УДК 543.42.061, 543.421/.424

Колодочка П. С.[1,2], Куликовская П. А.[1,2], Ходасевич М. А.[2]

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА К ГЛАВНЫМ КОМПОНЕНТАМ СПЕКТРОВ ОПТИЧЕСКОЙ ПЛОТНОСТИ ДЛЯ КЛАССИФИКАЦИИ САХАРОВ И ПЛАСТМАСС

[1]Белорусский государственный университет, Минск, Беларусь
[2]Институт физики НАН Беларуси, Минск, Беларусь

Разработаны многопараметрические модели классификации рафинированных тростниковых и свекловичных сахаров по спектрам оптической плотности в диапазоне длин волн от 200 до 1380 нм и классификации пяти видов пластмасс по спектрам оптической плотности в диапазоне от 1500 до 3100 нм. Выбор спектральных переменных по величине нагрузок в первую главную компоненту спектров позволил получить 100 % точность классификации 102 сахаров методом иерархического кластерного анализа и 95 % точность классификации 80 образцов пластмасс методом k ближайших соседей.

В настоящее время для повышения качества классификации объектов часто используют методы анализа многопараметрических спектральных данных. Целью настоящей работы является демонстрация возможностей многопараметрического спектрального анализа для качественной характеризации объектов исследования на примерах дискриминации рафинированных сахаров по типу растительного сырья для их производства и классификации видов пластмасс. Первым этапом многопараметрического анализа является сжатие данных методом главных компонент (PCA – principal component analysis) [1]. На втором этапе для качественного анализа в пространстве главных компонент применялись иерархический кластерный анализ (HCA – hierarchical cluster analysis) [2] и метод k ближайших соседей (kNN – k nearest neighbors) [3].

Спектры оптической плотности 25 % водных растворов сахаров были зарегистрированы на спектрофотометре Shimadzu UV-3101PC в диапазоне длин волн от 200 до 1380 нм с шагом 1 нм и шириной щели 1 нм. Из 102 исследованных образцов 45 являются растворами свекловичного сахара и 57 – тростникового. Измерения спектров пластиков проводились в диапазоне от 1500 до 3100 нм с интервалом 2 нм и шириной щели 1 нм. В данной работе исследуются пять видов пластика: полиэтилентерефталат (1), полиэтилен высокой плотности (2), полиэтилен низкой плотности (4), полипропилен (5) и полистирол (6). В рассматриваемой выборке содержатся 80 различных по толщине прозрачных и окрашенных образцов пластика.

В качестве первого метода предобработки всех спектров использовалось центрирование, приравнивающее нулю среднее по выборке образцов значение для каждой спектральной переменной. Для устранения разницы образцов пластиков по толщине было также использовано нормирование спектров. Среди рассмотренных методов (p-норма, масштабирование по стандартному отклонению, масштабирование по наибольшему значению, масштабирование по медианному абсолютному отклонению и другие) наилучшие результаты были получены при использовании p-нормировки спектров $\tilde{\mathrm{v}}_i = \frac{\mathrm{v}_i}{\left[\sum_{i=1}^{N} |\mathrm{v}_i|^p\right]^{1/p}}$ для $p = 2$, которая является евклидовой нормой и эффективно устраняет разницу в толщине образцов. Здесь p — положительное вещественное значение; v_i и $\tilde{\mathrm{v}}_i$ – ненормированное и нормированное значения i-ой спектральной переменной, $i = 1 \ldots N$, N – количество спектральных переменных.

На рисунке 1 представлены спектры нагрузок в первую главную компоненту предобработанных спектров оптической плотности сахаров и пластиков. Выбор спектральных переменных при вторичном построении пространства главных компонент, в котором проводился кластерный анализ, осуществлялся с помощью последовательного добавления переменных с уменьшающимися нагрузками. Оптимальное количество спектральных переменных (104 для сахаров, 46 для пластиков, представлены на рисунке 2) соответствует максимальной точно-

сти классификации объектов исследования с помощью применения кластерного анализа в пространстве главных компонент спектров с выбранными переменными.

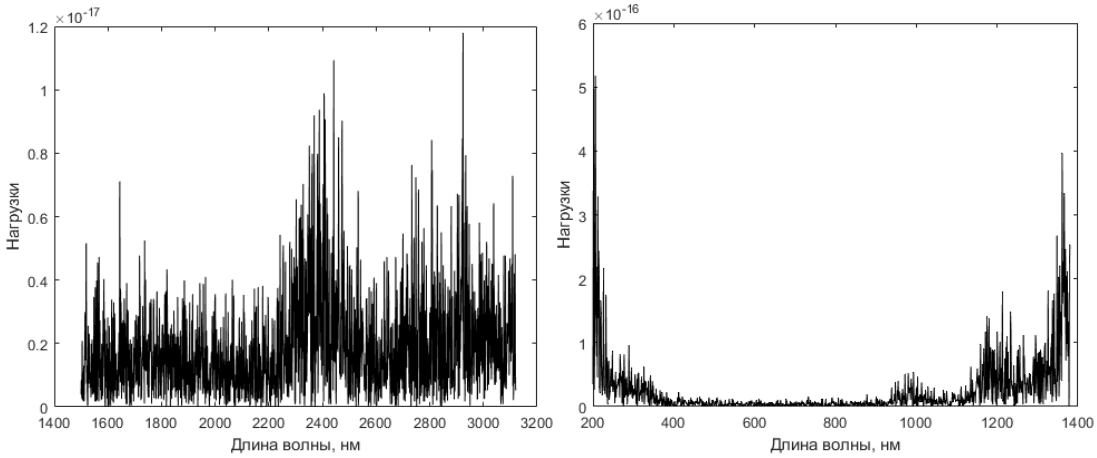

Рисунок 1 – Графики нагрузок для пластиков (слева) и сахаров (справа)

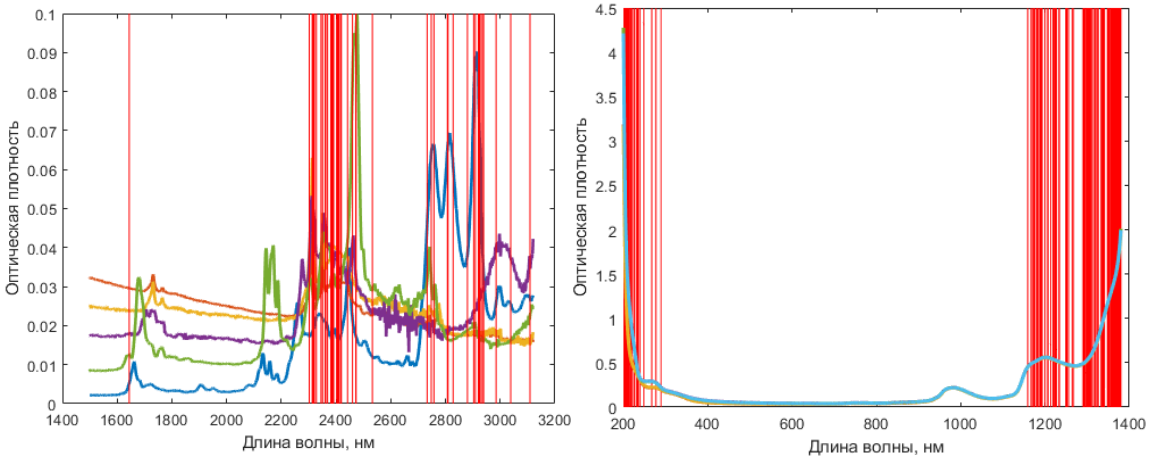

Рисунок 2 – Спектры оптической плотности и выбранные спектральные переменные для пластиков (слева) и сахаров (справа)

Для классификации свекловичных и тростниковых сахаров использован агломеративный HCA. Алгоритм рассматривает каждый объект как отдельный кластер, а затем последовательно объединяет кластера, являющиеся наиболее близкими в соответствии с выбранной метрикой пространства. На рисунке 3 изображены счета в пространстве второй, третьей и четвертой главных компонент 104 спектральных переменных из спектров оптической плотности сахаров, в котором проведена кластеризация образцов сахара агломеративным HCA при использовании метрики городских кварталов (расстояния измеряются параллельно осям координат) и 10-кратной перекрестной проверки. Точность классификации 100 %.

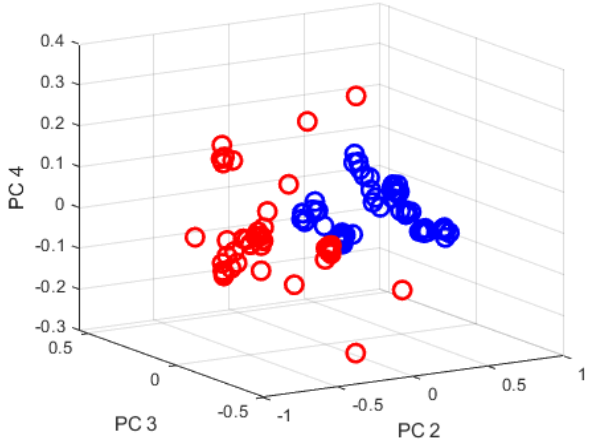

Рисунок 3 – График счетов тростниковых (красный цвет) и свекловичных (синий цвет) сахаров

Для классификации полимеров применялся метод kNN с выбором размерности пространства главных компонент при использовании евклидовой метрики. После случайной инициализации центроидов классов, присвоение членства происходит по правилу принадлежности большинства из k соседей с наименьшими расстояниями до классифицируемого объекта. Оптимизация количества учитываемых при классификации ближайших соседей (k=2) и размерности пространства главных компонент (PC1, PC2, PC3, PC4, PC6, PC10) позволила достичь 95 % точности классификации также при использовании 10-кратной перекрестной проверки. На рисунке 4 представлено двумерное пространство первой и второй главных компонент 46 спектральных переменных из спектров оптической плотности пластиков, в которых кругами отмечены 3 неправильно классифицированных образца.

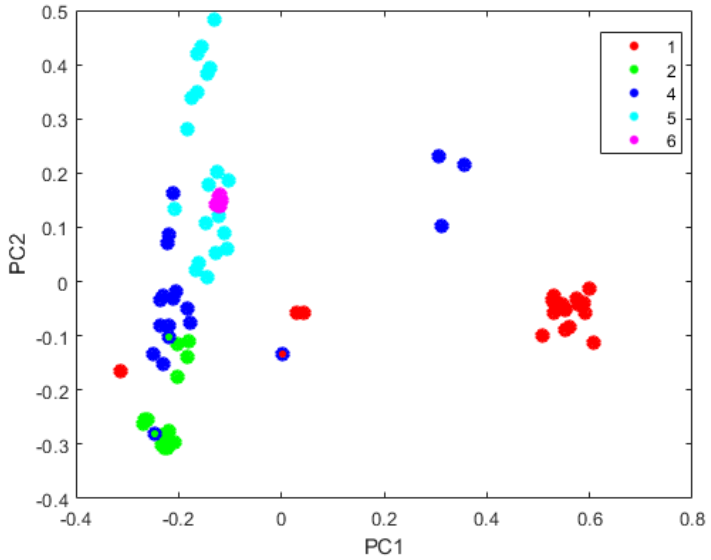

Рисунок 4 – График счетов правильно классифицированных (точки) и неправильно классифицированных (кружки) полимеров в пространстве первой и второй главных компонент 46 спектральных переменных

Высокая точность классификации исследуемых объектов на основе применения многопараметрического анализа с выбором спектральных переменных демонстрирует возможности замены стандартного дорогостоящего и трудоемкого метода релаксометрии ядерного магнитного резонанса на сравнительно дешевый и простой многопараметрический спектральный анализ при проведении классификации растительных источников при производстве рафинированных сахаров и пригодность для практического применения при предварительной сортировке пластмасс, необходимой для решения проблемы вторичного их использования.

Список литературы

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

2. Liao T. W. Clustering of time series data – a survey // Pattern Recognition. –2005. – Vol. 38. – P. 1857–1874.

3. Berrueta L. A., Alonso-Salces R. M., Héberger K. Supervised pattern recognition in food analysis // Journal of Chromatography A. – 2007. – Vol. 1158. – P. 196–214.