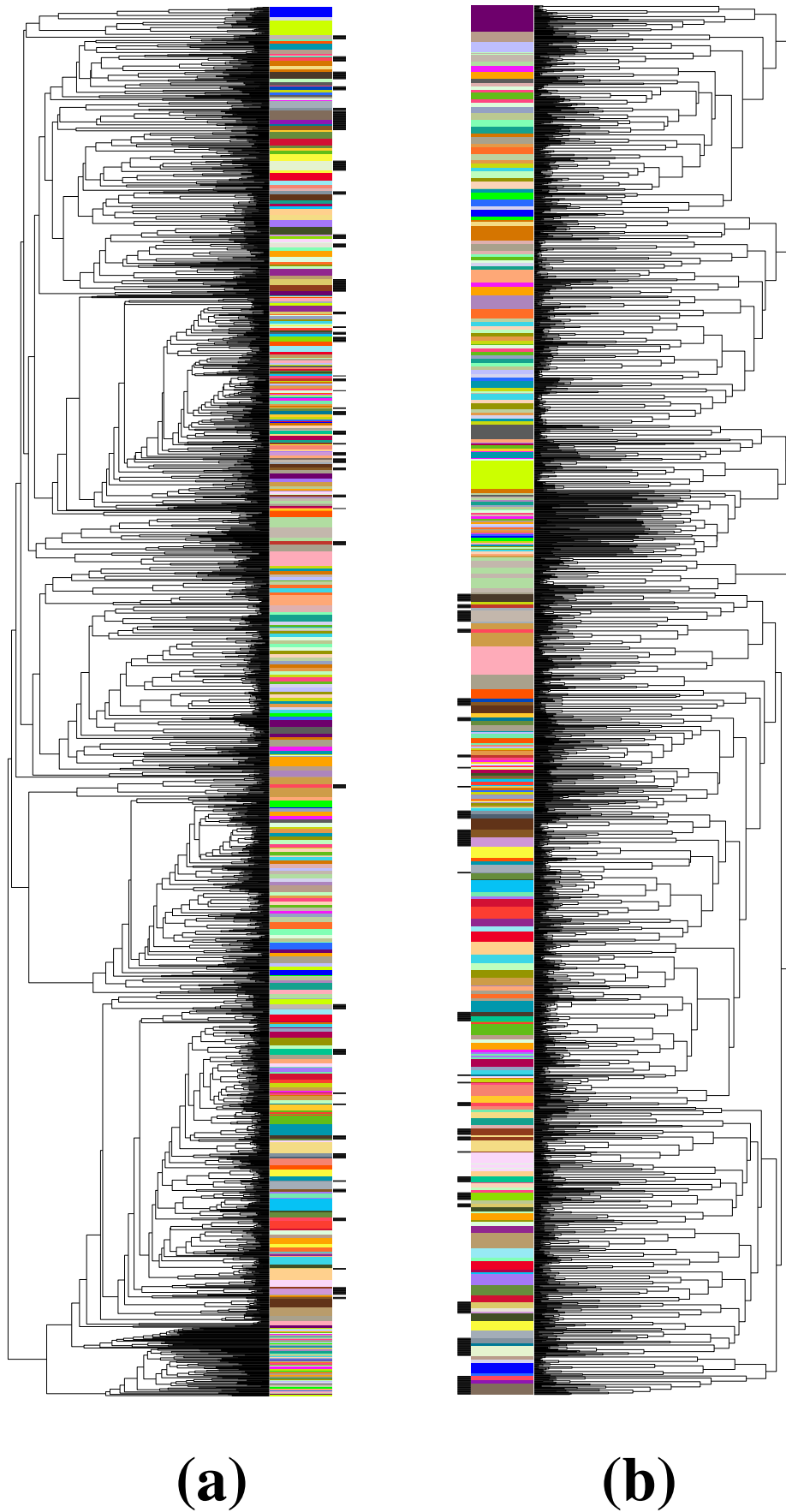


Рис. 35: Иерархический кластерный анализ всех образцов базы данных на основе исходных данных (a) и данных после кодирования (b); черные метки рядом с цветными индикаторами указывают на образцы негативного класса.

Тем не менее, данный вариант кластерного анализа мало подходит для анализа межвидовых дистанций. Поэтому НСА-дендрограммы строили также для усредненных (арифметическое среднее) образцов каждого из 76 классов (Рисунок 36).

В условиях, когда химические данные, собранные с использованием различных частей растений (корни, цветы, семена и т.д.), использовали для измерения межвидовых дистанций вполне естественно ожидать ограниченные результаты от подобного анализа. Среди множества использованных метрик связей и измерения расстояний (евклидово, Махаланобиса и т.д.) НСА в основном имел тенденцию группировать близкородственные виды (из одного рода или семейства) если для них использовались одни и те же части растения, тогда не было получено никакого сходства группировок более высоких порядков полученных после НСА с таковыми для филогенетического древа. Подобная ситуация наблюдалась как для первичных данных (1600 переменных) так и для кодированных (25 переменных).

Для визуализации взаимных расстояний между образцами выборки использовали алгоритм t-SNE (t-distributed Stochastic Neighbor Embedding) для первичных и кодированных данных (Рисунки 37 и 38).

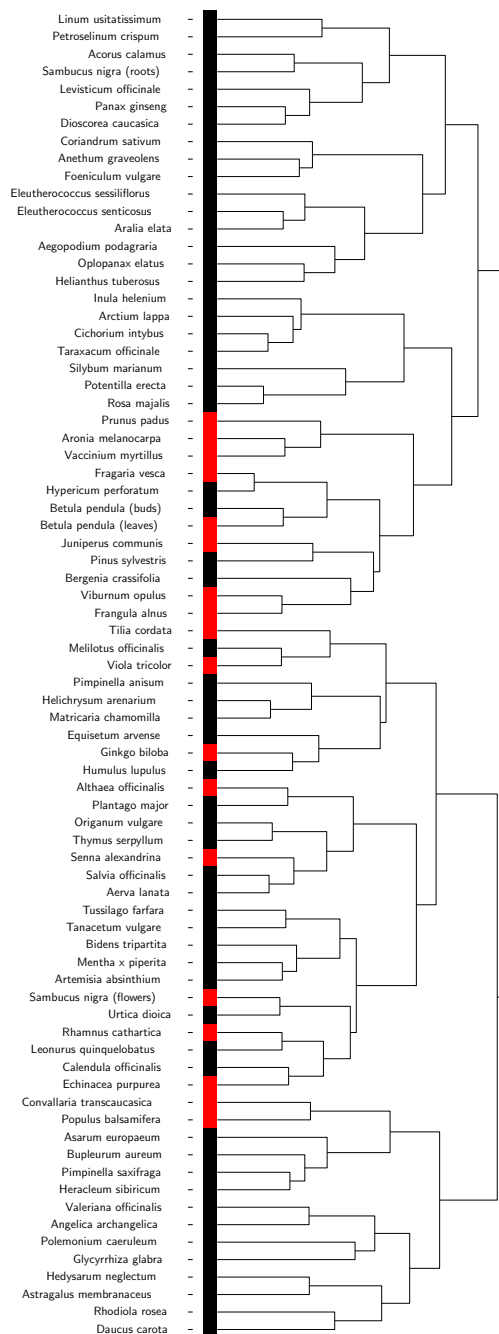


Рис. 36: Кластерный анализ: НСА-дендрограмма для усредненных векторов 76 классов. Красные маркеры обозначают виды, объединенные в негативный класс.

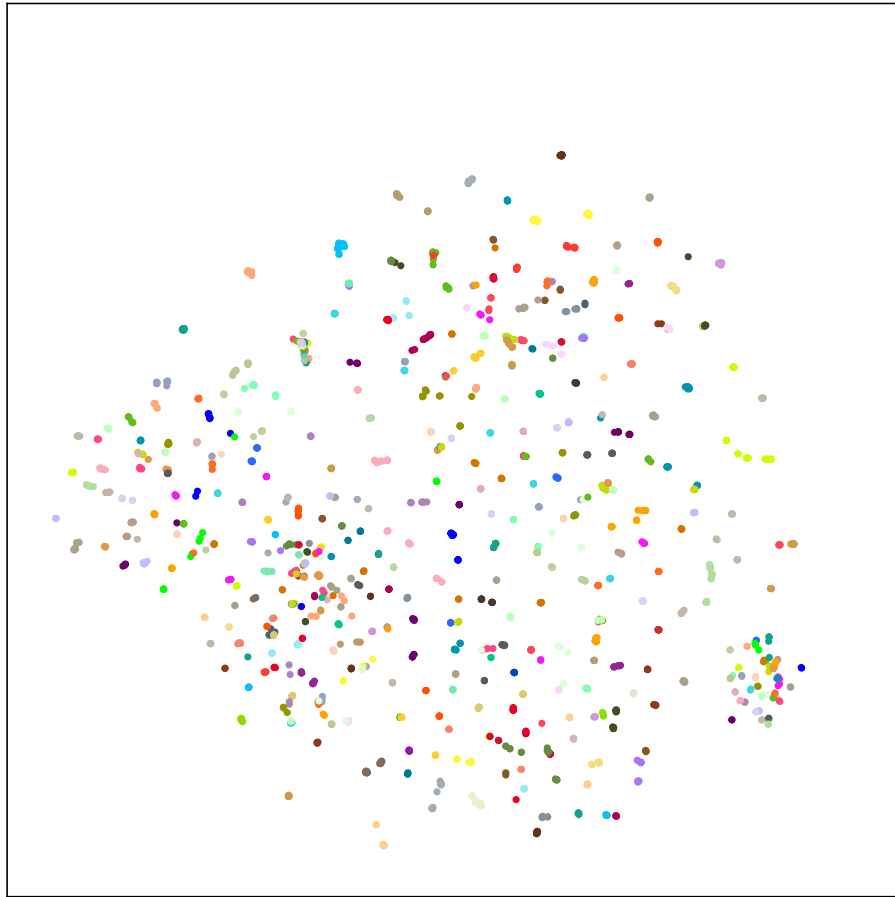


Рис. 37: t-SNE визуализация полной базы данных в исходной форме (1600 переменных). Цветовая схема соответствует приведённой ранее. Параметр Perplexity=10.

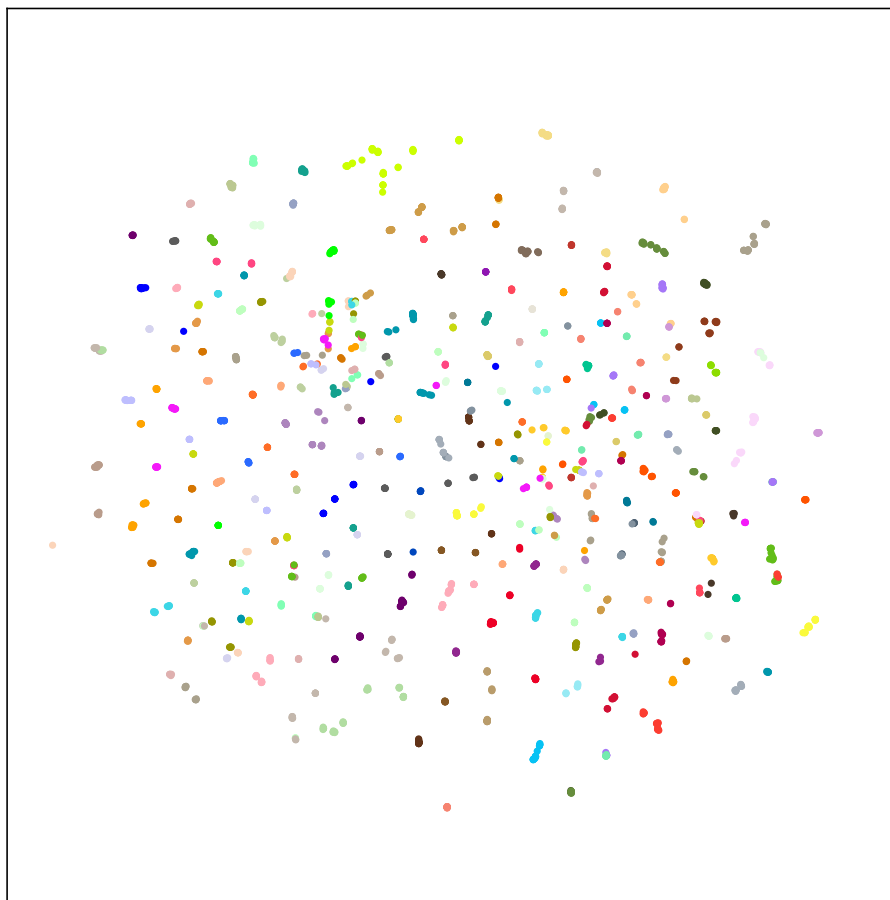


Рис. 38: t-SNE визуализация полной базы данных в кодированной форме (25 переменных). Цветовая схема соответствует приведённой ранее. Параметр Perplexity=10.

Визуализация с использованием t-SNE показывает, что кодирование приводит к большей разреженности структуры данных и исчезновению кластеров (Рисунок 37) некоторых классов имевшихся в исходных данных. Таким образом, помимо уменьшения признакового пространства и увеличения скорости расчётов алгоритмов, кодирование в том числе может помогать улучшать разделимость некоторых классов в случае похожих химических профилей.

Несмотря на то, что использованные для обучения алгоритмов классификации данные получали на ВЭЖХ-МС оборудовании, выбранная процедура предобработки данных оставляла в основном масс-спектрометрические данные - зна-

чения m/z и площадей хроматографических пиков обнаруженных соединений. При этом, среди тысяч видов высших растений, одним и тем же целочисленным значениям m/z могут соответствовать от нескольких до десятков и даже сотен различных соединений, не имеющих в общем случае ни функциональной, ни структурной общности. Поэтому результат применения НСА был вполне неудивителен. Для более ясного выражения данной позиции можно также использовать пример структуры большой Байесовской сети (Рисунок 39), обученной на данных.

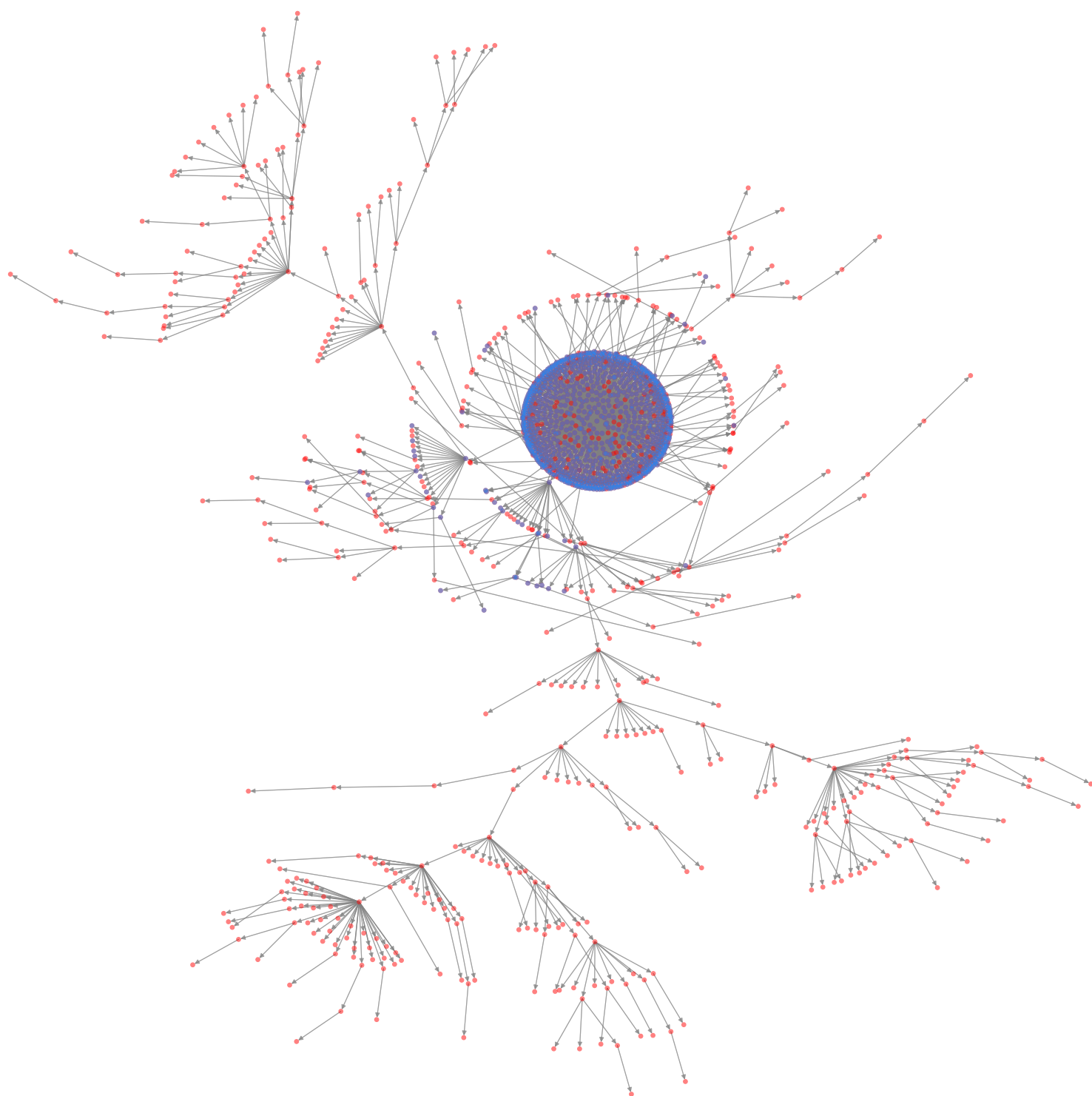


Рис. 39: Пример структуры дискретной БС обученной на 1600 переменных. Основной кластер переменных сгруппирован вокруг переменной 'identity', ответственной за классификацию. Все остальные узлы сети представляют значения m/z хроматографических пиков, а ребра графа отображают условную зависимость между переменными. Голубым цветом отмечена узловая структура общая для всех сетей в кросс-валидации.

Вершины графа, вовлеченные в сложные многоуровневые условные зависимости, которые при этом были бы устойчивы при различных разбиениях выборки в процессе кросс-валидации, составляют очень малую долю (~ 20 против 1600). Структура большей части переменных соответствовала случаю наивного байесовского классификатора, где значения вершин m/z строго зависело только от переменной класса. Если же переменные соответствовали какому-либо набору известных вторичных метаболитов растений, можно было бы с высокой вероятностью ожидать более сложной и осмысленной структуры сети при обучении её на данных. Из вышесказанного можно сделать вывод, что использованная в работе процедура предобработки данных приводила к векторам данных, расстояния между которыми мало коррелировали с соответствующими филогенетическими отношениями видов.

Немного иначе в данном аспекте выглядит ситуация с РНРТ и РНМР. Для данных моделей из фактор-матриц оси m/z были выбраны 3 строки (фактора), имеющие наибольшую внутривидовую и наименьшую межвидовую корреляцию (Рисунок 40 и 41):

$$\min_{i=\overline{1,r}; i \notin \Omega} \frac{1}{n_c - 1} \sum_{j \neq l} |B[j, i]| - \log(|A[i]| + \varepsilon), \quad (4.7)$$

где Ω это набор уже отобранных компонент, $B \in \mathbb{R}^{n_c-1 \times r}$ - матрицы корреляции между компонентами класса l и другими классами, и $A \in \mathbb{R}^{r \times 1}$ это вектор корреляции между компонентами и образцами данного класса l .

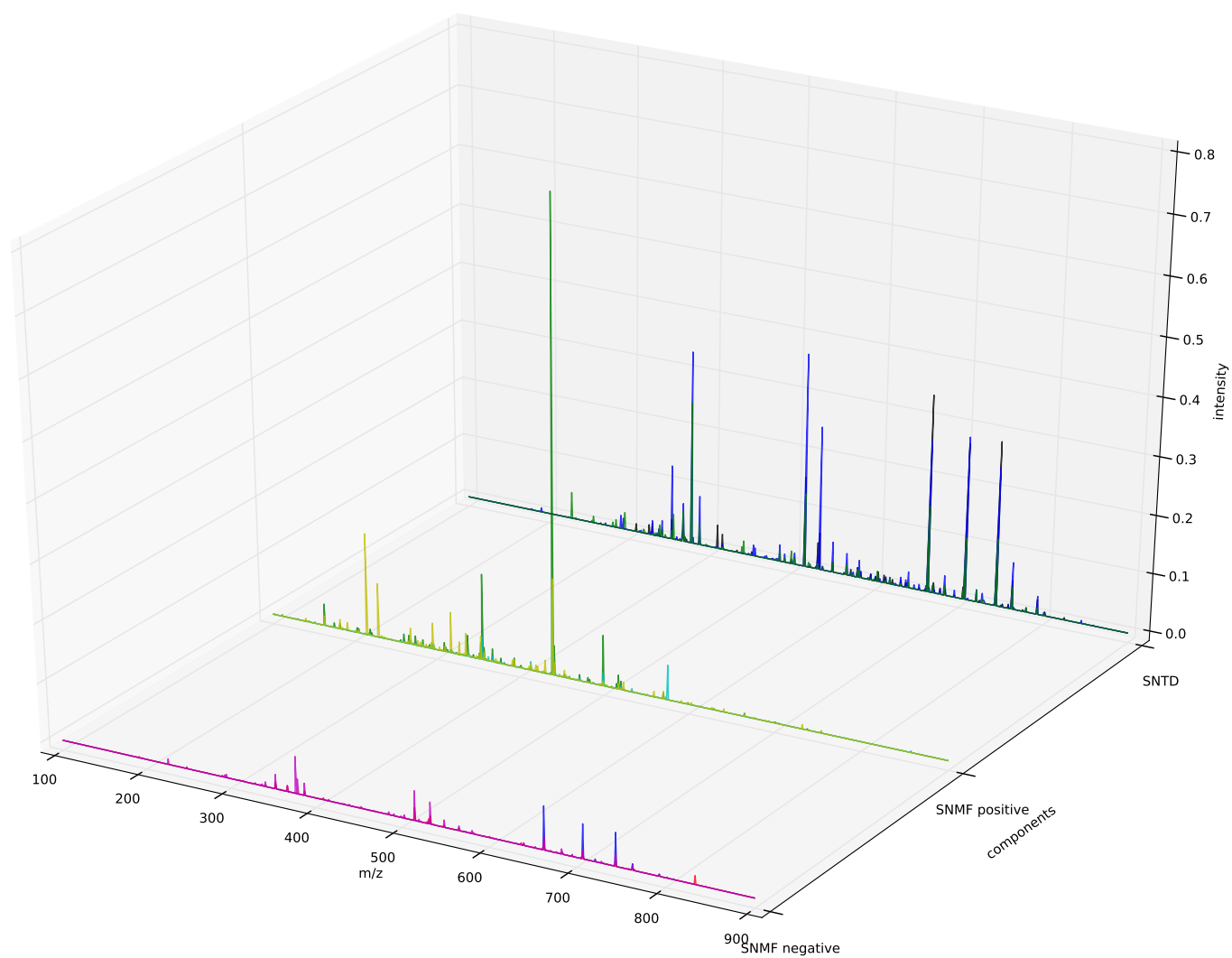


Рис. 40: Компоненты разложений РНРТ и РНМР для *Aralia elata*. Компоненты обозначены различными цветами.

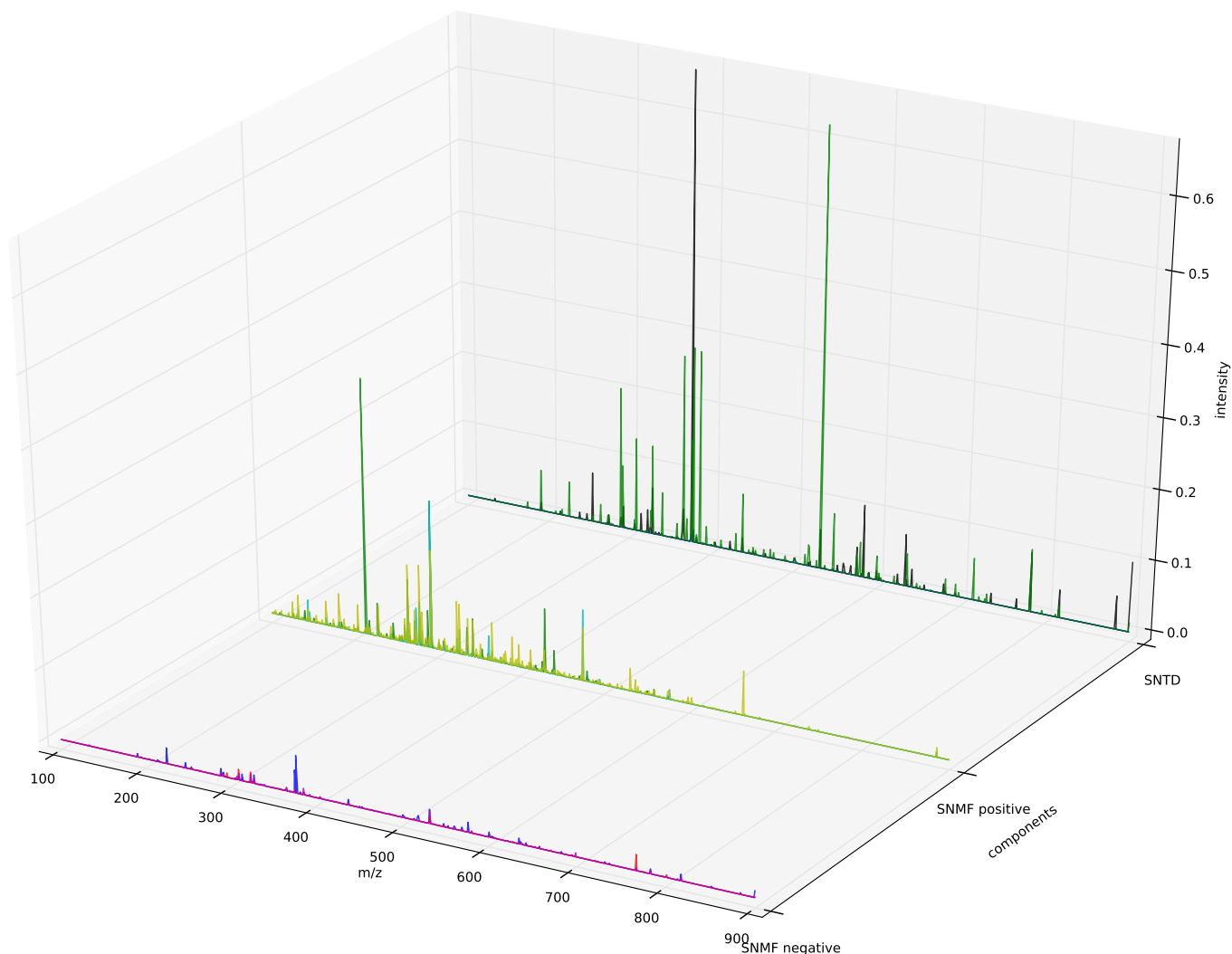


Рис. 41: Компоненты разложений РНРТ и РНМР для *Dioscorea caucasica* . Компоненты обозначены различными цветами.

Так как фактор-матрицы рассчитываются отдельно для каждого вида в выборке, они содержат информацию о характеристических наборах признаков с соответствующими относительными уровнями содержаний в экстрактах. Матричное и тензорное разложения показали одинаково высокую правильность идентификации на тестовой выборке Тест 1, тогда как на более гетерогенной выборке Тест 2 тензорное разложение значительно опередило матричное по всем характеристикам. Вполне вероятно, что РНМР имея в два раза больше переменных на каждый фактор по отношению к РНРТ и потому лучше отображает внутреннюю структуру обучающей выборки и схожей тестовой Тест 1, оно проигрывает в обобщающей

способности.

Если после ввода данных неизвестного образца вместо представления только самой вероятной метки класса, настроить алгоритмы на выдачу N наиболее вероятных кандидатов (стратегия выдачи результатов Топ N), эффективность обученных моделей при распознавании неизвестных образцов значительно возрастает (Таблица 7). Наиболее очевидное улучшение было достигнуто в случае большой дискретной БС на выборке Тест 2 (Таблица 8), где частота возникновения правильной метки возросла более чем на 20% по сравнению со стратегией "победитель получает всё". Несмотря на то, что точные показатели эффективности вполне вероятно упадут при увеличении размеров выборки и её гетерогенности, это говорит о высоком потенциале применения дискретных БС к подобным задачам. Таким образом, стратегию Топ N можно рассматривать как более предпочтительный подход представления результатов – сужение числа возможных кандидатов видовой принадлежности до 3-5 с ~95% или более точностью может быть более полезно на практике, чем ~80% точность в предсказании единственного кандидата.

Таблица 7: Сравнительные характеристики использованных подходов. В таблице представлены медианные значения, рассчитанные в процессе кросс-валидации. Подход Топ N, выборка Тест 1. Результат идентификации считается положительным, если правильная метка класса присутствует среди Топ N результатов.

Метод	Точность, %				
	Топ 1	Топ 2	Топ 3	Топ 4	Топ 5
Логистическая регрессия (а/э)	85.0	92.1	94.7	95.6	96.6
Наивный Баевсовский классификатор (а/э)	68.4	77.9	82.1	85.5	87.7
Большая дискретная БС	78.5	85.8	86.9	87.6	87.6
Разложение Таккера (главный угол)	77.9	82.8	84.2	86.0	86.9
Матричное разложение (главный угол)	75.4	79.4	80.7	82.0	82.7

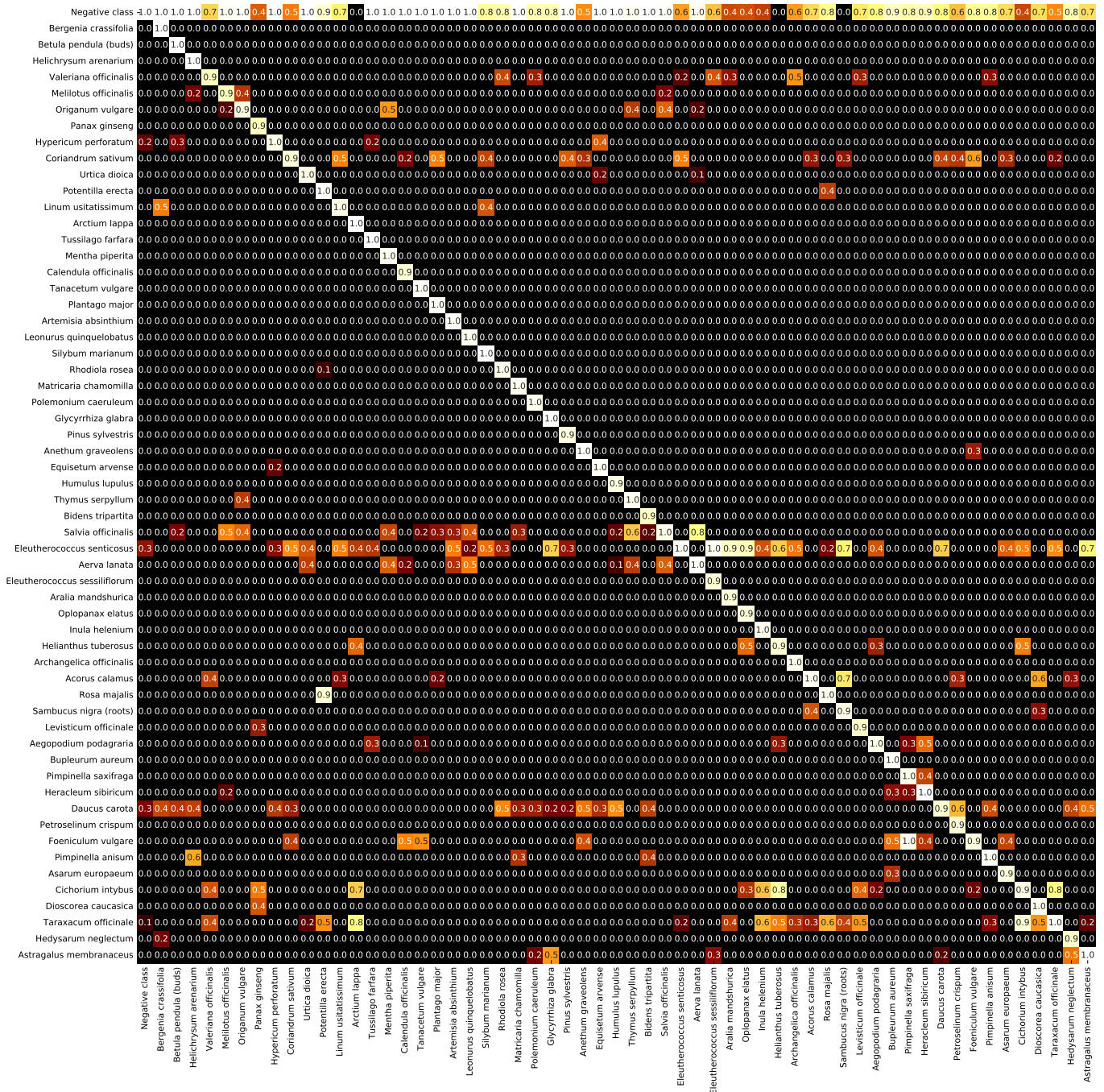
Таблица 8: Сравнительные характеристики использованных подходов. В таблице представлены медианные значения, рассчитанные в процессе кросс-валидации. Подход Топ N, выборка Тест 2. Результат идентификации считается положительным, если правильная метка класса присутствует среди Топ N результатов.

Метод	Точность, %				
	Топ 1	Топ 2	Топ 3	Топ 4	Топ 5
Логистическая регрессия (а/э)	72.7	79.6	84.1	86.4	88.6
Наивный Баевсовский классификатор (а/э)	77.3	86.4	88.6	93.2	93.2
Большая дискретная БС	72.7	81.8	88.6	90.9	93.2
Разложение Таккера (главный угол)	86.4	88.6	90.9	90.9	93.2
Матричное разложение (главный угол)	81.8	84.1	86.4	86.4	88.6

Для оценки того, какие виды растений алгоритм воспринимает как похожие друг на друга, применяли также так называемый "анализ близостей анализ частот возникновения тех или иных видов в Топ5 для каждого из видов. (Рисунки [42,43, 44, 45, 46](#)).

[illegible]

Рис. 42: Матрица близостей. Логистическая регрессия на кодированных данных.



[illegible]

Рис. 44: Матрица близостей. Большая дискретная БС на бинаризованных данных.

[illegible]

Рис. 45: Матрица близостей. Разреженное неотрицательное матризное разложение.

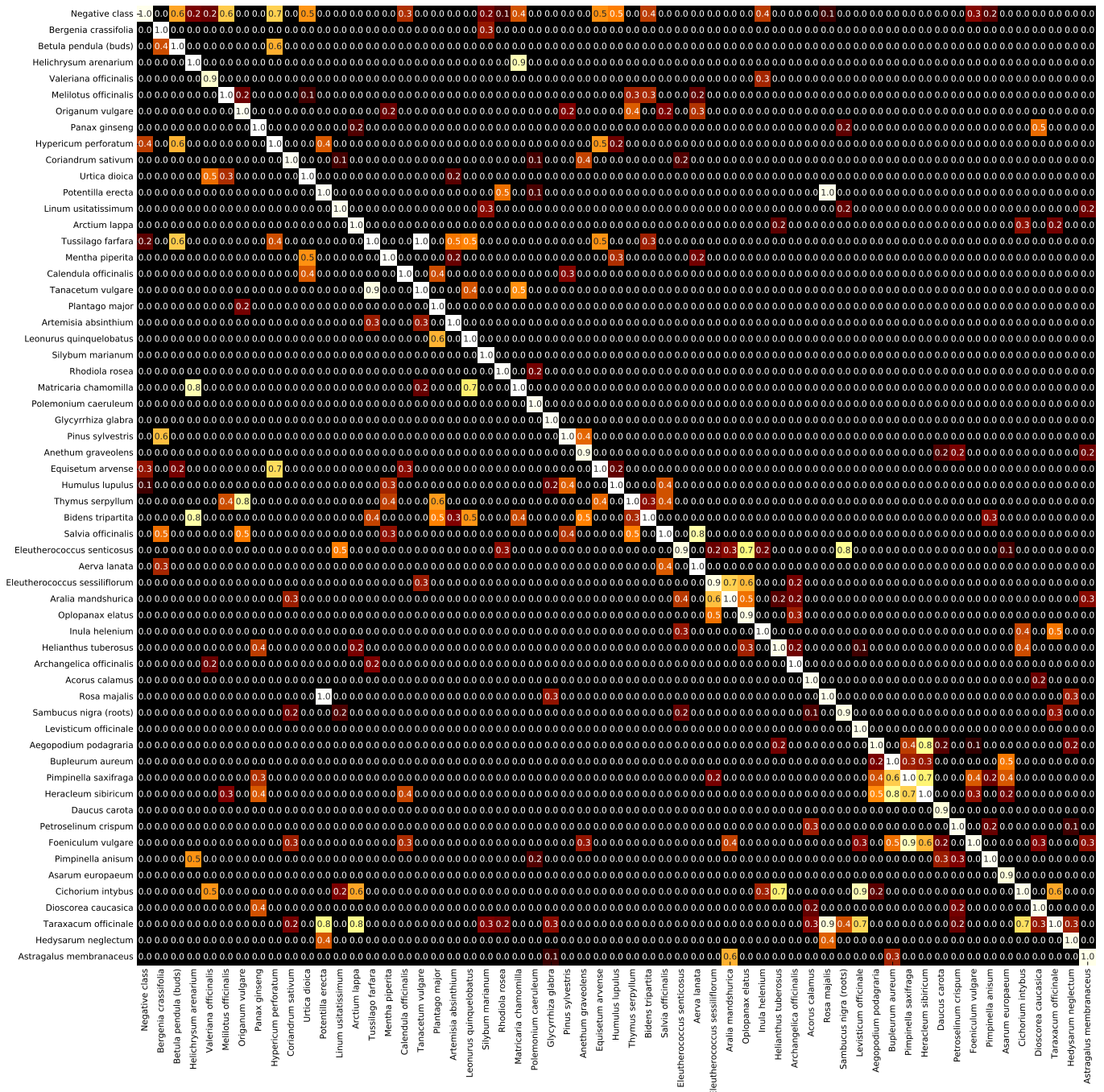


Рис. 46: Матрица близостей. Разреженное неотрицательное разложение Таккера.

Рассмотрение матриц для метода анализа близостей для случая РНРТ и РН-МР показывает тренд, схожий с результатами НСА - хотя бы часть видов-соседей для многих классов принадлежит к числу родственных (Рисунки 45, 46. В то же самое время дискретная БС показывает резко отличные результаты (Рисунок 44). Всего 4 вида - *Tussilago farfara*, *Polemonium caeruleum*, *Glycyrrhiza glabra* и *Astragalus membranaceus* являлись соседями почти для всех остальных видов в

базе. На данном примере можно явно увидеть недостаток бинаризации данных: информация о наличии соединения с определённым значением m/z без сведений о соотношении содержаний между различными пиками приводит к тому, что растения, содержащие набор наиболее статистически часто встречаемых видов автоматически имеют высокий приоритет появиться в результатах любого поискового запроса. Тем не менее, при использовании стратегии Топ N такой тип данных становится весьма устойчив к изменению условий получения данных, т.е. условий экстракции и ВЭЖХ-МС анализа. Несмотря на то, что соотношения интенсивностей соединений в профиле экстракта может колебаться в сильных пределах при экстракции различными растворителями и анализе на различном хроматографическом оборудовании, список характеристичных значений m/z не претерпевает значимых изменений, а значит и бинарная форма представления профиля оказывается устойчивой.

Логистическая регрессия (Рисунок 42) и НБК (Рисунок 43) оказались в промежуточной категории – в них присутствуют как и случаи родственных видов, оказывающихся в числе самых частых соседей друг друга, так и ситуация схожая с дискретной БС: класс *Eleutherococcus senticosus* в НБК попал в топ 5 соседей для почти половины классов.

Для оценки различий во вкладе переменных в различение видов использовали матрицы весов автоэнкодера. Несмотря на то, что вследствие нелинейных преобразований на каждом слое автоэнкодера нельзя провести проекцию 1600 исходных переменных в итоговые 25, за счёт использования матрицы весов первого кодирующего слоя (участвующего в преобразовании исходных данных) проводили непрямую проекцию важности переменных. Для этого список переменных отсортировали в соответствии с нормами столбцов матрицы весов W (размером 400×1600). Полученные индексы сортировки переменных использовали двумя способами: в виде сортировки по убыванию (Рисунок 47) и сортировки по возрастанию (Рисунок 48). Далее на основе сортированных списков переменных обучали логистическую регрессию с регуляризацией на кодированных переменных. Все переменные, кроме использованных на данном шаге обнуляли и кодировали, а полученные 25 переменных уже использовали как признаковое пространство. Данную процедуру проводили для 5 различных разбиений и строили графики на основе медианных значений. Рисунок 47 позволяет увидеть, что уже 200-300

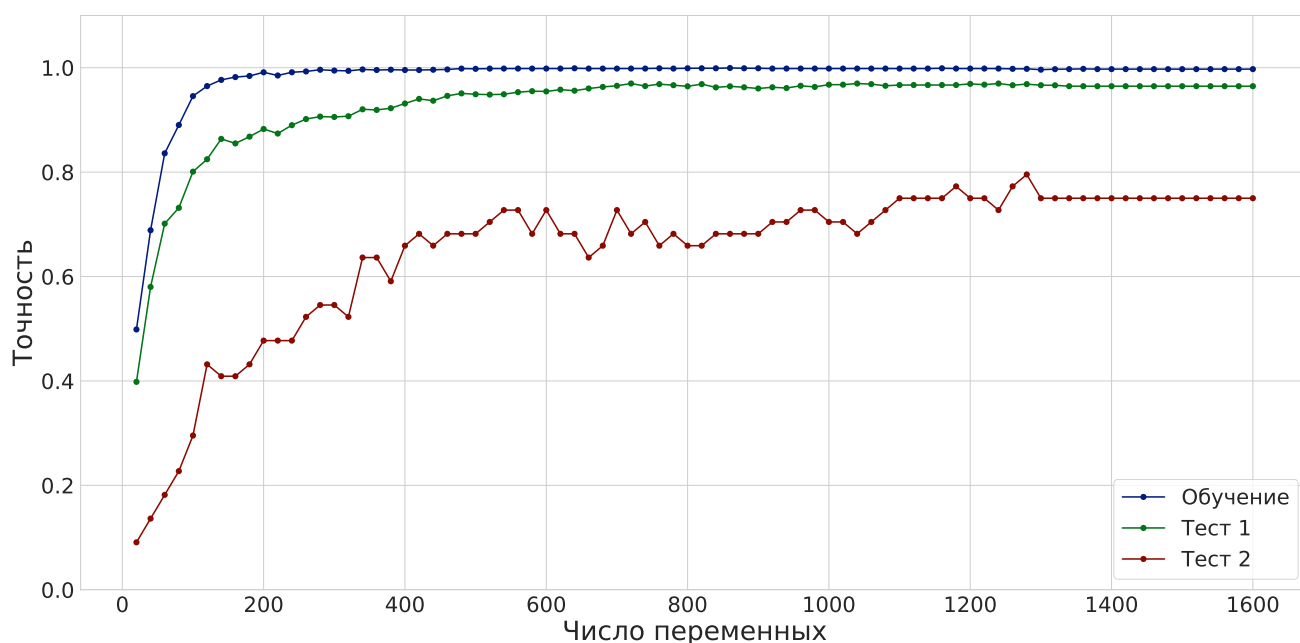


Рис. 47: Анализ значимости переменных в векторах данных. Точность модели логистической регрессии, обученной на пошагово увеличивающемся наборе переменных. Переменные были отсортированы на основе норм столбцов для матрицы весов 1 слоя автоэнкодера в порядке убывания.

выбранных вышеописанным способом переменных позволяют достичь более чем 90% точность предсказаний на Тесте 1. Данный результат с точки зрения химической интерпретации легко ассоциируется с ограниченным набором значений m/z , которые для имеющейся базы данных вносят наибольший вклад в различие видов, т.е. являются в некоторой мере характеристическими. При этом график построенный начиная с наборов переменных имеющих низкие соответствующие веса в матрице автоэнкодера ожидаемо показал малую эффективность даже не выборках более 600 переменных (Рисунок 48).

Несмотря на то, что возможно отбросить практически половину переменных (750-800 из 1600) без каких-либо заметных потерь в эффективности алгоритмов, подобная мера предобработки данных неустойчива по отношению к составу базы данных. Новые добавляемые виды могут иметь характеристические наборы переменных среди списка уже отфильтрованных. Таким образом, вместо попытки отфильтровать несколько сотен менее значимых переменных, более выгодной стратегией представляется использование автоэнкодера. Признаковое пространство размером всего 25 переменных и сохраняющее при этом структуру исход-

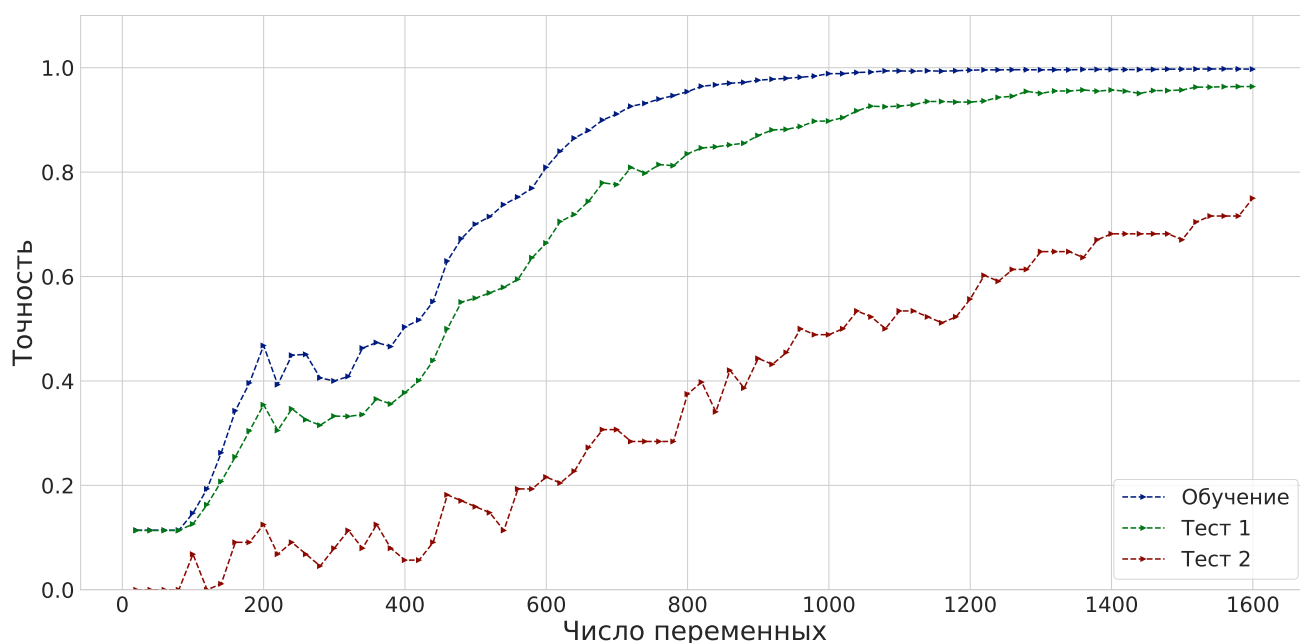


Рис. 48: Анализ значимости переменных в векторах данных. Точность модели логистической регрессии, обученной на пошагово увеличивающемся наборе переменных. Переменные были отсортированы на основе норм столбцов для матрицы весов 1 слоя автоэнкодера в порядке возрастания.

ных данных позволяет ускорить работу любой из использованных моделей, при этом структуру самого автоэнкодера достаточно обучить один раз для всей базы данных. Несмотря на то, что на сегодняшний день известны тысячи видов высших растений, значительная доля которых обладает той или иной физиологической активностью, активно в производстве различных препаратов применяют не более нескольких сотен. Такая ситуация позволяет без особых ограничений использовать комбинацию высокоэффективной жидкостной хроматографии и масс-спектрометрии низкого разрешения для рутинной видовой идентификации растительного материала. Безусловно, для этого требуется соответствующая база данных проанализированных экстрактов. При этом, неизвестные образцы, идентификацию которых необходимо провести, могут быть как в форме порошков, так и в виде экстрактов различными растворителями. Более того, скорее всего можно использовать любую форму препарата растения, которая в достаточно полной мере сохранила основной набор характеристических соединений, присущих данному виду.

ВЫВОДЫ

1. Получены данные ВЭЖХ-МС анализа экстрактов лекарственных растений 74 видов лекарственных растений, составлена и опубликована в интернете соответствующая база данных (БД). БД включает данные о 654 образцах и 2262 хроматограммах; ионный ток регистрировали при значениях m/z 100–900, в режиме регистрации как положительных, так и отрицательных ионов.
2. Разработан подход к видовой идентификации лекарственных растений на основе комбинации хроматомасс-спектрометрического анализа водно-спиртовых экстрактов растительного сырья и методов машинного обучения. Предложенный подход не требует применения индивидуальных стандартных соединений и использует данные масс-спектрометрии низкого разрешения.
3. Изучены подходы к обработке первичных аналитических данных и предложен подход с отбрасыванием значений времен удерживания компонентов пробы, позволяющий сохранить высокую эффективность алгоритма идентификации при малом числе используемых переменных.
4. Предложенный подход проверен на устойчивость к внесению искажений. Показано, что алгоритмы, разработанные на одном оборудовании и в одних условиях экстракции, способны показывать высокую правильность идентификации при использовании данных, полученных на другом приборном комплексе и/или с использованием других условий экстракции.

Литература

1. Williams P. Health benefits of herbs and spices: Public health // Medical Journal of Australia. — 2006. — Vol. 4, no. 4. — P. S17–S18.
2. Hostettmann K., Marston A. Twenty years of research into medicinal plants: results and perspectives // Phytochemistry Reviews. — 2002. — Vol. 1, no. 3. — P. 275–285.
3. Li P., Qi L.-W., Liu E.-H. et al. Analysis of chinese herbal medicines with holistic approaches and integrated evaluation models // TrAC Trends in Analytical Chemistry. — 2008. — Vol. 27, no. 1. — P. 66–77.
4. Goodacre R., York E. V., Heald J. K., Scott I. M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry // Phytochemistry. — 2003. — Vol. 62, no. 6. — P. 859–863.
5. Gorgulu S. T., Dogan M., Severcan F. The characterization and differentiation of higher plants by fourier transform infrared spectroscopy // Applied spectroscopy. — 2007. — Vol. 61, no. 3. — P. 300–308.
6. He K., Pauli G. F., Zheng B. et al. Cimicifuga species identification by high performance liquid chromatography–photodiode array/mass spectrometric/evaporative light scattering detection for quality control of black cohosh products // Journal of Chromatography A. — 2006. — Vol. 1112, no. 1-2. — P. 241–254.
7. Folashade O., Omoregie H., Ochogu P. et al. Standardization of herbal medicines—a review // International Journal of Biodiversity and Conservation. — 2012. — Vol. 4, no. 3. — P. 101–112.
8. Dahanukar S., Kulkarni R., Rege N. et al. Pharmacology of medicinal plants and natural products // Indian journal of pharmacology. — 2000. — Vol. 32, no. 4. — P. S81–S118.
9. European Parliament and of the Council. Directive 2004/24/ec. — 2004. — URL: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32004L0024&qid=1451884773824>.
10. Food and Drug Administration. Dietary supplements. — URL: <http://www.fda.gov/Food/DietarySupplements/>.
11. Kessler R. C., Davis R. B., Foster D. F. et al. Long-term trends in the use of complementary and alternative medical therapies in the united states // Annals of internal medicine. — 2001. — Vol. 135, no. 4. — P. 262–268.
12. Chaudhury R. R. Herbal remedies and traditional medicines in reproductive health care practices and their clinical evaluation // Journal of Reproductive Health and

- Medicine. — 2015. — Vol. 1, no. 1. — P. 44 – 46.
13. Petrovska B. B. Historical review of medicinal plants usage // Pharmacognosy reviews. — 2012. — Vol. 6, no. 11. — P. 1.
 14. Maroyi A. Traditional use of medicinal plants in south-central zimbabwe: re-view and perspectives // Journal of ethnobiology and ethnomedicine. — 2013. — Vol. 9, no. 1. — P. 31.
 15. Wang M.-W., Richard D. Y., Zhu Y. Pharmacology in china: a brief overview // Trends in pharmacological sciences. — 2013. — Vol. 34, no. 10. — P. 532–533.
 16. Jing J., Parekh H. S., Wei M. et al. Advances in analytical technologies to evaluate the quality of traditional chinese medicines // TrAC Trends in Analytical Chemistry. — 2013. — Vol. 44. — P. 39–45.
 17. Simmler C., Napolitano J. G., McAlpine J. B. et al. Universal quantitative nmr analysis of complex natural samples // Current opinion in biotechnology. — 2014. — Vol. 25. — P. 51–59.
 18. Bansal A., Chhabra V., Rawal R. K., Sharma S. Chemometrics: a new scenario in herbal drug standardization // Journal of Pharmaceutical Analysis. — 2014. — Vol. 4, no. 4. — P. 223–233.
 19. Liang X.-M., Jin Y., Wang Y.-p. et al. Qualitative and quantitative analysis in quality control of traditional Chinese medicines // Journal of Chromatography A. — 2009. — Vol. 1216, no. 11. — P. 2033–2044.
 20. Yu F., Kong L., Zou H., Lei X. Progress on the screening and analysis of bioactive compounds in traditional Chinese medicines by biological fingerprinting analysis // Combinatorial chemistry & high throughput screening. — 2010. — Vol. 13, no. 10. — P. 855–868.
 21. Hand D. J. Principles of data mining // Drug safety. — 2007. — Vol. 30, no. 7. — P. 621–622.
 22. Huang Y., Wu Z., Su R. et al. Current application of chemometrics in traditional Chinese herbal medicine research // J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. — 2016. — Jul. — Vol. 1026. — P. 27–35.
 23. Wang M. W., Ye R. D., Zhu Y. Pharmacology in China: a brief overview // Trends Pharmacol. Sci. — 2013. — Oct. — Vol. 34, no. 10. — P. 532–533.
 24. Jiang Y., David B., Tu P., Barbin Y. Recent analytical approaches in quality control of traditional Chinese medicines—a review // Anal. Chim. Acta. — 2010. — Jan. — Vol. 657, no. 1. — P. 9–18.
 25. Dong X., Wang R., Zhou X. et al. Current mass spectrometry approaches and

- challenges for the bioanalysis of traditional Chinese medicines // *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* — 2016. — Jul. — Vol. 1026. — P. 15–26.
26. Kunle O. F., Egharevba H. O., Ahmadu P. O. Standardization of herbal medicines—a review // *International Journal of Biodiversity and Conservation.* — 2012. — Vol. 4, no. 3. — P. 101–112.
 27. Liang Y.-Z., Xie P., Chan K. Quality control of herbal medicines // *Journal of Chromatography B.* — 2004. — Vol. 812, no. 1. — P. 53 – 70.
 28. Gad H. A., El-Ahmady S. H., Abou-Shoer M. I., Al-Azizi M. M. Application of chemometrics in authentication of herbal medicines: a review // *Phytochemical Analysis.* — 2013. — Vol. 24, no. 1. — P. 1–24.
 29. Mao J., Xu J. Discrimination of herbal medicines by molecular spectroscopy and chemical pattern recognition // *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy.* — 2006. — Vol. 65, no. 2. — P. 497–500.
 30. Zhao L., Huang C., Shan Z. et al. Fingerprint analysis of *Psoralea corylifolia* L. by HPLC and LC–MS // *Journal of Chromatography B.* — 2005. — Vol. 821, no. 1. — P. 67–74.
 31. Yue H., Ma L., Pi Z. et al. Fast screening of authentic ginseng products by surface desorption atmospheric pressure chemical ionization mass spectrometry // *Planta medica.* — 2013. — Vol. 29, no. 02. — P. 169–174.
 32. Tian R.-t., Xie P.-s., Liu H.-p. Evaluation of traditional chinese herbal medicine: Chaihu (*bupleuri radix*) by both high-performance liquid chromatographic and high-performance thin-layer chromatographic fingerprint and chemometric analysis // *Journal of Chromatography A.* — 2009. — Vol. 1216, no. 11. — P. 2150–2155.
 33. Schulz H., Baranska M., Quilitzsch R. et al. Characterization of peppercorn, pepper oil, and pepper oleoresin by vibrational spectroscopy methods // *Journal of agricultural and food chemistry.* — 2005. — Vol. 53, no. 9. — P. 3358–3363.
 34. Wang P., Yu Z. Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review // *J Pharm Anal.* — 2015. — Oct. — Vol. 5, no. 5. — P. 277–284.
 35. Farag M. A., Porzel A., Wessjohann L. A. Comparative metabolite profiling and fingerprinting of medicinal licorice roots using a multiplex approach of GC–MS, LC–MS and 1D NMR techniques // *Phytochemistry.* — 2012. — Vol. 76. — P. 60–72.
 36. Herrador M. A., Gonzalez A. G. Pattern recognition procedures for differentiation of green, black and oolong teas according to their metal content from in-

- ductively coupled plasma atomic emission spectrometry // *Talanta*. — 2001. — Vol. 53, no. 6. — P. 1249–1257.
37. Martín M. J., Pablos F., González A. Characterization of green coffee varieties according to their metal content // *Analytica chimica acta*. — 1998. — Vol. 358, no. 2. — P. 177–183.
 38. Kong W.-J., Zhao Y.-L., Xiao X.-H. et al. Spectrum–effect relationships between ultra performance liquid chromatography fingerprints and anti-bacterial activities of *Rhizoma coptidis* // *Analytica Chimica Acta*. — 2009. — Vol. 634, no. 2. — P. 279–285.
 39. Ning Z., Lu C., Zhang Y. et al. Application of plant metabonomics in quality assessment for large-scale production of traditional Chinese medicine // *Planta medica*. — 2013. — Vol. 79, no. 11. — P. 897–908.
 40. Tistaert C., Dejaegher B., Heyden Y. V. Chromatographic separation techniques and data handling methods for herbal fingerprints: A review // *Analytica Chimica Acta*. — 2011. — Vol. 690, no. 2. — P. 148 – 161.
 41. Riedl J., Esslinger S., Fauhl-Hassek C. Review of validation and reporting of non-targeted fingerprinting approaches for food authentication // *Analytica Chimica Acta*. — 2015. — Vol. 885. — P. 17 – 32.
 42. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks // *science*. — 2006. — Vol. 313, no. 5786. — P. 504–507.
 43. Goodfellow I., Bengio Y., Courville A., Bengio Y. Deep learning. — MIT press Cambridge, 2016. — Vol. 1.
 44. Springfield E. P., Eagles P. K., Scott G. Quality assessment of South African herbal medicines by means of HPLC fingerprinting // *J Ethnopharmacol*. — 2005. — Oct. — Vol. 101, no. 1-3. — P. 75–83.
 45. Fan C., Deng J., Yang Y. et al. Multi-ingredients determination and fingerprint analysis of leaves from *Ilex latifolia* using ultra-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry // *J Pharm Biomed Anal*. — 2013. — Vol. 84. — P. 20–29.
 46. Wan J. B., Bai X., Cai X. J. et al. Chemical differentiation of Da-Cheng-Qi-Tang, a Chinese medicine formula, prepared by traditional and modern decoction methods using UPLC/Q-TOFMS-based metabolomics approach // *J Pharm Biomed Anal*. — 2013. — Vol. 83. — P. 34–42.
 47. Sheridan H., Krenn L., Jiang R. et al. The potential of metabolic fingerprinting as a tool for the modernisation of tcm preparations // *Journal of Ethnopharmacology*. — 2012. — Vol. 140, no. 3. — P. 482 – 491.

48. Wang J., Kong H., Yuan Z. et al. A novel strategy to evaluate the quality of traditional chinese medicine based on the correlation analysis of chemical fingerprint and biological effect // *Journal of Pharmaceutical and Biomedical Analysis*. — 2013. — Vol. 83. — P. 57 – 64.
49. Chang Y. X., Ge A. H., Donnapee S. et al. The multi-targets integrated fingerprinting for screening anti-diabetic compounds from a Chinese medicine Jinqi Jiangtang Tablet // *J Ethnopharmacol.* — 2015. — Vol. 164. — P. 210–222.
50. Mazina J., Vaher M., Kuhtinskaja M. et al. Fluorescence, electrophoretic and chromatographic fingerprints of herbal medicines and their comparative chemometric analysis // *Talanta*. — 2015. — Vol. 139. — P. 233 – 246.
51. Cordero C., Rubiolo P., Cobelli L. et al. Potential of the reversed-inject differential flow modulator for comprehensive two-dimensional gas chromatography in the quantitative profiling and fingerprinting of essential oils of different complexity // *Journal of Chromatography A*. — 2015. — Vol. 1417. — P. 79 – 95.
52. Cheng H., Qin Z., Guo X. et al. Geographical origin identification of propolis using gc-ms and electronic nose combined with principal component analysis // *Food Research International*. — 2013. — Vol. 51, no. 2. — P. 813 – 822.
53. Tarachiwin L., Katoh A., Ute K., Fukusaki E. Quality evaluation of angelica acutiloba kitagawa roots by 1h nmr-based metabolic fingerprinting // *Journal of Pharmaceutical and Biomedical Analysis*. — 2008. — Vol. 48, no. 1. — P. 42 – 48.
54. Farag M. A., Porzel A., Wessjohann L. A. Unraveling the active hypoglycemic agent trigonelline in *Balanites aegyptiaca* date fruit using metabolite fingerprinting by NMR // *J Pharm Biomed Anal.* — 2015. — Vol. 115. — P. 383–387.
55. Fan Q., Chen C., Lin Y. et al. Fourier transform infrared (ft-ir) spectroscopy for discrimination of rhizoma gastrodiae (tianma) from different producing areas // *Journal of Molecular Structure*. — 2013. — Vol. 1051. — P. 66 – 71.
56. Liu W., Xu J., Zhu R. et al. Fingerprinting profile of polysaccharides from lycium barbarum using multiplex approaches and chemometrics // *International Journal of Biological Macromolecules*. — 2015. — Vol. 78. — P. 230 – 237.
57. Sandasi M., Kamatou G. P., Combrinck S., Viljoen A. M. A chemotaxonomic assessment of four indigenous south african lippia species using gc-ms and vibrational spectroscopy of the essential oils // *Biochemical Systematics and Ecology*. — 2013. — Vol. 51. — P. 142 – 152.
58. Huang H., Sun J., McCoy J.-A. et al. Use of flow injection mass spectrometric fingerprinting and chemometrics for differentiation of three black cohosh species // *Spectrochimica Acta Part B: Atomic Spectroscopy*. — 2015. — Vol. 105. — P. 121 – 129.

59. Nock C. J., Waters D. L., Edwards M. A. et al. Chloroplast genome sequences from total dna for plant identification // *Plant biotechnology journal*. — 2011. — Vol. 9, no. 3. — P. 328–333.
60. Jackson S., Rounsley S., Purugganan M. Comparative sequencing of plant genomes: choices to make // *The Plant Cell*. — 2006. — Vol. 18, no. 5. — P. 1100–1104.
61. Harris J. G., Harris M. W. *Plant identification terminology: an illustrated glossary*. — Spring Lake Publishing Spring Lake, Utah, 1994.
62. Chen S., Harmon A. C. Advances in plant proteomics // *Proteomics*. — 2006. — Vol. 6, no. 20. — P. 5504–5516.
63. Louden D., Handley A., Lafont R. et al. Hplc analysis of ecdysteroids in plant extracts using superheated deuterium oxide with multiple on-line spectroscopic analysis (uv, ir, 1h nmr, and ms) // *Analytical chemistry*. — 2002. — Vol. 74, no. 1. — P. 288–294.
64. Wolfender J.-L., Rodriguez S., Hostettmann K., Hiller W. Liquid chromatography/ultra violet/mass spectrometric and liquid chromatography/nuclear magnetic resonance spectroscopic analysis of crude extracts of gentianaceae species // *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*. — 1997. — Vol. 8, no. 3. — P. 97–104.
65. Dillard C. J., German J. B. *Phytochemicals: nutraceuticals and human health* // *Journal of the Science of Food and Agriculture*. — 2000. — Vol. 80, no. 12. — P. 1744–1756.
66. Proestos C., Chorianopoulos N., Nychas G.-J., Komaitis M. Rp-hplc analysis of the phenolic compounds of plant extracts. investigation of their antioxidant capacity and antimicrobial activity // *Journal of Agricultural and Food Chemistry*. — 2005. — Vol. 53, no. 4. — P. 1190–1195.
67. Klejdus B., Vacek J., Benešová L. et al. Rapid-resolution hplc with spectrometric detection for the determination and identification of isoflavones in soy preparations and plant extracts // *Analytical and bioanalytical chemistry*. — 2007. — Vol. 389, no. 7-8. — P. 2277–2285.
68. Kamiński M., Kartanowicz R., Kamiński M. M. et al. Hplc-dad in identification and quantification of selected coumarins in crude extracts from plant cultures of ammi majus and ruta graveolens // *Journal of separation science*. — 2003. — Vol. 26, no. 14. — P. 1287–1291.
69. Tarantilis P. A., Tsoupras G., Polissiou M. Determination of saffron (*crocus sativus* L.) components in crude plant extract using high-performance liquid chromatography-uv-visible photodiode-array detection-mass spectrometry // *Jour-*

- nal of Chromatography A. — 1995. — Vol. 699, no. 1-2. — P. 107–118.
70. Springfield E., Eagles P., Scott G. Quality assessment of south african herbal medicines by means of hplc fingerprinting // *Journal of ethnopharmacology*. — 2005. — Vol. 101, no. 1-3. — P. 75–83.
 71. Saucier C., Polidoro A. d. S., dos Santos A. L. et al. Comprehensive two-dimensional gas chromatography with mass spectrometry applied to the analysis of volatiles in artichoke (*cynara scolymus* l.) leaves // *Industrial Crops and Products*. — 2014. — Vol. 62. — P. 507–514.
 72. Ouyang X., Leonards P., Legler J. et al. Comprehensive two-dimensional liquid chromatography coupled to high resolution time of flight mass spectrometry for chemical characterization of sewage treatment plant effluents // *Journal of Chromatography A*. — 2015. — Vol. 1380. — P. 139–145.
 73. van der Horst A., Schoenmakers P. J. Comprehensive two-dimensional liquid chromatography of polymers // *Journal of Chromatography A*. — 2003. — Vol. 1000, no. 1-2. — P. 693–709.
 74. Reich G. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications // *Advanced drug delivery reviews*. — 2005. — Vol. 57, no. 8. — P. 1109–1143.
 75. Massart D. L. et al. *Handbook of chemometrics and qualimetrics*. — Elsevier, 1997.
 76. Du J.-X., Wang X.-F., Zhang G.-J. Leaf shape based plant species recognition // *Applied mathematics and computation*. — 2007. — Vol. 185, no. 2. — P. 883–893.
 77. Zhang S., Wang H., Huang W. Two-stage plant species recognition by local mean clustering and weighted sparse representation classification // *Cluster Computing*. — 2017. — Vol. 20, no. 2. — P. 1517–1525.
 78. Chi Z., Houqiang L., Chao W. Plant species recognition based on bark patterns using novel gabor filter banks // *Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on / IEEE*. — Vol. 2. — 2003. — P. 1035–1038.
 79. Belhumeur P. N., Chen D., Feiner S. et al. Searching the world herbaria: A system for visual identification of plant species // *European Conference on Computer Vision / Springer*. — 2008. — P. 116–129.
 80. Liang Y.-Z., Xie P., Chan K. Quality control of herbal medicines // *Journal of chromatography B*. — 2004. — Vol. 812, no. 1-2. — P. 53–70.
 81. Jiang Y., David B., Tu P., Barbin Y. Recent analytical approaches in quality control

- of traditional chinese medicines-a review // *Analytica chimica acta*. — 2010. — Vol. 657, no. 1. — P. 9–18.
82. Родионова О. Chemometric approach for massive datasets in chemistry // *Российский химический журнал*. — 2006. — Т. 50, № 2. — С. 128–144.
 83. Monakhova Y. B., Holzgrabe U., Diehl B. W. Current role and future perspectives of multivariate (chemometric) methods in nmr spectroscopic analysis of pharmaceutical products // *Journal of pharmaceutical and biomedical analysis*. — 2017. — Vol. 147. — P. 580–589.
 84. Kumar D. Nuclear magnetic resonance (nmr) spectroscopy for metabolic profiling of medicinal plants and their products // *Critical reviews in analytical chemistry*. — 2016. — Vol. 46, no. 5. — P. 400–412.
 85. Christopher M. B. Pattern recognition and machine learning. — Springer-Verlag New York, 2016.
 86. Bridges Jr C. C. Hierarchical cluster analysis // *Psychological reports*. — 1966. — Vol. 18, no. 3. — P. 851–854.
 87. Wold S., Esbensen K., Geladi P. Principal component analysis // *Chemometrics and intelligent laboratory systems*. — 1987. — Vol. 2, no. 1-3. — P. 37–52.
 88. Mimmack G. M., Mason S. J., Galpin J. S. Choice of distance matrices in cluster analysis: Defining regions // *Journal of climate*. — 2001. — Vol. 14, no. 12. — P. 2790–2797.
 89. Bai Y., Wang X., Lei J. et al. Discrimination of fructus forsythiae according to geographical origin with near-infrared spectroscopy // *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on / IEEE*. — 2012. — P. 175–178.
 90. Pan Y., Zhang J., Shen T. et al. Liquid chromatography tandem mass spectrometry combined with fourier transform mid-infrared spectroscopy and chemometrics for comparative analysis of raw and processed gentiana rigescens // *Journal of Liquid Chromatography & Related Technologies*. — 2015. — Vol. 38, no. 14. — P. 1407–1416.
 91. Abdi H., Williams L. J. Principal component analysis // *Wiley interdisciplinary reviews: computational statistics*. — 2010. — Vol. 2, no. 4. — P. 433–459.
 92. Chan C.-O., Chu C.-C., Mok D. K.-W., Chau F.-T. Analysis of berberine and total alkaloid content in cortex phellodendri by near infrared spectroscopy (nirs) compared with high-performance liquid chromatography coupled with ultra-visible spectrometric detection // *Analytica chimica acta*. — 2007. — Vol. 592, no. 2. — P. 121–131.

93. Daolio C., Beltrame F. L., Ferreira A. G. et al. Classification of commercial catuaba samples by nmr, hplc and chemometrics // *Phytochemical analysis*. — 2008. — Vol. 19, no. 3. — P. 218–228.
94. Flores I. S., Silva A. K., Furquim L. C. et al. Hr-mas nmr allied to chemometric on hancornia speciosa varieties differentiation // *Journal of the Brazilian Chemical Society*. — 2018. — Vol. 29, no. 4. — P. 708–714.
95. Li J.-R., Sun S.-Q., Wang X.-X. et al. Differentiation of five species of danggui raw materials by ftir combined with 2d-cos ir // *Journal of Molecular Structure*. — 2014. — Vol. 1069. — P. 229–235.
96. Wang M., Fu J., Guo H. et al. Discrimination of crude and processed rhubarb products using a chemometric approach based on ultra fast liquid chromatography with ion trap/time-of-flight mass spectrometry // *Journal of separation science*. — 2015. — Vol. 38, no. 3. — P. 395–401.
97. Shi X., Wu Y., Lv T. et al. A chemometric-assisted lc–ms/ms method for the simultaneous determination of 17 limonoids from different parts of xylocarpus granatum fruit // *Analytical and bioanalytical chemistry*. — 2017. — Vol. 409, no. 19. — P. 4669–4679.
98. Wang Y., Liu E., Li P. Chemotaxonomic studies of nine paris species from china based on ultra-high performance liquid chromatography tandem mass spectrometry and fourier transform infrared spectroscopy // *Journal of pharmaceutical and biomedical analysis*. — 2017. — Vol. 140. — P. 20–30.
99. Pan Y., Zhang J., Zhao Y.-L. et al. Chemotaxonomic studies of nine gentianaceae species from western china based on liquid chromatography tandem mass spectrometry and fourier transform infrared spectroscopy // *Phytochemical Analysis*. — 2016. — Vol. 27, no. 3-4. — P. 158–167.
100. Nigutová K., Kusari S., Sezgin S. et al. Chemometric evaluation of hypericin and related phytochemicals in 17 in vitro cultured hypericum species, hairy root cultures and hairy root-derived transgenic plants // *Journal of Pharmacy and Pharmacology*. — 2019. — Vol. 71, no. 1. — P. 46–57.
101. Oliveira I., Pinto T., Faria M. et al. Morphometrics and chemometrics as tools for medicinal and aromatic plants characterization // *Journal of Applied Botany and Food Quality*. — 2017. — Vol. 90.
102. Bittner M., Schenk R., Springer A., Melzig M. F. Economical, plain, and rapid authentication of actaea racemosa l.(syn. cimicifuga racemosa, black cohosh) herbal raw material by resilient rp-pda-hplc and chemometric analysis // *Phytochemical Analysis*. — 2016. — Vol. 27, no. 6. — P. 318–325.
103. Zimmermann B., Kohler A. Infrared spectroscopy of pollen identifies plant species

- and genus as well as environmental conditions // PLoS One. — 2014. — Vol. 9, no. 4. — P. e95417.
104. Schulz H., Özkan G., Baranska M. et al. Characterisation of essential oil plants from turkey by ir and raman spectroscopy // *Vibrational Spectroscopy*. — 2005. — Vol. 39, no. 2. — P. 249–256.
 105. Al-Musayeib N., Ebada S. S., Gad H. A. et al. Chemotaxonomic diversity of three ficus species: Their discrimination using chemometric analysis and their role in combating oxidative stress // *Pharmacognosy magazine*. — 2017. — Vol. 13, no. Supl 3. — P. S613.
 106. Fan G., Zhang M. Y., Zhou X. D. et al. Quality evaluation and species differentiation of rhizoma coptidis by using proton nuclear magnetic resonance spectroscopy // *Analytica chimica acta*. — 2012. — Vol. 747. — P. 76–83.
 107. Mesquita P. R., Nunes E. C., dos Santos F. N. et al. Discrimination of eugenia uniflora l. biotypes based on volatile compounds in leaves using hs-spme/gc–ms and chemometric analysis // *Microchemical Journal*. — 2017. — Vol. 130. — P. 79–87.
 108. Yudthavorasit S., Wongravee K., Leepipatpiboon N. Characteristic fingerprint based on gingerol derivative analysis for discrimination of ginger (*zingiber officinale*) according to geographical origin using hplc-dad combined with chemometrics // *Food chemistry*. — 2014. — Vol. 158. — P. 101–111.
 109. Gad H. A., Bouzabata A. Application of chemometrics in quality control of turmeric (*curcuma longa*) based on ultra-violet, fourier transform-infrared and 1h nmr spectroscopy // *Food chemistry*. — 2017. — Vol. 237. — P. 857–864.
 110. Viapiana A., Struck-Lewicka W., Konieczynski P. et al. An approach based on hplc-fingerprint and chemometrics to quality consistency evaluation of *matri-caria chamomilla l.* commercial samples // *Frontiers in plant science*. — 2016. — Vol. 7. — P. 1561.
 111. Chu B.-w., Zhang J., Li Z.-m. et al. Evaluation and quantitative analysis of different growth periods of herb–arbor intercropping systems using hplc and uv–vis methods coupled with chemometrics // *Journal of natural medicines*. — 2016. — Vol. 70, no. 4. — P. 803–810.
 112. Chen N.-D., Chen N.-F., Li J. et al. Rapid authentication of different ages of tissue-cultured and wild *dendrobium huoshanense* as well as wild *dendrobium henanense* using ftir and 2d-cos ir // *Journal of molecular structure*. — 2015. — Vol. 1101. — P. 101–108.
 113. Zaini N. N., Osman R., Juahir H., Saim N. Development of chromatographic fingerprints of *eurycoma longifolia* (tongkat ali) roots using online solid phase

- extraction-liquid chromatography (spe-lc) // *Molecules*. — 2016. — Vol. 21, no. 5. — P. 583.
114. Chen X., Wu D., He Y., Liu S. Nondestructive differentiation of panax species using visible and shortwave near-infrared spectroscopy // *Food and Bioprocess Technology*. — 2011. — Vol. 4, no. 5. — P. 753–761.
 115. Zhu Y., Tan A. T. L. Discrimination of wild-grown and cultivated ganoderma lucidum by fourier transform infrared spectroscopy and chemometric methods // *American Journal of Analytical Chemistry*. — 2015. — Vol. 6. — P. 480–491.
 116. Lever J., Krzywinski M., Altman N. Points of significance: Principal component analysis. — 2017. — Vol. 14. — P. 641–642.
 117. Refaeilzadeh P., Tang L., Liu H. Cross-validation // *Encyclopedia of database systems*. — Springer, 2009. — P. 532–538.
 118. Witten I. H., Frank E., Hall M. A., Pal C. J. *Data Mining: Practical machine learning tools and techniques*. — Morgan Kaufmann, 2016.
 119. Fan Q., Wang Y., Sun P. et al. Discrimination of ephedra plants with diffuse reflectance ft-nirs and multivariate analysis // *Talanta*. — 2010. — Vol. 80, no. 3. — P. 1245–1250.
 120. Chen Y., Xie M.-Y., Yan Y. et al. Discrimination of ganoderma lucidum according to geographical origin with near infrared diffuse reflectance spectroscopy and pattern recognition techniques // *Analytica chimica acta*. — 2008. — Vol. 618, no. 2. — P. 121–130.
 121. Wold S., Sjöström M., Eriksson L. Pls-regression: a basic tool of chemometrics // *Chemometrics and intelligent laboratory systems*. — 2001. — Vol. 58, no. 2. — P. 109–130.
 122. Geladi P., Kowalski B. R. Partial least-squares regression: a tutorial // *Analytica chimica acta*. — 1986. — Vol. 185. — P. 1–17.
 123. Tarachiwin L., Katoh A., Ute K., Fukusaki E. Quality evaluation of angelica acutiloba kitagawa roots by 1h nmr-based metabolic fingerprinting // *Journal of pharmaceutical and biomedical analysis*. — 2008. — Vol. 48, no. 1. — P. 42–48.
 124. Li Y., Zhang J., Zhao Y. et al. Characteristic fingerprint based on low polar constituents for discrimination of wolffiporia extensa according to geographical origin using uv spectroscopy and chemometrics methods // *Journal of analytical methods in chemistry*. — 2014. — Vol. 2014. — P. 519424.
 125. Zhao Y., Zhang J., Jin H. et al. Discrimination of gentiana rigescens from different origins by fourier transform infrared spectroscopy combined with chemometric methods // *Journal of AOAC International*. — 2015. — Vol. 98, no. 1. — P. 22–

26.

126. Nsuala B. N., Kamatou G. P., Sandasi M. et al. Variation in essential oil composition of *leonotis leonurus*, an important medicinal plant in south africa // *Biochemical systematics and ecology*. — 2017. — Vol. 70. — P. 155–161.
127. Hu Y., Kong W., Yang X. et al. Gc–ms combined with chemometric techniques for the quality control and original discrimination of *curcuma longae* rhizome: Analysis of essential oils // *Journal of separation science*. — 2014. — Vol. 37, no. 4. — P. 404–411.
128. Pan Y., Zhang J., Li H. et al. Characteristic fingerprinting based on macamides for discrimination of *maca (lepidium meyenii)* by lc/ms/ms and multivariate statistical analysis // *Journal of the Science of Food and Agriculture*. — 2016. — Vol. 96, no. 13. — P. 4475–4483.
129. Pan Y., Zhang J., Shen T. et al. Comparative metabolic fingerprinting of *gentiana rhodantha* from different geographical origins using lc-uv-ms/ms and multivariate statistical analysis // *BMC biochemistry*. — 2015. — Vol. 16, no. 1. — P. 9.
130. Hoffmann J. F., Carvalho I. R., Barbieri R. L. et al. *Butia* spp.(arecaceae) lc-ms-based metabolomics for species and geographical origin discrimination // *Journal of agricultural and food chemistry*. — 2017. — Vol. 65, no. 2. — P. 523–532.
131. Zheng S., Jiang X., Wu L. et al. Chemical and genetic discrimination of *cistanche herba* based on uplc-qtof/ms and dna barcoding // *PloS one*. — 2014. — Vol. 9, no. 5. — P. e98061.
132. Shevchuk A., Jayasinghe L., Kuhnert N. Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of lc-ms data // *Food Research International*. — 2018. — Vol. 109. — P. 387–402.
133. da Silva G. S., Canuto K. M., Ribeiro P. R. V. et al. Chemical profiling of guarana seeds (*paullinia cupana*) from different geographical origins using uplc-qtof-ms combined with chemometrics // *Food Research International*. — 2017. — Vol. 102. — P. 700–709.
134. Tan T., Zhang J., Xu X. et al. Geographical discrimination of *glechoma herba* based on fifteen phenolic constituents determined by lc–ms/ms method combined with chemometric methods // *Biomedical Chromatography*. — 2018. — Vol. 32, no. 8. — P. e4239.
135. He S., Liu X., Zhang W. et al. Discrimination of the *coptis chinensis* geographic origins with surface enhanced raman scattering spectroscopy // *Chemometrics and Intelligent Laboratory Systems*. — 2015. — Vol. 146. — P. 472–477.

136. Chen C.-w., Yan H., Han B.-x. Rapid identification of three varieties of chrysanthemum with near infrared spectroscopy // *Revista Brasileira de Farmacognosia*. — 2014. — Vol. 24, no. 1. — P. 33–37.
137. Lee B.-J., Kim H.-Y., Lim S. R. et al. Discrimination and prediction of cultivation age and parts of panax ginseng by fourier-transform infrared spectroscopy combined with multivariate statistical analysis // *PloS one*. — 2017. — Vol. 12, no. 10. — P. e0186664.
138. Fu H.-Y., Huang D.-C., Yang T.-M. et al. Rapid recognition of chinese herbal pieces of areca catechu by different concocted processes using fourier transform mid-infrared and near-infrared spectroscopy combined with partial least-squares discriminant analysis // *Chinese Chemical Letters*. — 2013. — Vol. 24, no. 7. — P. 639–642.
139. Wang M., Avula B., Wang Y.-H. et al. An integrated approach utilising chemometrics and gc/ms for classification of chamomile flowers, essential oils and commercial products // *Food chemistry*. — 2014. — Vol. 152. — P. 391–398.
140. Shikanga E. A., Viljoen A. M., Vermaak I., Combrinck S. A novel approach in herbal quality control using hyperspectral imaging: Discriminating between *sceletium tortuosum* and *sceletium crassicaule* // *Phytochemical Analysis*. — 2013. — Vol. 24, no. 6. — P. 550–555.
141. Millán L., Sampedro M. C., Sánchez A. et al. Liquid chromatography–quadrupole time of flight tandem mass spectrometry–based targeted metabolomic study for varietal discrimination of grapes according to plant sterols content // *Journal of Chromatography a*. — 2016. — Vol. 1454. — P. 67–77.
142. Mncwangi N. P., Viljoen A. M., Zhao J. et al. What the devil is in your phyto-medicine? exploring species substitution in harpagophytum through chemometric modeling of 1h-nmr and uhplc-ms datasets // *Phytochemistry*. — 2014. — Vol. 106. — P. 104–115.
143. Mavimbela T., Viljoen A., Vermaak I. Differentiating between *agathosma betulina* and *agathosma crenulata*—a quality control perspective // *Journal of Applied Research on Medicinal and Aromatic Plants*. — 2014. — Vol. 1, no. 1. — P. e8–e14.
144. Liaw A., Wiener M. et al. Classification and regression by randomforest // *R news*. — 2002. — Vol. 2, no. 3. — P. 18–22.
145. de Santana F. B., Mazivila S. J., Gontijo L. C. et al. Rapid discrimination between authentic and adulterated andiroba oil using ftir-hatr spectroscopy and random forest // *Food Analytical Methods*. — 2018. — Vol. 11, no. 7. — P. 1–9.
146. Steinwart I., Christmann A. Support vector machines. — Springer Science & Business Media, 2008.

147. Zheng L., Watson D., Johnston B. et al. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with pca, support vector machines and random forest data modeling // *Analytica Chimica Acta*. — 2009. — Vol. 642, no. 1-2. — P. 257–265.
148. Ni Y., Mei M., Kokot S. One-and two-dimensional gas chromatography–mass spectrometry and high performance liquid chromatography–diode-array detector fingerprints of complex substances: A comparison of classification performance of similar, complex rhizoma curcumae samples with the aid of chemometrics // *Analytica chimica acta*. — 2012. — Vol. 712. — P. 37–44.
149. Yao S., Li T., Liu H. et al. Traceability of boletaceae mushrooms using data fusion of uv–visible and ftir combined with chemometrics methods // *Journal of the Science of Food and Agriculture*. — 2018. — Vol. 98, no. 6. — P. 2215–2222.
150. Dall’Acqua Y. G., Cunha Júnior L. C., Nardini V. et al. Discrimination of euterpe oleracea mart.(açai) and euterpe edulis mart.(juçara) intact fruit using near-infrared (nir) spectroscopy and linear discriminant analysis // *Journal of food processing and preservation*. — 2015. — Vol. 39, no. 6. — P. 2856–2865.
151. Wold S., SJÖSTRÖM M. Simca: a method for analyzing chemical data in terms of similarity and analogy // *Chemometrics: Theory and Application*. — ACS Publications, 1977. — P. 243–282.
152. Wang P., Yu Z. Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review // *Journal of Pharmaceutical Analysis*. — 2015. — Vol. 5, no. 5. — P. 277–284.
153. Li W., Cheng Z., Wang Y., Qu H. Quality control of Ionicerae japonicae flos using near infrared spectroscopy and chemometrics // *Journal of pharmaceutical and biomedical analysis*. — 2013. — Vol. 72. — P. 33–39.
154. Gad H. A., El-Ahmady S. H., Abou-Shoer M. I., Al-Azizi M. M. A modern approach to the authentication and quality assessment of thyme using uv spectroscopy and chemometric analysis // *Phytochemical Analysis*. — 2013. — Vol. 24, no. 6. — P. 520–526.
155. Deconinck E., Aouadi C., Bothy J., Courselle P. Detection and identification of multiple adulterants in plant food supplements using attenuated total reflectance - infrared spectroscopy // *Journal of pharmaceutical and biomedical analysis*. — 2018. — Vol. 152. — P. 111–119.
156. Zhao Y., Zhang J., Yuan T. et al. Discrimination of wild paris based on near infrared spectroscopy and high performance liquid chromatography combined with multivariate analysis // *PloS one*. — 2014. — Vol. 9, no. 2. — P. e89100.
157. Kudo M., Watt R. A., Moffat A. C. Rapid identification of digitalis purpurea us-

- ing near-infrared reflectance spectroscopy // *Journal of pharmacy and pharmacology*. — 2000. — Vol. 52, no. 10. — P. 1271–1277.
158. Zhang Z., Lam T.-N., Zuo Z. Radix puerariae: an overview of its chemistry, pharmacology, pharmacokinetics, and clinical use // *The Journal of Clinical Pharmacology*. — 2013. — Vol. 53, no. 8. — P. 787–811.
 159. Lau C.-C., Chan C.-O., Chau F.-T., Mok D. K.-W. Rapid analysis of radix puerariae by near-infrared spectroscopy // *Journal of Chromatography A*. — 2009. — Vol. 1216, no. 11. — P. 2130–2135.
 160. Wang C., Xiang B., Zhang W. Application of two-dimensional near-infrared (2d-nir) correlation spectroscopy to the discrimination of three species of dendrobium // *Journal of Chemometrics*. — 2009. — Vol. 23, no. 9. — P. 463–470.
 161. Woo Y.-a., Cho C.-h., Kim H.-j. et al. Classification of cultivation area of ginseng by near infrared spectroscopy and icp-aes // *Microchemical Journal*. — 2002. — Vol. 73, no. 3. — P. 299–306.
 162. Woo Y.-A., Kim H.-J., Cho J. Identification of herbal medicines using pattern recognition techniques with near-infrared reflectance spectra // *Microchemical journal*. — 1999. — Vol. 63, no. 1. — P. 61–70.
 163. Wu Y., Zheng Y., Li Q. et al. Study on difference between epidermis, phloem and xylem of radix ginseng with near-infrared and infrared spectroscopy coupled with principal component analysis // *Vibrational Spectroscopy*. — 2011. — Vol. 55, no. 2. — P. 201–206.
 164. Lucio-Gutiérrez J. R., Coello J., MasPOCH S. Application of near infrared spectral fingerprinting and pattern recognition techniques for fast identification of eleutherococcus senticosus // *Food research international*. — 2011. — Vol. 44, no. 2. — P. 557–565.
 165. Cui X., Zhang Z., Ren Y. et al. Quality control of the powder pharmaceutical samples of sulfaguanidine by using nir reflectance spectrometry and temperature-constrained cascade correlation networks // *Talanta*. — 2004. — November. — Vol. 64, no. 4. — P. 943–948.
 166. Kanakis C. D., Petrakis E. A., Kimbaris A. C. et al. Classification of greek mentha pulegium l.(pennyroyal) samples, according to geographical location by fourier transform infrared spectroscopy // *Phytochemical Analysis*. — 2012. — Vol. 23, no. 1. — P. 34–43.
 167. Kramer O. K-nearest neighbors // *Dimensionality Reduction with Unsupervised Nearest Neighbors*. — Springer, 2013. — P. 13–23.
 168. Li C., Yang S.-C., Guo Q.-S. et al. Geographical traceability of marsdenia

- tenacissima by fourier transform infrared spectroscopy and chemometrics // *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. — 2016. — Vol. 152. — P. 391–396.
169. Lee L. C., Liong C.-Y., Jemain A. A. A contemporary review on data preprocessing (dp) practice strategy in atr-ftir spectrum // *Chemometrics and Intelligent Laboratory Systems*. — 2017. — Vol. 163. — P. 64–75.
 170. Kokalj M., Rihtarič M., Kreft S. Commonly applied smoothing of ir spectra showed unappropriate for the identification of plant leaf samples // *Chemometrics and Intelligent Laboratory Systems*. — 2011. — Vol. 108, no. 2. — P. 154–161.
 171. Gudi G., Krahmer A., Kruger H., Schulz H. Attenuated total reflectance–fourier transform infrared spectroscopy on intact dried leaves of sage (*salvia officinalis* l.): Accelerated chemotaxonomic discrimination and analysis of essential oil composition // *Journal of agricultural and food chemistry*. — 2015. — Vol. 63, no. 39. — P. 8743–8750.
 172. Chuang Y.-K., Yang I.-C., Lo Y. M. et al. Integration of independent component analysis with near-infrared spectroscopy for analysis of bioactive components in the medicinal plant *gentiana scabra bunge* // *journal of food and drug analysis*. — 2014. — Vol. 22, no. 3. — P. 336–344.
 173. Hyvärinen A., Karhunen J., Oja E. Independent component analysis. — John Wiley & Sons, 2004. — Vol. 46.
 174. Belščak-Cvitanović A., Valinger D., Benković M. et al. Integrated approach for bioactive quality evaluation of medicinal plant extracts using hplc-dad, spectrophotometric, near infrared spectroscopy and chemometric techniques // *International Journal of Food Properties*. — 2018. — Vol. 20, no. sup3. — P. S2463–S2480.
 175. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine learning in Python // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — P. 2825–2830.
 176. Analytics C. Anaconda software distribution. — Computer software. Vers. 2-2.4.0. — 2015. — Nov. — URL: <https://continuum.io>.
 177. McKinney W. et al. Data structures for statistical computing in python // *Proceedings of the 9th Python in Science Conference / Austin, TX*. — Vol. 445. — 2010. — P. 51–56.
 178. Kluyver T., Ragan-Kelley B., Pérez F. et al. Jupyter notebooks – a publishing format for reproducible computational workflows // *Positioning and Power in Academic Publishing: Players, Agents and Agendas / Ed. by F. Loizides, B. Schmidt ; IOS Press*. — 2016. — P. 87 – 90.

179. Hunter J. D. Matplotlib: A 2D graphics environment // Computing in science & engineering. — 2007. — Vol. 9, no. 3. — P. 90–95.
180. Oliphant T. E. A guide to NumPy. — Trelgol Publishing USA, 2006. — Vol. 1.
181. Nanda V., Sarkar B., Sharma H., Bawa A. Physico-chemical properties and estimation of mineral content in honey produced from different plants in northern india // Journal of Food Composition and Analysis. — 2003. — Vol. 16, no. 5. — P. 613–619.
182. Bishop C. M. Pattern Recognition and Machine Learning (Information Science and Statistics). — 2006.
183. Ng A. Machine learning. stanford course. — 2015. — URL: <http://cs229.stanford.edu/materials.html>.
184. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. — 2nd edition. — New York : Springer, 2009.
185. Breiman L. Random forests // Mach. Learn. — 2001. — Vol. 45, no. 1. — P. 5–32.
186. Netzer Y., Wang T., Coates A. et al. Reading digits in natural images with unsupervised feature learning // NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011. — 2011. — URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.
187. Deming S., Michotte Y., Massart D. L. et al. Chemometrics: a textbook. — Elsevier, 1988. — Vol. 2.
188. Deng X., Geng H., Ali H. H. Cross-platform analysis of cancer biomarkers: a Bayesian network approach to incorporating mass spectrometry and microarray data // Cancer informatics. — 2007. — Vol. 3. — P. 117693510700300001.
189. Yu J., Chen X.-W. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data // Bioinformatics. — 2005. — Vol. 21, no. suppl_1. — P. i487–i494.
190. Lukman S., He Y., Hui S.-C. Computational methods for traditional Chinese medicine: a survey // Computer methods and programs in biomedicine. — 2007. — Vol. 88, no. 3. — P. 283–294.
191. Young J., Graham P., Penny R. Using Bayesian networks to create synthetic data // Journal of Official Statistics. — 2009. — Vol. 25, no. 4. — P. 549.
192. The plant list project. — <https://theplantlist.org>.
193. Frolov E., Oseledets I. Fifty shades of ratings: How to benefit from a negative feed-

- back in top-n recommendations tasks // Proceedings of the 10th ACM Conference on Recommender Systems / ACM. — 2016. — P. 91–98.
194. Siegel D., Permentier H., Reijngoud D.-J., Bischoff R. Chemical and technical challenges in the analysis of central carbon metabolites by liquid-chromatography mass spectrometry // *Journal of Chromatography B*. — 2014. — Vol. 966. — P. 21–33.
 195. Scutari M. Learning Bayesian networks with the bnlearn R package // *Journal of Statistical Software*. — Vol. 35, no. 3. — P. 1–22.
 196. Chow C., Liu C. Approximating discrete probability distributions with dependence trees // *IEEE Trans. Inf. Theory*. — 1968. — Vol. 14, no. 3. — P. 462–467.
 197. Schreiber J. Pomegranate: fast and flexible probabilistic modeling in Python // *Journal of Machine Learning Research*. — 2018. — Vol. 18, no. 164. — P. 1–6.
 198. Hagberg A., S Chult D., Swart P. Exploring network structure, dynamics, and function using NetworkX // *Proceedings of the 7th Python in Science Conference*. — 2008. — P. 11–15.
 199. Kingma D. P., Ba J. Adam: a method for stochastic optimization // *arXiv preprint arXiv:1412.6980*. — 2014.
 200. PyTorch: Tensors and dynamic neural networks in Python. — Computer software. Vers. 0.3.1, <http://pytorch.org/>.
 201. Kolda T. G., Bader B. W. Tensor decompositions and applications // *SIAM review*. — 2009. — Vol. 51, no. 3. — P. 455–500.
 202. Zhou G., Cichocki A., Zhao Q., Xie S. Efficient nonnegative Tucker decompositions: algorithms and uniqueness // *IEEE Trans. Image Process*. — 2015. — Vol. 24, no. 12. — P. 4990–5003.
 203. Xu Y. Alternating proximal gradient method for sparse nonnegative Tucker decomposition // *Mathematical Programming Computation*. — 2015. — Vol. 7, no. 1. — P. 39–70.
 204. Bjorck A., Golub G. H. Numerical methods for computing angles between linear subspaces // *Mathematics of computation*. — 1973. — Vol. 27, no. 123. — P. 579–594.
 205. Zhou G., Cichocki A., Zhang Y., Mandic D. P. Group component analysis for multiblock data: common and individual feature extraction // *IEEE Trans. Neural Netw. Learn. Syst*. — 2016. — Vol. 27, no. 11. — P. 2426–2439.
 206. Székely G. J., Rizzo M. L., Bakirov N. K. Measuring and testing dependence by correlation of distances // *The annals of statistics*. — 2007. — P. 2769–2794.

207. Hay A. The derivation of global estimates from a confusion matrix // International Journal of Remote Sensing. — 1988. — Vol. 9, no. 8. — P. 1395–1398.
208. Phylot: a phylogenetic tree generator, based on NCBI taxonomy. — <https://phylot.biobyte.de/>.