УДК 535.343.32, 543.421/.424

П. С. Колодочка[1,2], М. А. Ходасевич[2]
([1]Физический факультет БГУ, [2]Институт физики НАН Беларуси, г. Минск)

КЛАССИФИКАЦИЯ СВЕКЛОВИЧНОГО И ТРОСТНИКОВОГО САХАРА С ПОМОЩЬЮ ПРИМЕНЕНИЯ МЕТОДА БЛИЖАЙШИХ СОСЕДЕЙ К ГЛАВНЫМ КОМПОНЕНТАМ СПЕКТРОВ ПРОПУСКАНИЯ ВОДНЫХ РАСТВОРОВ

Рафинированные сахара содержат в себе более 99 % сахарозы, по этой причине достаточно сложной задачей является определение растительного источника сахара. Стандартным в Европе методом определения вида сахара является релаксометрия ядерного магнитного резонанса, позволяющая определить тип сахарной примеси в меде [1], а также классифицировать и выявить фальсификацию тростникового, свекловичного и кокосового сахаров [2]. Однако ядерно-магнитный резонанс является дорогостоящим и трудоемким методом. Целесообразно найти более простой и дешевый метод определения свекловичного и тростникового сахаров или их фальсификации. Исследование по выявлению фальсификации черного чая [3] является примером успешной работы по выявлению подделок с помощью спектрального анализа.

Целью данной исследовательской работы является построение классификационной модели с применением метода k ближайших соседей (kNN – k nearest neighbors) [4] для определения растительного источника сахаров. Модель основывается на применении метода главных компонент (PCA – principal component analysis) [5] к УФ, видимому и ближнему ИК спектрам оптической плотности водных растворов сахаров.

Спектры оптической плотности регистрировали на спектрофотометре Shimadzu UV-3101PC в диапазоне длин волн от 365 до 1000 нм с шагом 1 нм и шириной щели 1 нм. В работе [6] было показано, что спектры сахаров за пределами этого диапазона не содержат существенных различий. Объектами исследования являлись 25 % водные растворы сахара, оптическая плотность которых вкладывается в пределы динамического диапазона используемого спектрофотометра.

Использовались образцы сахара из 7 стран: Пакистан, Португалия, Польша, Румыния, Италия, Сербия, Беларусь. Из 102 исследованных образцов 45 являются растворами свекловичного сахара и 57 – тростникового. Таким образом, для построения пространства главных компонент используется матрица данных размером 102 на 635, где

102 – количество образцов, а 635 – количество спектральных переменных.

В данной работе применяется метод PCA для анализа информации, выявления выбросов и уменьшения размерности. Вместо исходного множества спектральных переменных набор данных может быть описан с использованием нескольких первых главных компонент без значительной потери данных.

Перед применением метода PCA спектры подверглись предобработке с помощью центрирования. После этого было проведено понижение размерности матрицы исходных данных и выявлено необходимое количество главных компонент. Рассмотрение можно ограничить шестью главными компонентами, которые описывают более 99,96 % информации, содержащейся в спектрах (рис. 1).

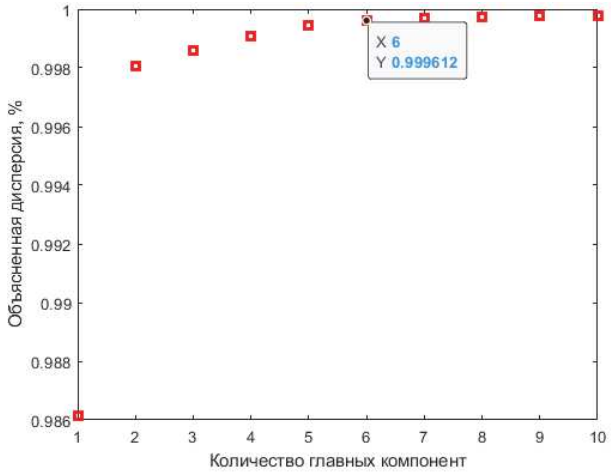

Рисунок 1 – Зависимость объясненной дисперсии от количества главных компонент спектров

Для классификации свекловичного и тростникового сахара использован метод k ближайших соседей. Это один из простейших методов классификации, который основан на оценивании сходства объектов, определяемого в данном случае величиной евклидового расстояния между ними в пространстве главных компонент. Класс неизвестного объекта определяется по принадлежности к известным классам большинства из k ближайших объектов.

На рис. 2 изображены счета в пространстве третьей и шестой компоненты, в котором проведена дальнейшая кластеризация образцов сахара по их спектрам. При этом ошибка валидации при перепроверке с помощью обучающей выборки будет минимальной – менее 3 %. Такой результат достигнут при учете 5 ближайших соседей (рис. 3).

Далее модель была подвергнута дополнительной 10-кратной перекрестной проверке. Полученная точность классификации составляет около 94 %.

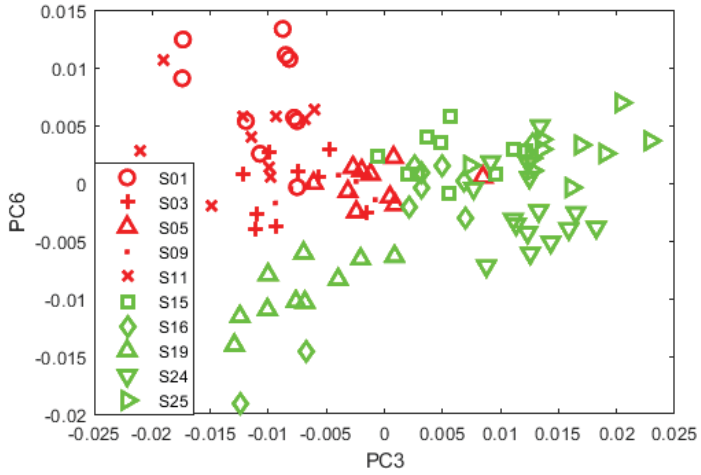

Рисунок 2 – График счетов в 3 и 6 главные компоненты (тростниковый сахар: S01-S11, свекловичный: S15-S25)

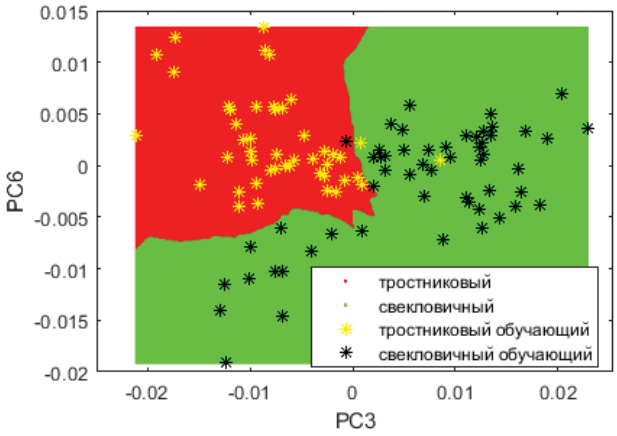

Рисунок 3 – График классификационной модели kNN в пространстве 3 и 6 главных компонент с наложением графика счетов

Таким образом, в ходе исследования были измерены спектры оптической плотности 25 % водных растворов свекловичного и тростникового сахаров, применен метод главных компонент и построена классификационная модель методом kNN. Полученная модель характеризуется точностью около 94 %, которая уступает показателю классификации (около 98 %) с помощью метода построения классификационных деревьев [7], проведенной нами ранее [6]. На рис. 4 представлено оптимальное дерево принятия решений, позволяющее достичь указанной точности классификации растительного источника рафинированного сахара, которая является приемлемой для практического применения вместо эталонного метода релаксометрии ядерного магнитного резонанса.

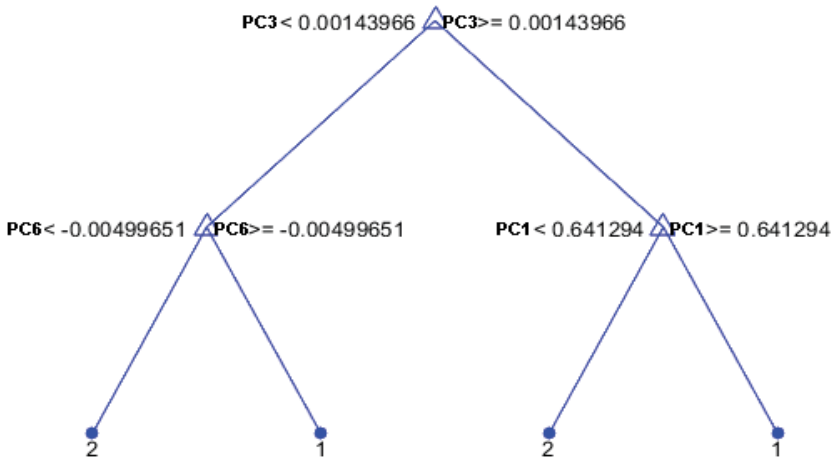

Рисунок 4 – Древо классификации тростникового (1) и свекловичного (2) сахаров

Финансирование работы

Исследование выполнено при частичной поддержке Государственной программы научных исследований Республики Беларусь «Фотоника и электроника для инноваций» в рамках выполнения задания 1.1.

ЛИТЕРАТУРА

1. K. Rachineni [et al.] Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification // Current Research in Food Science – 2022. – Vol. 5. – P. 272-277.

2. R. Bachmann [et al.] Minor metabolites as chemical marker for the differentiation of cane, beet and coconut blossom sugar. From profiling towards identification of adulterations // Food Control – 2022. – Vol.135. – P. 108832.

3. L. Wei, Y. Yang, D. Rapid detection of carmine in black tea with spectrophotometry coupled predictive modelling // Sun Food Chemistry – 2020. – Vol. 329. – P. 127177.

4. 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.

5. R. Bro, A.K. Principal component analysis // Smilde Analytical Methods – 2014. – Vol. 6. – P. 2812–2831.

6. M. Khodasevich, P. Kolodochka [et al.] Classification of sugar types by UV-VIS-NIR spectroscopy and multivariate analysis // The 12th international conference on photonics and applications, Vietnam, September 28 – October 1, 2022 – P. 159.

7. W.-Y. Loh, Fifty Years of Classification and Regression Trees // International Statistical Review – 2014. – Vol. 82. – P. 329-348.