

УДК 634.0.813.1:543.422.4

М. А. Зильберглейт, В. М. Резников

Белорусский технологический институт им. С. М. Кирова

ПРИМЕНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ОБРАЗОВ В ИК-СПЕКТРОСКОПИИ КОМПОНЕНТОВ РАСТИТЕЛЬНОЙ ТКАНИ

Методы распознавания образов нашли свое применение для решения ряда важных задач, стоящих перед различными областями знаний. В химии эти методы широко используются для выяснения и предсказания структуры и свойств вещества на основе данных ГЖХ, масс-, ИК-, ЯМР- и электрохимической спектроскопии [1].

Как известно [2, 3], ИК-спектры компонентов растительной ткани (полисахаридов, лигнинов, лигноуглеводных комплексов) содержат широкие и перекрывающиеся полосы, что затрудняет идентификацию этих полимеров. В связи с этим в настоящей работе предпринята попытка использования методов распознавания образов в целях классификации таких объектов по их ИК-спектрам. Кроме того, была оценена возможность идентификации неизвестного объекта по его ИК-спектру при наличии обучающих выборок. Объектом, который надо было идентифицировать, служил продукт А, выделенный из реакционной среды в ходе окислительной деструкции древесного комплекса ели надуксусной кислотой.

Всего рассматривалось 14 объектов: глюкоманнан (1), ксилан (2), альгиновая кислота (3), лигноуглеводный комплекс (ЛУК) ели (4), ЛУК березы (5—8), «нативный» лигнин пихты (9), диоксанлигнин березы (10), ЛМР ели (11), ЛМР сосны (12), «нативный» лигнин красной сосны (13), продукт А (14).

Все ИК-спектры рассматриваемых объектов, согласно литературным данным, были получены путем запрессовки образцов в таблетку с КВг, т. е. по единой методике.

Из анализа исходных данных следует, что число конечных классов может колебаться от 3+1 до 6+1. Возможное увеличение числа классов на единицу объясняется включением в выборку спектра альгиновой кислоты, присутствие которой не характерно для компонентов древесных растительных тканей. Включение альгиновой кислоты в исходную выборку было продиктовано возможностью ее сходства с продуктом А.

При использовании в качестве классифицирующего правила алгоритмов, основанных на использовании функции расстояния, исходные данные преобразовывали по следующему правилу. В изучаемых спектрах выделяли область $1852-719 \text{ см}^{-1}$, которую разбивали на интервалы по 0,15 и 0,2 мкм в диапазонах $1852-1010 \text{ см}^{-1}$ и $1010-719 \text{ см}^{-1}$ соответственно. Такое разбиение дало 50 полос поглощения. Интенсивность максимального пика в полосе шириной 0,15 мкм ($1852-1010 \text{ см}^{-1}$) или 0,2 мкм ($1010-719 \text{ см}^{-1}$) принимали равной 3. Разбивая области $1852-1010 \text{ см}^{-1}$ и $1010-719 \text{ см}^{-1}$ на полосы шириной 1,5 и 2 мкм соответственно, находили в каждой из полос наибольший пик, высоту которого принимали равной 2. Отсутствие пиков в той или иной по-

лосе шириной 0,15 мкм (1852—1010 см⁻¹) и 0,2 мкм (1010—719 см⁻¹) отвечало нулевому значению интенсивности; всем остальным пикам приписывали интенсивность единичной длины.

В случае же использования в классифицирующем правиле метрики, основанной на сходстве, преобразование исходных данных осуществлялось следующим образом. Область разбиения оставалась прежней, а интенсивностям приписывались значения дихотомических (булевых) признаков, т.е. наличие полосы поглощения характеризовалось интенсивностью равной 1, а отсутствие — 0.

В качестве классифицирующего правила, использующего как меру близости расстояние, была применена иерархическая кластер-процедура.

Расстояние между объектами определяли как расстояние по Хеммингу:

$$d_{ij} = \sum_{k=1}^m |X_i^k - X_j^k|.$$

Здесь d_{ij} — расстояние между объектами i и j ;

k — размерность исследуемого пространства, $m = 50$;

X — величина признака.

Алгоритму классификации с помощью последовательной кластер-процедуры отвечают различные версии [4]. В работе использованы три версии, известные под названиями «ближайший сосед», «дальний сосед», «простое среднее». Классификация с использованием данного метода и методов, описанных ниже, показала, что альгиновая кислота чаще всего образует самостоятельный класс, поэтому все результаты распознавания содержат от 4 до 7 классов. Эти данные представлены в табл. 1, там же приведены результаты классификации, выполненной по алгоритму кратчайшего незамкнутого пути (КНП) [5].

Таблица 1

РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ, ИСПОЛЬЗУЮЩЕЙ В КАЧЕСТВЕ МЕРЫ БЛИЗОСТИ РАССТОЯНИЕ

Алгоритм	Классы						
	1	2	3	4	5	6	7
Ближайший сосед	1	3	5	8	14	6, 7	2, 4, 9, 10, 11, 12, 13
Дальний сосед	1	2	3	10	8, 14	5, 6, 7	4, 9, 11, 12, 13
Простое среднее	1	2	3	10	8, 14	5, 6, 7	4, 9, 11, 12, 13
КНП	1	10	3	5	8	6, 7, 14	2, 4, 9, 11, 12, 13
КНП	1	3	5	8	14	6, 7	2, 4, 9, 11, 12, 13, 10
КНП	1	3	5	10	14	6, 7	2, 4, 8, 9, 11, 12, 13

	1	2	3	4	5
Ближайший сосед	1	3	5	7, 6, 14	2, 4, 8, 9, 10, 11, 12, 13
Дальний сосед	3	2, 10	1, 8, 14	5, 6, 7	4, 9, 11, 12, 13
КНП	1	3	5	6, 7, 14	2, 4, 8, 9, 10, 11, 12, 13
Простое среднее	3	2, 10	1, 8, 14	5, 6, 7	4, 9, 11, 12, 13

	1	2	3	4
Ближайший сосед	1	3	5, 6, 7, 14	2, 4, 8, 9, 10, 11, 12, 13
Дальний сосед	2, 10	5, 6, 7	1, 3, 8, 14	4, 9, 11, 12, 13

Используя в качестве меры близости сходство и рассматривая матрицу наблюдений как матрицу, описывающую структуру взаимосвязей объектов наблюдений, можно попытаться классифицировать объекты, применив метод корреляционных плеяд [6].

В качестве коэффициентов сходства были использованы следующие величины:

$$S'_{ij} = \frac{p_{ij}}{m};$$

$$S''_{ij} = \frac{p_{ij}}{m + g_{ij}},$$

где p_{ij} — общее число совпадающих признаков (булевых);

g_{ij} — общее число несовпадающих признаков (булевых).

Результаты классификации с использованием коэффициента сходства приведены в табл. 2.

Таблица 2

РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ, ИСПОЛЬЗУЮЩЕЙ В КАЧЕСТВЕ МЕРЫ БЛИЗОСТИ СХОДСТВО

Мера сходства	Классы						
	1	2	3	4	5	6	7
S'_{ij}	1	3	4	8	10	6, 7, 14	2, 5, 9, 11, 12, 13
S'_{ij}	1	3	5	8	10	6, 7, 14	2, 4, 9, 11, 12, 13
S'_{ij}	1	3	4	5	10	6, 7, 14	2, 8, 9, 11, 12, 13
S'_{ij}	1	3	5	10	11	6, 7, 14	2, 4, 8, 9, 12, 13
S''_{ij} *	1	3	5	10	11	6, 7, 14	2, 4, 8, 9, 12, 13

* Один из возможных вариантов классификации.

Анализ полученных данных свидетельствует о том, что с помощью методов распознавания удастся выделить достаточно достоверные классы соединений¹. Что же касается продукта А, выделенного из реакционной смеси, то исходя из результатов классификации его следует отнести к лигноуглеводным комплексам. Для проверки этого предположения был использован эвристический алгоритм Журавлева [7], который дает возможность получить однозначное решение о принадлежности объекта к определенному классу обучающей выборки. В качестве обучающей выборки были использованы ИК-спектры 11 лигнинов, 5 лигноуглеводных комплексов и 5 полисахаридов. Приняв порог $\varepsilon=0$ и используя в качестве меры близости расстояние по Хеммингу, а в качестве ограничений на систему опорных множеств положение о том, что они совпадают с системой всех непустых подмножеств множества рассматриваемых признаков, получим

$$\Gamma_{\text{ЛУК}}(A) > \Gamma_{\text{полисах}}(A) > \Gamma_{\text{лигнин}}(A).$$

Следовательно, алгоритм относит объект А к классу лигноуглеводного комплекса.

Для подтверждения справедливости предположения о принадлежности продукта А к ЛУК продукт А был исследован методом гель-хроматографии. Установлена синхронность кривых распределения лигнина и углеводов, причем эта синхронность сохраняется с повышением температуры хроматографирования до 60°С. Мы полагаем, что этот

¹ Аналогичные результаты получены при использовании для классификации методов факторного анализа (центроидное решение).

факт служит веским аргументом в пользу того, что продукт А является ЛУК [8, с. 291].

Кроме того, следует отметить и тот интересный факт, что продукт А, выделенный из хвойной древесины, был отнесен к комплексам, выделенным из лиственной древесины. Можно предположить, что при окислительной деструкции древесины надуксусной кислотой происходит избирательное окисление или отщепление от лигнина глюкоманнана, а ксилан более устойчив к деструкции. И действительно, из литературы [9] известно, что деструкция полисахаридов надуксусной кислотой приводит к уменьшению содержания глюкоманнана в сахарах, выделенных из оксидата. Учитывая, что вывод о селективном действии надуксусной кислоты сделан только на основании ИК-спектров, можно, по-видимому, отметить, что методы распознавания позволяют выдвигать и гипотезы.

СПИСОК ЛИТЕРАТУРЫ

1. Джурс П., Айзенауэр Т. Распознавание образов в химии. Пер. с англ. М., 1977. 230 с.
2. Жбанков Р. Г. Инфракрасные спектры и структура углеводов. Минск, 1972. 456 с.
3. Сарканен К. В., Людвиг К. Х. Лигнины. Пер. с англ. М., 1975. 632 с.
4. Cunningham K. M., Ogilvie J. G. Evaluation of hierarchical techniques: a preliminary study. — Computer J., 1972, vol. 15, N 3, p. 209—213.
5. Елкина В. Н., Загоруйко Н. Г. Количественные критерии качества таксономии и их использование в процессе принятия решений. — В кн.: Вычислительные системы. Новосибирск, 1969, с. 36—38.
6. Выханду Л. К. Об исследовании многопризнаковых биологических систем. — В кн.: Применение математических методов в биологии. Вып. 3. Л., 1964, с. 19—22.
7. Журавлев Ю. И., Никифоров В. В. Алгоритмы распознавания, основанные на вычислении оценок. — Кибернетика, 1971, № 3, с. 1—11.
8. Рафиков С. Р., Будтов В. П., Монаков Ю. Б. Введение в физикохимию растворов полимеров. М., 1978. 328 с.
9. Glinski A. J., Nichols G. A. Peroxyacetic acid delignification of white birch and new evidence for lignin carbohydrate bonds. — Paperi ja Puu, 1977, t. 59, N 11, s. 745—760.

Поступило 25 XII 1978