Оценка соответствия. Аккредитация лабораторий

Compliance verification. Laboratory accreditation

DOI: 10.26896/1028-6861-2018-84-10-67-78

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ В АНАЛИТИЧЕСКОМ КОНТРОЛЕ ПРЕПАРАТОВ ЛЕКАРСТВЕННЫХ РАСТЕНИЙ

© **Дмитрий Владимирович Назаренко, Игорь Александрович Родин, Олег Алексеевич Шпигун**

Московский государственный университет имени М. В. Ломоносова, Москва, Россия;
e-mail: dmitro.nazarenko@gmail.com

Статья поступила 5 июня 2018 г.

Несмотря на то что объем мирового рынка лекарственных растений составляет сотни миллиардов долларов, государственный контроль за качеством подобных препаратов в большинстве стран мира практически отсутствует. Отчасти это объясняется сложным составом растительного сырья: традиционная аналитическая методология основана на применении стандартных образцов сравнения для каждого определяемого вещества. При этом препараты на основе лекарственных растений могут содержать десятки и сотни физиологически активных компонентов. Выделение данных соединений в чистом виде на практике осуществляют с помощью препаративной хроматографии, что приводит к их высокой стоимости. Более того, варьирование химического состава растительных препаратов в зависимости от географического происхождения сырья делает малореальным установление строгих диапазонов допустимых содержаний для всех физиологически активных компонентов. Совокупность вышеперечисленных факторов ограничивает возможности использования традиционных подходов к анализу, требующих строгой стандартизации, списка соединений для каждого типа растения, уровней содержаний и наличия стандартных образцов сравнения. Это привело к исследованию возможностей внедрения различных математических подходов как вспомогательной методологии. В отличие от традиционной методологии, подходы с использованием машинного обучения основаны на правильном сборе выборок данных. В такой выборке должны присутствовать группы образцов, отвечающие состояниям объекта, которые должен будет различить разрабатываемый алгоритм: аутентичный/поддельный, чистый/содержащий примеси, действенный/не содержащий определенного уровня активных компонентов и т.д. Данный обзор посвящен рассмотрению приложения машинного обучения к задачам химического анализа и производственного контроля сырья лекарственных растений и препаратов на его основе за последние 15 лет.

Ключевые слова: машинное обучение; лекарственные растения; производственный и технологический контроль; фармакология; аналитическая химия.

THE USE OF MACHINE LEARNING IN THE ANALYTICAL CONTROL OF THE PREPARATIONS OF MEDICINAL PLANTS

© **Dmitry V. Nazarenko, Igor A. Rodin, Oleg A. Shpigun**

Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia;
e-mail: dmitro.nazarenko@gmail.com

Submitted June 5, 2018.

Despite the fact that the global market for medicinal plants amounts to hundreds of billions of dollars, there is almost no government control over the quality of such pharmaceuticals in most countries of the world. This is partly attributed to the complex composition of plant materials: traditional analytical

methodology is based on the use of standard reference samples for each analyte. In this case, preparations based on medicinal plants may contain tens and hundreds of physiologically active components. Isolation of those compounds in a pure form in practice is carried out using preparative chromatography, which leads to their high cost. Moreover, varying of the chemical composition of the medicinal plants depending on the geographical origin of the raw materials interfere with prescribing strict ranges of permissible contents for all physiologically active components. Combination of the above factors limits the possibilities of using traditional approaches to analysis, requiring strict standardization, the list of compounds for each type of plant, levels of contents and the availability of the reference materials and standards of comparison. This led to the study of the possibility of introducing various mathematical approaches as an auxiliary methodology. Unlike traditional methodologies, machine learning approaches are based on the correct collection of the data samples. Such a sample should contain groups of the samples that correspond to the states of the object which the developed algorithm must distinguish: authentic/fake, pure/containing impurities, effective/not containing a certain level of active components, etc. This review is devoted to consideration of the application of machine learning technique to the problems of chemical analysis and production control of raw materials of medicinal plants and preparations on their base for the last 15 years.

Keywords: machine learning; medicinal plants; production and technological control; pharmacology; analytical chemistry.

Лекарственные растения представляют практически неисчерпаемый источник для поиска новых физиологически активных веществ, пригодных для медицинского использования [1, 2]. По разным оценкам доля новых лекарственных препаратов, в той или иной мере основанных на соединениях растительного происхождения, составляет от 20 до 60 %. Тем не менее большинство современных фармацевтических препаратов содержит в своем составе только 1 – 2 индивидуальных действующих вещества, по которым были проведены регламентированные клинические испытания. Контроль качества подобных препаратов почти всегда представляет собой определение действующего вещества в лекарственной форме методом внешнего стандарта, а также контроль небольшого числа известных побочных продуктов или примесей. По сравнению с этим контроль качества лекарственных препаратов на основе частей растений или растительных экстрактов, содержащих от десятков до сотен соединений, многие из которых обладают той или иной физиологической активностью, представляет собой гораздо более сложную задачу [3 – 6]. Во-первых, одновременное определение большого числа соединений требует использования большого числа дорогостоящих стандартных образцов, во-вторых, нет четких критериев того, какие уровни содержаний тех или иных соединений в растениях устанавливать как допустимые [7]. В дополнение к этому необходимо принимать в расчет трудность проведения и однозначной интерпретации результатов клинических исследований для сложных по составу растительных препаратов [8].

Вышеперечисленные сложности 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 том что объем их 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 лекарственных средств [13, 14]. Весьма выгодно в этом направлении выделяются работы китайских ученых, так как фармакология Китая стремится интегрировать традиционные препараты на основе лекарственных растений в русло современной медицины [15]. Поэтому несмотря на препятствия и трудности процесса интеграции растительных препаратов в систему научной фармакологии, в настоящее время ведется множество исследований, посвященных как клиническим испытаниям, так и разработке методик и подходов для контроля качества лекарственного растительного сырья и препаратов на его основе [16 – 18].

Одним из наиболее популярных направлений в данной области является совместное использование методов аналитической химии и машинного обучения для создания эффективных и недорогих методов контроля качества препаратов сложного состава [19 – 23].

В настоящем обзоре рассмотрены работы в данном направлении, опубликованные за 15-летний период, особое внимание уделено исследованиям за последние пять лет.

Обучение без учителя

Машинное обучение представляет собой раздел науки на стыке математики и программирования, который охватывает создание и применение алгоритмов, способных решать различные задачи без четких заранее прописанных инструк-

ций [24]. Обучение без учителя — раздел машинного обучения, посвященный анализу данных, для 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, поделить выборку на группы 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) [25] и метод главных компонент (*PCA*, Principal Component Analysis) [26]. Несмотря на то что на практике использование методов обучения без учителя в основном ограничено предварительными этапами анализа структуры выборки, имеет смысл кратко рассмотреть их приложение в исследованиях.

PCA и *HCA*

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

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

2. Находят наиболее близкую пару кластеров и объединяют их в один. Таким образом, остается $N - 1$ 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.

В итоге данной процедуры получают иерархическое дерево, отражающее близость объектов выборки друг к другу. Его анализ позволяет выявлять группы похожих между собой и отличных от остальных объектов. Так как критерием близости кластеров в процессе объединения является расстояние, выбор способа его измерения составляет важный аспект применения кластеризации. Не существует универсальных критериев и рекомендаций по выбору метрик расстояния, однако чаще всего применяют Евклидово расстояние или расстояние Махаланобиса [27].

Типичный пример использования *HCA* приведен в работе [28], где иерархический кластерный анализ применяли к данным химического анализа корней женьшеня методами молекулярной спектроскопии. Спектры в ближней инфракрасной области (БИК) в диапазоне 12 000 – 4000 $см^{-1}$, ИК-спектры диффузного отражения в диапазоне 4000 – 400 $см^{-1}$ и спектры комбинационного рассеяния в диапазоне 3700 – 100 $см^{-1}$ использовали для дифференциации трех типов женьшеня (*Radix ginseng*, *Radix ginseng rubra* и *Radix panacis quinquefolii*) и двух типов псевдоженьшеня (*Radix codonopsis* и *Radix platycodi*). В результате применения кластеризации ко вторым производным спектров данные были разделены на четыре кластера: три из них соответствовали трем типам женьшеня, в четвертый кластер попали образцы псевдоженьшеней. Простота, скорость и неразрушающий анализ позволили авторам сделать вывод о конкурентоспособности их методологии по отношению к традиционным и хроматографическим методам анализа. Аналогичный подход с использованием ИК-спектроскопии в диапазоне 4000 – 400 $см^{-1}$ применяли для дифференциации образцов листьев растений трех различных семейств: Ranunculaceae, Plumbaginaceae и Leguminoseae [5]. Листья растений измельчали с использованием жидкого азота и прессовали с KBr. Благодаря различиям в липидном метаболизме и содержании 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 специального проведения кластеризации выборок образцов различного географического происхождения. Отличные друг от друга климатические условия произрастания приводят к значительным отличиям в химическом составе растений и, соответственно, определяют их различную пригодность как сырья для производства лекарственных препаратов. Этим обусловлено значительное количество работ по применению машинного обучения для установления происхождения сырья. Так, в работе [29] проводили неразрушающий анализ плодов форзиции с использованием БИК-спектроскопии: 133 образца плодов, собранных в трех провинциях Китая, подвергли иерархической кластеризации с использованием вторых производных ИК-спектров и сглаживания Норриса. Несмотря на то что большая часть спектров была корректно разнесена по трем кластерам, часть выборки оказалась не в «своих» кластерах. В работе [30] *HCA* использовали для рассмотрения различий в химическом составе образцов черного перца, а в работе [31] — для сравнения образцов *Gentiana rigescens*, подвергнутых различной технологической обработке. Приведенные примеры иллюстрируют достоинства

кластерного анализа: с его помощью можно быстро и достоверно оценить, насколько тот или иной метод анализа позволяет разделить различные объекты 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 традиционно применяется как метод понижения размерности данных [32]. *PCA* отчасти схож с кластеризацией, так как позволяет обнаружить закономерности и группировки в данных без какой-либо предварительной информации о состояниях объектов в выборке. С математической точки зрения *PCA* представляет собой проекцию пространства данных в новое пространство меньшей размерности, обладающее определенным свойствами.

В качестве примера использования *PCA* для видовой дискриминации можно привести работу [33], где с помощью 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 посредством *PCA*. Выбранный диапазон позволил получить полностью корректное разделение образцов по двум классам. В работе [34] описано применение данных ЯМР-спектроскопии и ВЭЖХ-УФ для видовой дискриминации афродизиаков из Бразилии. ЯМР-спектроскопию под «магическим» углом использовали для анализа различий четырех разновидностей *Hancornia speciosa* методами *PCA* и *HCA* [35]. В аналогичном исследовании *PCA* использовали для дифференциации 26 образцов корней пяти видов растений рода *Angelica* [36]. При схожих названиях данные растения обладают различным составом и характеризуются разными физиологическими эффектами при приеме, поэтому их корректная дифференциация — важная аналитическая задача. С использованием фильтра Савцикого – Голея получали вторые производные спектров для последующего *PCA*. Проекция *PCA* на двумерную плоскость показала наличие пяти явных кластеров образцов. Наибольший вклад в нагрузки *PCA* внес диапазон 1600 – 900 $см^{-1}$. В работе [37] *PCA* использовали для идентификации препаратов «сырого» и обработанного различными способами *Rheum rhabarbarum* на основе данных ВЭЖХ-МС высокого разрешения. Проводили также *PCA* данных ВЭЖХ-МС определения 17 лимоноидов для дискриминации различных частей плода *Xylocarpus granatum* [38].

Популярно также использование методов обучения без учителя для хемотаксономических исследований, в том числе, для подтверждения данных генетического анализа. Так, в работе [39] с использованием ИК-спектроскопии и ультравысокоэффективной жидкостной хроматографии с масс-спектрометрическим детектированием (УВЭЖХ-МС) проводили хемотаксонономический анализ 70 образцов, принадлежащих 9 видам рода *Paris*. Применение *PCA* к данным ИК-спектроскопии позволило увидеть явное разбиение на три группы, состоящие из пяти, трех и одного вида, что было подтверждено после рассмотрения химического состава экстрактов по данным хромато-масс-спектрометрического анализа. *PCA* также показал наличие явных различий внутри группы из трех видов, что позволило различить их на основе только ИК-спектроскопических данных. Хемотаксономическое исследование 9 видов семейства Gentianaceae с приложением *PCA* к данным ВЭЖХ-МС и ИК-спектроскопического анализа описано в работе [40]. В работе [41] *PCA* использовали для поиска корреляций в содержании 8 вторичных метаболитов между образцами 17 видов *Hipericum*, а в работе [42] — для рассмотрения различий между пятью видами широко используемых лекарственных ароматических растений. Работа [43] была посвящена видовой идентификации *Actaea racemosa* на основе *PCA* как способа определения поддельных препаратов.

Интересно обратиться также к работе [44], где авторы использовали ИК-спектроскопию нарушенного полного внутреннего отражения и классический вариант ИК-спектроскопии пропускания для анализа 813 образцов цветочной пыльцы, принадлежащей 300 видам растений. Для анализа полученного массива данных был применен как иерархический кластерный анализ, так и *PCA*. Оба метода использовали для анализа вариаций содержания питательных веществ в пыльце (в особенности — триглицеридов и белков), рассмотрения различий стратегий опыления, межвидовых и межклассовых различий. По итогам анализа данных с использованием *PCA* и *HCA* был сделан вывод о необходимости включать также мужскую пыльцу в состав выборок для экологических и эволюционных исследований. *PCA* и *HCA* также использовали совместно для видовой дискриминации летучих масел различных лекарственных растений, произрастающих в Турции [45], видов *Ficus* [46] и *Coptidis* [47], подвидов *Eugenia uniflora* L. [48],

а также для дискриминации образцов имбиря по региону произрастания [49]. В работе [50] *PCA* и *HCA* применяли для оценки качества 30 образцов *Curcuma longa* на основе данных УФ-, ИК- и ЯМР-спектроскопии, а в работе [51] — для оценки корреляции антиоксидантной активности и данных ВЭЖХ-УФ анализа образцов *Matricaria chamomilla*. По идентичной схеме *PCA* также проводили для определения возраста образцов *Gentiana rigescens* [52] и образцов рода *Dendrobium* [53]. Наряду с этим можно отметить использование *PCA* в процессе разработки экстракционного метода для получения репрезентативного профиля растений вида *Eurycoma longifolia* [54].

Таким образом, в хемометрике методы обучения без учителя оказываются наиболее полезны для поиска различных закономерностей в массивах данных. Подобные стратегии часто применяются на предварительных стадиях для исследования какого-либо класса объектов и данных. Свойство *PCA* понижать размерность данных часто оказывается удобным для проекции многомерных данных на двумерную или трехмерную поверхность и получения наглядной визуализации объектов внутри исследуемой выборки. Более того, *PCA* иногда служит удобной отправной точкой для дальнейшего использования методов обучения с учителем на данных меньшей размерности [55 – 57].

Безусловно, подобные подходы к анализу выборок и массивов данных не лишены некоторых ограничений. Так, например, *PCA* для матрицы данных проводится таким образом, чтобы отобразить как можно бóльшую 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 информации, необходимой для корректной дифференциации объектов внутри выборки [58]. Еще один важный нюанс 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 предложенного подхода для дискриминации изучаемых групп объектов. Этот 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.

Обучение с учителем

Данный раздел машинного обучения включает алгоритмы, которым для решения какой-либо задачи на вход подается заранее размеченная выборка данных. Задачей алгоритма при 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 подают 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 [59]. Алгоритм далее обучается по предоставленной обучающей выборке за счет выявления и обобщения внутренней структуры данных и нахождения ее 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 [59, 60].

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 какому-либо критерию по результатам его 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 традиционно используют в случае спектральных данных, например, данных ИК-спектроскопии и спектроскопии комбинационного рассеяния.

Задача классификации

В качестве простого примера можно привести работу по идентификации трех видов рода *Ephedra* на основе данных БИК-спектроскопического

анализа измельченного порошка цельных растений [61]. После предобработки полученного массива ИК-спектры 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 для решения данной проблемы. Для *LDA* эффективность рассчитанных моделей колебалась в диапазоне 84 – 92 %, для двух оставшихся методов она была чуть более низкой. В данном примере можно увидеть, что несмотря на относительную простоту вычислительной задачи (малое число классов), авторам не удалось получить высоких показателей правильности идентификации ни для одного из примененных методов. Такая ситуация с высокой вероятностью указывает на то, что подаваемые на вход алгоритмов данные не содержали достаточное для однозначного определения видовой принадлежности количество химической информации. При условии, что авторы не «потеряли» важную информацию в процессе предобработки, можно заключить, что использованный метод анализа непригоден для эффективного решения поставленной задачи. Более успешным применением схожего подхода можно считать работу [62], 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 первые и вторые производные. Применив дискриминантный анализ на основе метода проекции на латентные структуры (*PLS-DA*, Partial Least Square — Discriminant Analysis) [63, 64] и *LDA*, получили значения эффективности на тестовой выборке 100 и 96,6 % соответственно. Вообще дискриминация образцов по признаку географического происхождения является одним из традиционных применений *PLS-DA* в хемометрике. Так, его использовали для классификации образцов, проанализированных методами ЯМР-спектроскопии [65], УФ- [66] и ИК-спектроскопии [67], ГХ-МС [68, 69] и ВЭЖХ-МС [70 – 76], а также спектроскопии комбинационного рассеяния [77]. В другой работе *PLS-DA* и *LDA* применяли уже для дискриминации образцов культивированного и дикого *Ganoderma lucidum* [57]. Аналогичным образом *PLS-DA* проводили для видовой идентификации растений рода *Chrysantemum* [78]. Данные БИК-спектроскопии 139 образцов (92 для обучающей выборки и 47 для тестовой) трех разных видов использовали для построения классификационного алгоритма. В работе [79] PLS использовали для дифференциации образцов *Panax ginseng* пяти- и шестилетнего возраста на момент сбора, а в работе [80] — для дифференциации образцов *Areca catechu*, экстрагированных разными способами.

Идентификационные алгоритмы на основе *PLS-DA* также использовали для дискриминации видов родов *Chamomile* [81], *Sceletium* [82] и сортов винограда [83]. Совместные данные ВЭЖХ-МС и ЯМР ^{1}H анализа применяли для определения близкородственных примесных видов в сырье *Harpagophytum procumbens* [84]. Тонкослойную хроматографию в комбинации с *PLS-DA* применяли для дифференциации видов *Agathosma betulina* и *Agathosma crenulata* [85]. Случайный лес (*RF*, Random Forest) [86] и *PLS-DA* использовали для дифференциации настоящих и поддельных образцов масла амазонского растения *Carapa guianensis* методом ИК-спектроскопии [87]. Интересно привести также работу [4], где различные внешние условия для растений симулировались самими исследователями. В качестве данных использовали масс-спектры прямого ввода экстрактов листьев *Pharbitis nil*. Рост растений происходил при различной продолжительности светового дня (шесть вариантов). 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 удалось успешно построить алгоритм, способный устойчиво различать образцы, выращенные при наибольшей продолжительности (одна неделя против двух дней у ближайшей группы) 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, причем в том случае, когда метод главных компонент и кластеризация не позволяли увидеть какой-либо «полезной» структуры данных.

Еще один популярный метод классификации объектов — метод опорных векторов (*SVM*, Support Vector Machine) [88]. 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 [89] для идентификации 30 экстрактов шести сортов чая (по пять образцов каждого сорта). Отличительной особенностью данной работы является использование ВЭЖХ-УФ анализа для получения выборки химических данных. Данными для классификации служила 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 этих различий оказалась недостаточной для надежного решения поставленной задачи. После применения *SVM* правильность идентификации по некоторым сортам составила не более 80 %. Лучше всего показал себя алгоритм *RF*, при использовании которого правильность идентификаций всех сортов была максимальной.

Более эффективно использовать *SVM* удалось при исследовании хроматографических профилей трех видов корней рода *Curcuma* методами одномерной и двумерной газовой хроматографии, а также ВЭЖХ [90]. 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 объединенных данных ГХ и ВЭЖХ *SVM* показал 100 %-ную правильность предсказания на тестовой выборке. Данная работа иллюстрирует возможность создания высокоэффективного алгоритма при объединении данных различных методов анализа, содержащих комплементарную химическую информацию, даже если индивидуальная эффективность каждого из методов не очень высока.

Интересно отметить работу [56] по дискриминации трех сортов *Panax ginseng* на основе спектрофотомерии и БИК-спектроскопии (376 – 1025 нм). 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 %-ную правильность идентификации. С высокой эффективностью *SVM* и *PLS-DA* использовали также для видовой и географической дискриминации грибов семейства Boletaceae, комбинируя данные спектрофотомерии и БИК-спектроскопии [91]. С применением неразрушающего анализа (БИК-спектроскопия) и *LDA* была разработана экспрессная методика дифференциации плодов подвидов *Euterpe oleracea* [92]. Правильность идентификации составила 93,2 % на тестовой выборке. Еще один популярный в хемометрическом анализе метод — *SIMCA* (Soft Independent Modeling of Class Analogy) [93], который 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 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 [94].

Так, например, *SIMCA* использовали для классификации 140 образцов *Lonicera japonica*, собранных в семи разных провинциях Китая [95]. Для всех образцов были получены БИК-спектры в диапазоне 10 000 – 4000 $см^{-1}$ с шагом 4 $см^{-1}$ (1500 точек). Построенный на основе *SIMCA* классификатор обладал 100 %-ной точностью предсказаний в пределах тестовой выборки. В работе [96] *SIMCA* совместно с УФ-спектроскопией использовали для видовой дискриминации 50 образцов рода *Thymus*. Использовали *SIMCA* и *PLS-DA* данных спектроскопии нарушенного полного внутреннего отражения для определения пяти видов лекарственных растений в порошках [97].

Интересно отметить также использование нестандартных математических методов. В работе [98] применили температурно-ограниченные сети каскадных корреляций (*TCCCNs*, Temperature-Constrained Cascade Correlation Networks) для решения бинарной задачи классификации в анализе образцов растений рода *Rheum* методом БИК-спектроскопии. Из 52 образцов 25 относились к видам, признаваемым фармакопеей Китая как официальные, и 27 — к «неофициальным» видам. Несмотря на сложность задачи дифференциации групп видов, метод *TCCCNs* позволил добиться 100 %-й правильности идентификации на тестовой выборке, обойдя при этом метод искусственных нейронных сетей с обратным распространением ошибки.

Относительно редко в исследованиях применяют метод *k*-ближайших соседей (*k-NN*, *k*-Nearest Neighbors) [99], который относится к одним из наиболее простых методов обучения с учителем. Метод *k-NN* не включает никакой предварительной оптимизации модели по обучающей выборке, все расчеты проводятся уже на этапе классификации неизвестных объектов. Рассчитывается расстояние от неизвестного объекта (векто-

ра данных) до всех объектов выборки, и класс объекта определяется голосованием *k*-ближайших соседей. В работе [100] на основе данных тонкослойной хроматографии (ТСХ) и ВЭЖХ-УФ проводили классификацию 31 неизвестного образца лекарственного препарата из корней растений рода *Bupleurum*. Только два вида данного рода одобрены в фармакопее Китая для изготовления препарата. При этом существует еще пять близкородственных видов *Bupleurum*, которые периодически встречаются в поддельных препаратах. На основе размеченной выборки из 33 образцов сертифицированных 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 %. В случае ВЭЖХ-УФ оба метода показали 100 %-ю правильность идентификации. С высокой эффективностью метод *k-NN* применили для дискриминации географического происхождения 128 образцов *Marsdenia tenacissima* на основе данных ИК-спектроскопии [101]. Канонический дискриминантный анализ применяли для классификации образцов *Mentha pulegium* по признаку места произрастания на основе ИК-спектроскопии в диапазоне 4000 – 400 $см^{-1}$ [102]. Правильность идентификации составила 90 % на тестовой выборке.

Остается не до конца ясным вопрос корректной предобработки хемометрических данных, так как это может приводить к невысокой эффективности конечных алгоритмов [103]. Интересным шагом в данном направлении можно считать работу [104], где авторы исследовали различные варианты обработки ИК-данных на примере 12 видов лекарственных растений — 60 образцов шести видов рода *Hypericum* и 40 образцов шести 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, так как построение градуировочной зависимости методом внешнего стандарта обычно проводят с использованием регрессии по методу наименьших квадратов. Однако использование прямых градуировочных зависимостей малоприменимо 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 работе [102] проводили ИК-спектроскопическое определение 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 методом газовой хроматографии. Прямое определение валового содержания эфирных масел и основных действующих веществ (α-туйон и β-туйон) в *Salvia officinalis* проводили в работе [105]. Метод основан на *PLS* и не требовал никакой пробоподготовки, кроме высушивания листьев, а длительность анализа составила не более 5 мин, тогда как рутинный анализ методом ГХ-МС с дериватизацией занимает несколько часов. Более необычный вариант использования регрессии на основе *PLS* описан в работе [65]. Регрессию применяли для установления зависимости между данными ЯМР ^1H 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. использовали для установления принадлежности образца к одной из пяти категорий качества. 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 работе [106] метод независимых компонент (*ICA*, Independent Component Analysis) [107] использовали для определения уровней содержания гентиопикрозида и свертиамарина в лекарственном растении вида *Gentiana scabra*. На основе производных БИК-спектров авторам удалось получить коэффициенты корреляции 0,85 и 0,95 соответственно. Авторам работы [95] удалось получить устойчивый (переносимый на разные ИК-спектрометры) алгоритм для полуколичественной оценки содержания шести основных активных компонентов лекарственного растения *Lonicera japonica*. Аналогичным образом *PLS* регрессию использовали в работе [33] для определения валового содержания алкалоидов в *Cortex phellodendri* методом БИК-спектроскопии. Более масштабное исследование было проведено в работе [108], 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.

Таким образом, процесс внедрения препаратов сложного состава в практику современной фармакологии будет неуклонно вести ко все более широкому применению методов машинного обучения при решении задач контроля качества. Проведение количественного анализа для определения каждого целевого соединения будет дополняться новыми гибридными подходами, основанными на обработке «сырых» химических данных методами машинного обучения. Данная тенденция в аналитической химии связана с развитием и расширением применения аналитических методов с получением больших объемов данных (масс-спектрометрия высокого разрешения, ЯМР-спектроскопия и т.д.), для анализа сложных объектов природного и техногенного происхождения (образцы биологических тканей, сточных вод, продукты питания и т.п.).

Методы машинного обучения позволяют находить в данных большого объема информацию, способную помочь ответить на различные вопросы: например, различить состояния чистый продукт/продукт с примесью, аутентичный продукт/подделка, качественное/некачественное сырье и т.п. Такой постепенный переход будет продолжаться еще многие годы. Дальнейшие исследования в этой области и накапливаемый опыт применения позволят сказать, насколько надежной альтернативой подобные подходы станут по отношению к классической методологии химического анализа.

Совмещение аналитической химии с машинным обучением требует плотного сотрудничества специалистов-химиков со специалистами по машинному обучению, так как эффективные и корректные решения могут быть найдены только с пониманием принципиальной структуры и особенностей обрабатываемых данных. Схема исследований за последние 15 лет практически не претерпела значимых изменений. В большинстве случаев авторам удается 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*/*HCA*, на основе чего делают выводы о сходствах или различиях имеющихся групп либо 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, это вариант линейного *DA* или *PLS-DA*) для построения классификационного или регрессионного алгоритма. Несмотря на большой опыт, накопленный научным сообществом в данной области, практическое применение подобных подходов пока не получило распространения. Препараты лекарственных растений в большинстве стран мира по-прежнему не подвергаются практически никакому аналитическому контролю. Помимо небольшого размера выборок, еще одной проблемой является отсутствие практики у авторов исследований выкладывать в открытый доступ первичные данные, полученные непосредственно после физико-химического анализа. Такая ситуация значительно осложняет сравнение 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.

REFERENCES (ЛИТЕРАТУРА)

1. **Williams P.** Health benefits of herbs and spices: Public health / M. J. Australia. 2006. Vol. 4. N 4. P. S17 – S18.
2. **Hostettmann K., Marston A.** Twenty years of research into medicinal plants: Results and perspectives / Phytochem. Rev. 2002. Vol. 1. N 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 Anal. Chem. 2008. Vol. 27. N 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 / Phytochem. 2003. Vol. 62. N 6. P. 859 – 863.
5. **Gorgulu S. T., Dogan M., Severcan F.** The characterization and differentiation of higher plants by Fourier transform infrared spectroscopy / Appl. Spectrosc. 2007. Vol. 61. N 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 / J. Chromatogr. A. 2006. Vol. 1112. N 1 – 2. P. 241 – 254.
7. **Folashade O., Omoregie H., Ochogu P.** Standardization of herbal medicines-a review / Int. J. Biodiv. Conserv. 2012. Vol. 4. N 3. P. 101 – 112.
8. **Dahanukar S., Kulkarni R., Rege N.** Pharmacology of medicinal plants and natural products / Indian J. Pharmacol. 2000. Vol. 32. N 4. P. S81 – S118.
9. European Parliament and of the Council Directive 2004/24/ec; 2004. http://eur-lex.europa.eu/legal-content/EN/TXT/? uri=CELEX:32004L0024&qid=1451884773824 (accessed June 5, 2018).
10. Food and Drug Administration Dietary supplements; http://www.fda.gov/Food/DietarySupplements (accessed June 5, 2018).
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. N 4. P. 262 – 268.
12. **Chaudhury R. R.** Herbal remedies and traditional medicines in reproductive health care practices and their clinical evaluation / J. Reproductive Health and Medicine. 2015. Vol. 1. N 1. P. 44 – 46.
13. **Petrovska B. B.** Historical review of medicinal plants' usage / Pharmacognosy Rev. 2012. Vol. 6. N 11. P. 1.

14. **Maroyi A.** Traditional use of medicinal plants in south-central Zimbabwe: Review and perspectives / J. Ethnobiol. Ethnomed. 2013. Vol. 9. N 1. P. 31.
15. **Wang M.-W., Richard D. Y., Zhu Y.** Pharmacology in China: A brief overview / Trends Pharmacol. Sci. 2013. Vol. 34. N 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 Anal. Chem. 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 Biotechnol. 2014. Vol. 25. P. 51 – 59.
18. **Bansal A., Chhabra V., Rawal R. K., Sharma S.** Chemometrics: A new scenario in herbal drug standardization / J. Pharm. Anal. 2014. Vol. 4. N 4. P. 223 – 233.
19. **Liang Y.-Z., Xie P., Chan K.** Quality control of herbal medicines / J. Chromatogr. B. 2004. Vol. 812. N 1 – 2. P. 53 – 70.
20. **Jiang Y., David B., Tu P., Barbin Y.** Recent analytical approaches in quality control of traditional chinese medicines — a review / Anal. Chim. Acta. 2010. Vol. 657. N 1. P. 9 – 18.
21. **Rodionova O. E.** Chemometric approach to big data in chemistry / Ross. Khim. Zh. 2006. Vol. 50. N 2. P. 128 – 144 [in Russian].
22. **Monakhova Y. B., Holzgrabe U., Diehl B. W.** Current role and future perspectives of multivariate (chemometric) methods in NMR spectroscopic analysis of pharmaceutical products / J. Pharm. Biomed. Anal. 2017. Vol. 147. P. 580 – 589.
23. **Kumar D.** Nuclear magnetic resonance (NMR) spectroscopy for metabolic profiling of medicinal plants and their products / Critical Rev. Anal. Chem. 2016. Vol. 46. N 5. P. 400 – 412.
24. **Christopher M. B.** Pattern recognition and machine learning. — New York: Springer-Verlag, 2016.
25. **Bridges Jr. C. C.** Hierarchical cluster analysis / Psychological Reports. 1966. Vol. 18. N 3. P. 851 – 854.
26. **Wold S., Esbensen K., Geladi P.** Principal component analysis / Chemometrics and intelligent laboratory systems. 1987. Vol. 2. N 1 – 3. P. 37 – 52.
27. **Mimmack G. M., Mason S. J., Galpin J. S.** Choice of distance matrices in cluster analysis: Defining regions / J. Climate. 2001. Vol. 14. N 12. P. 2790 – 2797.
28. **Mao J., Xu J.** Discrimination of herbal medicines by molecular spectroscopy and chemical pattern recognition / Spectrochim. Acta. Part A: Molecular and Biomolecular Spectroscopy. 2006. Vol. 65. N 2. P. 497 – 500.
29. **Bai Y., Wang X., Lei J., et al.** Discrimination of fructus forsythiae according to geographical origin with near-infared spectroscopy / 33 Biomed. Eng. Biotechnol. (iCBEB). 2012. P. 175 – 178.
30. **Schulz H., Baranska M., Quilitzsch R., et al.** Characterization of peppercorn, pepper oil, and pepper oleoresin by vibrational spectroscopy methods / J. Agr. Food Chem. 2005. Vol. 53. N 9. P. 3358 – 3363.
31. **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* / J. Liquid Chromatogr. Relat. Technol. 2015. Vol. 38. N 14. P. 1407 – 1416.
32. **Abdi H., Williams L. J.** Principal component analysis / Wiley Interdisciplinary Reviews: Computational Statistics. 2010. Vol. 2. N 4. P. 433 – 459.
33. **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 / Anal. Chim. Acta. 2007. Vol. 592. N 2. P. 121 – 131.
34. **Daolio C., Beltrame F. L., Ferreira A. G., et al.** Classification of commercial catuaba samples by NMR, HPLC and chemometrics / Phytochem. Anal. 2008. Vol. 19. N 3. P. 218 – 228.
35. **Flores I. S., Silva A. K., Furquim L. C., et al.** HR-MAS NMR allied to chemometric on *Hancornia speciosa* varieties differentiation / J. Brazil. Chem. Soc. 2018. Vol. 29. N 4. P. 708 – 714.
36. **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 / J. Mol. Structure. 2014. Vol. 1069. P. 229 – 235.
37. **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 / J. Sep. Sci. 2015. Vol. 38. N 3. P. 395 – 401.
38. **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 / Anal. Bioanal. Chem. 2017. Vol. 409. N 19. P. 4669 – 4679.
39. **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 / J. Pharm. Biomed. Anal. 2017. Vol. 140. P. 20 – 30.
40. **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 / Phytochem. Anal. 2016. Vol. 27. N 3 – 4. P. 158 – 167.
41. **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 / J. Pharmacy Pharmacol. 2017. Vol. 69. DOI: 10.1111/jph.p.12782.
42. **Oliveira I., Pinto T., Faria M., et al.** Morphometrics and chemometrics as tools for medicinal and aromatic plants characterization / J. Appl. Botany Food Quality. 2017. Vol. 90. P. 31 – 42.
43. **Bittner M., Schenk R., Springer A., Melzig M. F.** Economical, plain, and rapid authentication of *Actaea acemosa* L. (syn. *Cimicifuga acemosa*, black cohosh) herbal raw material by resilient RP-PDA-HPLC and chemometric analysis / Phytochem. Anal. 2016. Vol. 27. N 6. P. 318 – 325.
44. **Zimmermann B., Kohler A.** Infrared spectroscopy of pollen identifies plant species and genus as well as environmental conditions / PLoS One. 2014. Vol. 9. N 4. P. e95417.
45. **Schulz H., Özkan G., Baranska M., et al.** Characterisation of essential oil plants from Turkey by IR and Raman spectroscopy / Vibr. Spectrosc. 2005. Vol. 39. N 2. P. 249 – 256.
46. **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 Mag. 2017. Vol. 13. Suppl. 3. P. S613.
47. **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 / Anal. Chim. Acta. 2012. Vol. 747. P. 76 – 83.
48. **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 / Microchem. J. 2017. Vol. 130. P. 79 – 87.
49. **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 Chem. 2014. Vol. 158. P. 101 – 111.
50. **Gad H. A., Bouzabata A.** Application of chemometrics in quality control of turmeric (*Curcuma longa*) based on ultra-violet, Fourier transform-infrared and ^{1}H NMR spectroscopy / Food Chem. 2017. Vol. 237. P. 857 – 864.
51. **Viapiana A., Struck-Lewicka W., Konieczynski P., et al.** An approach based on HPLC-fingerprint and chemometrics to quality consistency evaluation of *Matricaria chamomilla* L. commercial samples / Front. Plant Sci. 2016. Vol. 7. P. 1561.
52. **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 / J. Natur. Med. 2016. Vol. 70. N 4. P. 803 – 810.
53. **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 / J. Mol. Struct. 2015. Vol. 1101. P. 101 – 108.
54. **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. N 5. P. 583.
55. **Hinton G. E., Salakhutdinov R. R.** Reducing the dimensionality of data with neural networks / Science. 2006. Vol. 313. N 5786. P. 504 – 507.
56. **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. N 5. P. 753 – 761.
57. **Zhu Y., Tan A. T. L.** Discrimination of wild-grown and cultivated *Ganoderma lucidum* by Fourier transform infrared spectroscopy and chemometric methods / American J. Anal. Chem. 2015. Vol. 6. N 5. P. 480 – 491.
58. **Lever J., Krzywinski M., Altman N.** Points of significance: Principal component analysis / Nature Methods. 2017. Vol. 14 N. 14. P. 641 – 642.
59. **Refaeilzadeh P., Tang L., Liu H.** Cross-validation / Encyclopedia of database systems. — Springer, 2009. P. 532 – 538.
60. **Witten I. H., Frank E., Hall M. A., Pal C. J.** Data mining: Practical machine learning tools and techniques / Morgan Kaufmann. 2016.
61. **Fan Q., Wang Y., Sun P., et al.** Discrimination of ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis / Talanta. 2010. Vol. 80. N 3. P. 1245 – 1250.
62. **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 / Anal. Chim. Acta. 2008. Vol. 618. N 2. P. 121 – 130.
63. **Wold S., Sjöström M., Eriksson L.** *PLS*-regression: A basic tool of chemometrics / Chemometrics and Intelligent Laboratory Systems. 2001. Vol. 58. N 2. P. 109 – 130.
64. **Geladi P., Kowalski B. R.** Partial least-squares regression: A tutorial / Anal. Chim. Acta. 1986. Vol. 185. P. 1 – 17.
65. **Tarachiwin L., Katoh A., Ute K., Fukusaki E.** Quality evaluation of *Angelica acutiloba kitagawa* roots by ^{1}H NMR-based metabolic fingerprinting / J. Pharm. Biomed. Anal. 2008. Vol. 48. N 1. P. 42 – 48.
66. **Li Y., Zhang J., Zhao Y., et al.** Characteristic fingerprint based on low polar constituents for discrimination of *Wolfiporia extensa* according to geographical origin using UV spectroscopy and chemometrics methods / J. Anal. Meth. Chem. 2014. Vol. 2014.
67. **Zhao Y., Zhang J., Jin H., et al.** Discrimination of *Gentiana rigescens* from different origins by Fourier transform infrared spectroscopy combined with chemometric methods / J. AOAC Int. 2015. Vol. 98. N 1. P. 22 – 26.
68. **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 / Biochem. System. Ecol. 2017. Vol. 70. P. 155 – 161.
69. **Hu Y., Kong W., Yang X., et al.** GC-MS combined with chemometric techniques for the quality control and original discrimination of *Curcuma longa* rhizome: Analysis of essential oils / J. Sep. Sci. 2014. Vol. 37. N 4. P. 404 – 411.
70. **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 / J. Sci. Food Agr. 2016. Vol. 96. N 13. P. 4475 – 4483.
71. **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 Biochem. 2015. Vol. 16. N 1. P. 9.
72. **Hoffmann J. F., Carvalho I. R., Barbieri R. L., et al.** *Butia* spp. (Arecaceae) LC-MS-based metabolomics for species and geographical origin discrimination / J. Agr. Food Chem. 2017. Vol. 65. N 2. P. 523 – 532.
73. **Zheng S., Jiang X., Wu L., et al.** Chemical and genetic discrimination of *Cistanches herba* based on UPLC-QTOF/MS and DNA barcoding / PloS One. 2014. Vol. 9. N 5. P. e98061.
74. **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 Res. Int. 2018. Vol. 109. P. 387 – 402.
75. **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 Res. Int. 2017. Vol. 102. P. 700 – 709.
76. **Tan T., Zhang J., Xu X., et al.** Geographical discrimination of *Glechomae herba* based on fifteen phenolic constituents determined by LC-MS/MS method combined with chemometric methods / Biomed. Chromatogr. 2018. P. e4239.
77. **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.
78. **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. N 1. P. 33 – 37.
79. **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. N 10. P. e0186664.
80. **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 Chem. Lett. 2013. Vol. 24. N 7. P. 639 – 642.
81. **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 Chem. 2014. Vol. 152. P. 391 – 398.
82. **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* / Phytochem. Anal. 2013. Vol. 24. N 6. P. 550 – 555.
83. **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 / J. Chromatogr. A. 2016. Vol. 1454. P. 67 – 77.
84. **Mncwangi N. P., Viljoen A. M., Zhao J., et al.** What the devil is in your phytomedicine? Exploring species substitution in *Harpagophytum* through chemometric modeling of ^{1}H-NMR and UHPLC-MS datasets / Phytochem. 2014. Vol. 106. P. 104 – 115.
85. **Mavimbela T., Viljoen A., Vermaak I.** Differentiating between *Agathosma betulina* and *Agathosma crenulata*. A quality control perspective / J. Appl. Res. Med. Arom. Plants. 2014. Vol. 1. N 1. P. e8 – e14.
86. **Liaw A., Wiener M.** Classification and regression by randomForest / R News. 2002. Vol. 2. N 3. P. 18 – 22.
87. **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 Anal. Meth. 2018. Vol. 11. N 7. P. 1927 – 1935.
88. **Steinwart I., Christmann A.** Support vector machines. — New York: Springer-Verlag, 2008. — 601 p.
89. **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 / Anal. Chim. Acta. 2009. Vol. 642. N 1 – 2. P. 257 – 265.

90. **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 / Anal. Chim. Acta. 2012. Vol. 712. P. 37 – 44.
91. **Yao S., Li T., Liu H., et al.** Traceability of Boletaceae mushrooms using data fusion of UV-visible and FTIR combined with chemometrics methods / J. Sci. Food Agr. 2018. Vol. 98. N 6. P. 2215 – 2222.
92. **Dall'Acqua Y. G., Cunha Júnior L. C., Nardini V., et al.** Discrimination of *Euterpe oleracea* Mart. (Açaí) and *Euterpe edulis* Mart. (Juçara) intact fruit using near-infrared (NIR) spectroscopy and linear discriminant analysis / J. Food Proc. Preser. 2015. Vol. 39. N 6. P. 2856 – 2865.
93. **Wold S., Sjöström M.** SIMCA: A method for analyzing chemical data in terms of similarity and analogy. — Wash., D.C.: ACS Publications, 1977. P. 243 – 282.
94. **Wang P., Yu Z.** Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review / J. Pharm. Anal. 2015. Vol. 5. N 5. P. 277 – 284.
95. **Li W., Cheng Z., Wang Y., Qu H.** Quality control of *Lonicera japonica* flos using near infrared spectroscopy and chemometrics / J. Pharm. Biomed. Anal. 2013. Vol. 72. P. 33 – 39.
96. **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 / Phytochem. Anal. 2013. Vol. 24. N 6. P. 520 – 526.
97. **Deconinck E., Aouadi C., Bothy J., Courselle P.** Detection and identification of multiple adulterants in plant food supplements using attenuated total reflectance — Infrared spectroscopy / J. Pharm. Biomed. Anal. 2018. Vol. 152. P. 111 – 119.
98. **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. Vol. 64. N 4. P. 943 – 948.
99. **Kramer O.** *K*-nearest neighbors / Dimensionality Reduction with Unsupervised Nearest Neighbors. — Springer, 2013. P. 13 – 23.
100. **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 / J. Chromatogr. A. 2009. Vol. 1216. N 11. P. 2150 – 2155.
101. **Li C., Yang S.-C., Guo Q.-S., et al.** Geographical traceability of *Marsdenia tenacissima* by Fourier transform infrared spectroscopy and chemometrics / Spectrochim. Acta. Part A. 2016. Vol. 152. P. 391 – 396.
102. **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 / Phytochem. Anal. 2012. Vol. 23. N 1. P. 34 – 43.
103. **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.
104. **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. N 2. P. 154 – 161.
105. **Gudi G., Krähmer A., Krüger 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 / J. Agr. Food Chem. 2015. Vol. 63. N 39. P. 8743 – 8750.
106. **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 / J. Food Drug Anal. 2014. Vol. 22. N 3. P. 336 – 344.
107. **Hyvärinen A., Karhunen J., Oja E.** Independent component analysis. — John Wiley & Sons. 2004.
108. **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 / Int. J. Food Properties. 2018. Vol. 20. Suppl. 3. P. 1 – 18.