УДК 543.42.061, 543.421/.424

П.А. Куликовская[1], М.А. Ходасевич[2]
([1]Институт физики НАН Беларуси, БГУ, г. Минск;
[2]Институт физики НАН Беларуси, г. Минск)

# КЛАССИФИКАЦИЯ ВИДОВ ПЛАСТМАСС С ПОМОЩЬЮ ПРИМЕНЕНИЯ МЕТОДА БЛИЖАЙШИХ СОСЕДЕЙ К ГЛАВНЫМ КОМПОНЕНТАМ СПЕКТРОВ ОПТИЧЕСКОЙ ПЛОТНОСТИ В ИК-ДИАПАЗОНЕ

**Введение.** На данный момент в мире остро стоит вопрос о вторичном использовании пластмасс. Непереработанные отходы разлагаются на микропластик, который является токсичным. Он уже встречается в продуктах питания и даже в крови человека, что пагубно влияет на здоровье. При этом переработке подвергается около 15 % отходов пластмасс, что обусловлено трудностями, связанными с сортировкой.

**Целью работы** является попытка улучшить качество классификации пластиковых отходов путем предварительной обработки измеренных в NIR-диапазоне спектров оптической плотности пластмасс, использования метода главных компонент PCA (principal component analysis) [1] для уменьшения размерности спектральных данных и одного из методов кластерного анализа – метода k ближайших соседей kNN (k nearest neighbours) [2] – для классификации различных видов бывших в употреблении пластмасс.

**Задачи исследования:** составить представительную выборку образцов пластмасс, измерить спектры оптической плотности образцов в выбранном спектральном диапазоне; применить метод PCA к измеренным и предобработанным спектрам и построить модель классификации образцов методом kNN.

В данной работе исследуются шесть видов пластика (*1. PET* – полиэтилентерефталат*, 2. HDPE* – полиэтилен высокой плотности*, 4. LDPE* – полиэтилен низкой плотности*, 5. PP* – полипропилен*, 6. PS* – полистирол*, 7. OTHER* – прочие виды пластика). В рассматриваемой выборке содержатся 82 различных по толщине прозрачных и окрашенных образца.

**Результаты и обсуждение.** Измерения спектров проводились в диапазоне 1500 до 3100 нм [3] с интервалом 2 нм, разрешающая способность спектрофотометра Shimadzu PC-3101 составляла 0,1 нм, ширина щели равнялась 1 нм. Выбор данного диапазона обусловлен тем, что он не зашумлен и содержит значительные отличия различных ви-

дов пластика. Спектральные данные были объединены в матрицу размерами 82 на 1621, где 1621 – количество спектральных переменных.

Метод главных компонент – один из наиболее распространенных многопараметрических методов, используемый для упрощения больших массивов данных с наименьшими потерями информации. Особенностью PCA является отсутствие нулевой главной компоненты PC0 (principal component), которая могла бы описывать смещение относительно начала системы координат. Поэтому в качестве первого метода предобработки данных использовалось центрирование, приравнивающее нулю среднее по выборке образцов значение для каждой спектральной переменной.

Для выбора второго метода предобработки данных были рассмотрены следующие методы предварительной обработки спектра: p-норма, масштабирование по стандартному отклонению, масштабирование по наибольшему значению, масштабирование по медианному абсолютному отклонению, масштабирование по первому элементу данных и масштабирование по межквартильному размаху. Наилучшие результаты были получены при использовании p-нормировки спектров для p=3, которая эффективно устраняет разницу в толщине образцов пластмасс:

$$\tilde{v}_i = \frac{v_i}{\left[\sum_{i=1}^{N} |v_i|^p\right]^{1/p}}.$$

Здесь $p$ – любое положительное вещественное значение; $v_i$ и $\tilde{v}_i$ – ненормированное и нормированное значения i-ой спектральной переменной, i = 1…N;

Найденные с помощью применения метода PCA к предобработанным спектрам первые четыре главные компоненты описывают 94 % суммарной дисперсии данных. Двумерные графики счетов по этим компонентам представлены на рис. 1. Нумерация видов пластмасс на графике соответствует описанной выше.

Из рис. 1 видно, что визуальное разделение образцов по видам пластмасс неполное, поэтому необходимо применение методов классификации. В данной работе используется метод кластерного анализа kNN. Особенностью его применения является определение расстояний между объектами в пространстве PC. После случайной инициализации центроидов классов присвоение членства происходит по правилу принадлежности большинства из k соседей с наименьшими расстояниями до классифицируемого объекта. В настоящей работе используется евклидово расстояние в двумерном пространстве главных компонент. Вторым этапом применения метода kNN является оптимизация числа рассматриваемых соседей путем вычисления предсказа-

тельной способности модели классификации с различными значениями k.

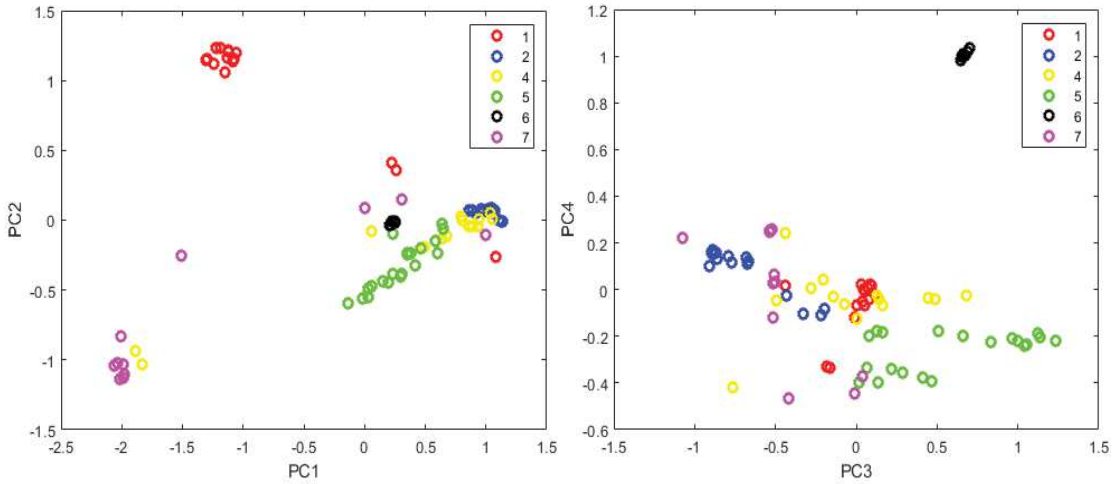


**Рисунок 1 – Графики счетов в пространствах *PC1-PC2* и *PC1-PC3***

Результаты оптимизации метода kNN представлены в табл. 1.

**Таблица 1 – Зависимость ошибки классификации от числа ближайших соседей**

| Количество ближайших соседей | Ошибка классификации, % | | | | |
|---|---|---|---|---|---|
| 2 | 15,85 | 20,73 | 21,95 | 19,51 | 17,07 |
| 3 | 24,39 | 24,39 | 24,39 | 24,39 | 24,39 |
| 4 | 25,61 | 23,17 | 21,95 | 23,17 | 24,95 |
| 5 | 23,17 | 24,39 | 24,39 | 23,17 | 24,39 |
| 6 | 24,39 | 19,51 | 24,39 | 20,73 | 23,17 |

Обычно классификация ограничивается небольшими значениями k и производится несколько раз вследствие случайного выбора начальных условий.

Для выделенных цветом результатов моделей классификации с минимальными ошибками для двух и шести ближайших соседей построены двумерные графики счетов в пространстве PC1-PC2 с указанием принадлежности образцов к классам.

Данные графики визуально отображают кластеризацию методом kNN. Цветами выделены области пространства двух первых PC, положение образцов в которых свидетельствует о принадлежности к определенному классу пластмасс.

**Выводы.** В ходе работы нами была проведена классификация пластиков шести различных видов методом главных компонент и построены модели классификации методом ближайших соседей. Полученная при рассмотрении двух ближайших соседей точность классификации около 84 % является достаточной для практического приме-

нения, но уступает достигнутой ранее точности около 87 % при классификации методом CART (деревья классификации и регрессии) [4].

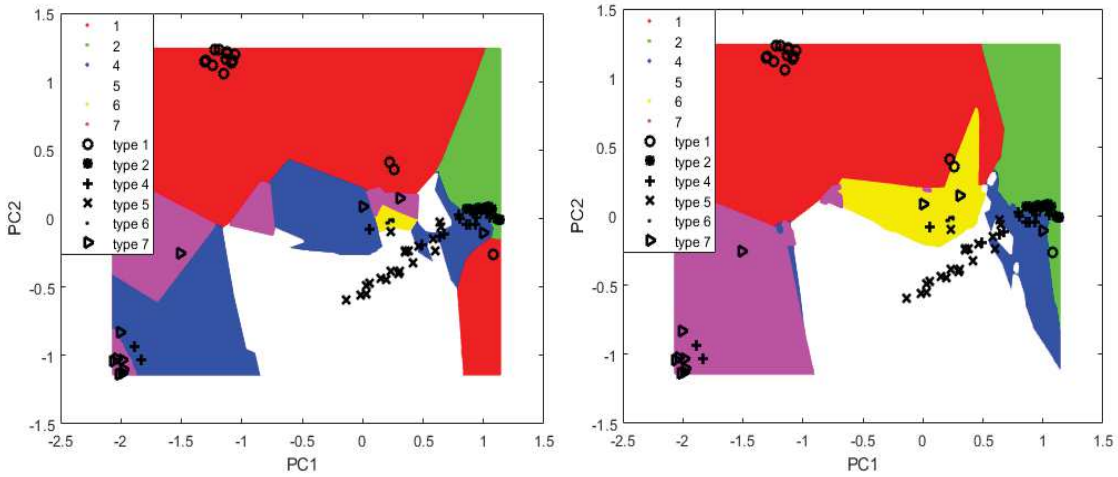


**Рисунок 2 – Графики классификации образцов методом kNN при k=2 и k=6 в пространстве *PC1-PC2***

Дальнейшее повышение точности классификации возможно с помощью применения отличных от евклидовой метрик расстояний между объектами в пространстве главных компонент.

**Финансирование работы**

Исследование выполнено при поддержке Белорусского фонда фундаментальных исследований в рамках выполнения проекта Ф22-031.

ЛИТЕРАТУРА

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

2. 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.

3. Plastic solid waste identification system based on near infrared spectroscopy in combination with support vector machine / S. Zhu [et al.] // Adv. Ind. Eng. Polym. Res. – 2019. – Vol. 2. – P. 77–81.

4. Khodasevich M. A., Kulikovskaya P. A. Application of NIR spectroscopy, principal component analysis and classification tree for plastic sorting // The 12$^{th}$ international conference on photonics and applications, Vietnam, September 28 – October 1, 2022 – P. 158.