УДК 543.422:633.88

П.С. Колодочка,
А.В. Ляхнович,
М. А. Ходасевич, д-р физ.-мат. наук
(ИФ НАН Беларуси, г. Минск, Беларусь);
Си Хунджу, PhD,
Шу Джей,
Сюй Вэй,
(Аньхойский Вост.-Кит. НИИ «Фотоэлектроника», г. Уху, Китай)

ПОВЫШЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ ГЕОГРАФИЧЕСКОГО ПРОИСХОЖДЕНИЯ ЛЕКАРСТВЕННОГО РАСТИТЕЛЬНОГО СЫРЬЯ ПРИ ПОМОЩИ ВЫБОРА СПЕКТРАЛЬНЫХ ПЕРЕМЕННЫХ В ТГЦ СПЕКТРАХ ПРОПУСКАНИЯ МЕТОДОМ ПОИСКА КОМБИНАЦИИ ОКОН ОПТИМАЛЬНОЙ ШИРИНЫ

Несмотря на богатую историю использования лекарственного растительного сырья (ЛРС) в традиционных медицинских системах различных культур, научный интерес к ним зачастую оказывается недостаточно высоким. Основной причиной этого является ряд факторов, которые ограничивают интенсивность и масштаб исследований в данной области. Во-первых, лекарственные растения часто рассматриваются как устаревшие или недостаточно эффективные по сравнению с современными синтетическими препаратами, что снижает их популярность среди фармацевтических компаний и исследовательских институтов, ориентированных на коммерциализацию результатов. Во-вторых, сложности, связанные со стандартизацией и контролем качества растительного сырья, а также вариативность его химического состава в зависимости от региона происхождения, способа сбора и обработки, создают дополнительные барьеры для проведения верифицируемых исследований [1]. Для унификации средств традиционной медицины необходимо определять состав и количественное содержание компонентов растительных продуктов, чтобы обеспечить эффективность, надежность и повторяемость действия лекарств и снизить их возможные побочные эффекты.

На предыдущем этапе работы нами разработаны многопараметрические классификационные модели для определения географического происхождения клубнеобразных корней *Gastrodia* [2], используемых в традиционной китайской медицине для лечения головных болей, головокружения, потери чувствительности конечностей, судорогах, неврастении, эпилепсии, невралгии и других болезней. Классифи-

кация осуществляется на основе анализа многопараметрических данных, полученных с помощью ТГц импульсной спектроскопии поглощения. Процессы подготовки образцов исследуемого лекарственного сырья, измерения временных форм прошедших через них ТГц импульсов, вычисления спектров пропускания и применения методов машинного обучения для классификации географического происхождения описаны в [2]. В данной работе описано повышение точности модели k ближайших соседей (kNN – k nearest neighbors) [2], которая классифицировала 50 образцов сырья из пяти географических локаций (Чанбайшань, Цзиньчжай, Шэньнунцзя, Сицзан и Юньнань) с точностью, прецизионностью и чувствительностью не хуже 0.98, с помощью применения выбора спектральных переменных методом поиска комбинации движущихся окон оптимальной ширины (scmwi – search combination moving window interval) [3].

На первом этапе задается ширина спектрального окна, содержащего на одну спектральную переменную больше, чем размерность пространства главных компонент [4], в котором производится кластерный анализ. Вычисленные спектры пропускания исследуемых образцов обрезаются с высокочастотной стороны до количества переменных, кратного ширине окна.

Следующий этап – определение оптимального положения первого спектрального окна по максимуму точности классификационной модели. Модель строится на каждом шаге сдвига окна на одну спектральную переменную. Из пространства спектральных переменных в составе окна выбирается пространство главных компонент размерности, определенной на этапе построения широкополосной модели. Эти главные компоненты служат входными данными для модели kNN [5], где k определяется по максимальной точности классификации модели без выбора спектральных переменных. При построении модели определяется матрица ошибок, отношение суммы диагональных элементов которой к полному количеству образцов определяет точность классификации. Максимальная точность в зависимости от положения первого окна определяет его фиксированное положение на все следующие этапы выполнения алгоритма выбора спектральных переменных.

Второе окно последовательно сдвигается на одну спектральную переменную в пределах оставшихся объединенных переменных. Построение классификационных моделей происходит по множеству спектральных переменных, состоящих из первого и второго окон. Целью выбора положения второго окна является как и на первом этапе достижение максимальной точности. Затем этот этап повторяется до использования в моделировании всех спектральных переменных.

Зависимость точности классификации от количества выбранных спектральных окон приведена на рисунке 1 для следующих параметров модели: из рассмотрения исключена полоса частот 0,6–1,6 ТГц; применена нормировка спектров на стандартное отклонение и сглаживание фильтром Савицкого-Голея полиномом третьей степени по 17 отсчетам; рассмотрено пятимерное пространство главных компонент с метрикой Махаланобиса; учитываются 2 ближайших соседа.

Видно, что при учете 2 окон шириной 6 спектральных переменных (всего 12 переменных, рисунок 2) классификация становится достоверной. Для описанной выше модели точность, прецизионность (усредненная доля объектов, правильно классифицированных среди фактически принадлежащих классу), чувствительность (усредненная доля объектов, фактически принадлежащих классу среди классифицированных моделью в него) и коэффициент корреляции Мэтьюса (обобщение коэффициента корреляции Пирсона на случай рассмотрения более двух классов) равны 1. Стоит отметить, что выбранные спектральные переменные находятся вблизи двух пиков предобработанных спектров.

Реализован метод поиска комбинации движущихся окон в спектрах пропускания ТГц частотного диапазона для повышения точности решения задачи контроля качества лекарственного растительного сырья методом k ближайших соседей. Достигнуты единичные точность, прецизионность и чувствительность классификации пяти географических локаций корней *Gastrodia*.

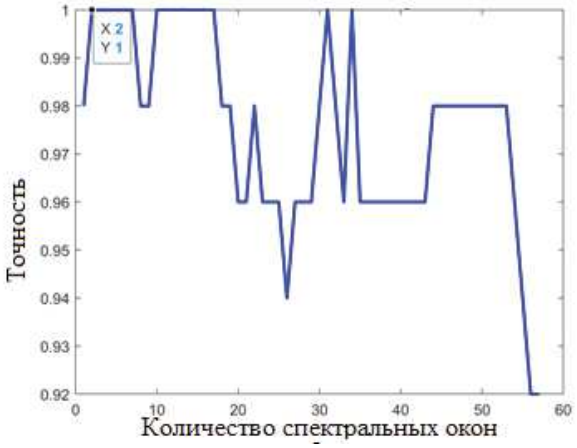

Рисунок 1 – Точность классификации образцов лекарственного растительного сырья *Gastrodia* из пяти географических локаций методом kNN в пятимерном пространстве главных компонент в зависимости от количества учитываемых в методе scmwi спектральных окон шириной 6 переменных

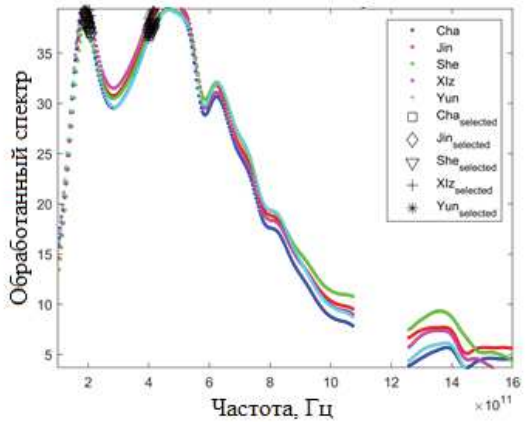

Рисунок 2 – Нормированные на среднеквадратичное отклонение спектры пропускания лекарственного растительного сырья Gastrodia при изъятии из рассмотрения частотного диапазона 0,6–1,6 ТГц и выбранные 12 спектральных переменных для построения классификационной модели k ближайших соседей (k=2) в пятимерном пространстве главных компонент с метрикой Махаланобиса

ЛИТЕРАТУРА

1. Liang, Y-Z. Quality control of herbal medicines / Y-Z. Liang, P. Xie, K. Chan. // Journal of Chromatography B. – 2004. – V. 812(1–2). – P. 53–70.

2. Многопараметрическая классификация географического происхождения лекарственного растительного сырья по ТГц спектрам пропускания / [П. Колодочка и др.] // Информационные технологии. Физика и математика : материалы 89-й научно-технической конференции профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием), Минск, 3-18 февраля 2025 г. / Белорусский государственный технологический университет. – Минск: БГТУ, 2025. – 393 с.

3. Ходасевич М.А., Асеев В.А. Выбор спектральных переменных и повышение точности калибровки температуры методом проекции на латентные структуры по спектрам флуоресценции Yb3+:CaF2 // Оптика и спектроскопия. – 2018 – Т. 124, № 5. – С. 713–717.

4. Esbensen, K.H. Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice / K.H. Esbensen, P. Geladi // Comprehensive Chemometrics / ed.: S. Brown, R.Tauler, B. Walczak. – 2009. – P. 211–226.

5. Ходасевич, М.А. Многопараметрический подход в методах оптической диагностики: основы и применения / М. А. Ходасевич; Нац. акад. наук Беларуси, Ин-т физики им. Б.И. Степанова. – Минск: Беларуская навука, 2024. – 114 с.