Применение метода главных компонент при построении кластерной структуры обучающей выборки молекул.

Григорьева С. С., Кумсков М. И., Захаров А. М.
qsar_msu@mail.ru
Москва, Кафедра Вычислительной математики механико-математического факультета МГУ им.Ломоносова

В работе предложен метод выбора метрик на молекулярных графах на основе главных компонент матрицы «молекула-дескриптор», формирующийся полным перечислением двоек (троек) особых точек на молекулярной поверхности.

Постановка задачи

Задача «структура-свойство» — это задача распознавания образов [1], где объектами являются молекулы, векторное описание которых заранее не задано. Решение этой задачи разбивается на два этапа:

1) этап построения описания обучающей выборки: формируется матрица «молекула-дескриптор» (МД);
2) этап поиска по матрице МД функциональной зависимости.

При представлении пространственной структуры молекул в виде МД матрицы она может получиться очень «широкой», например, в методе CoMFA (Comparative Molecular Field Analysis) [2] или при использовании структурного символьного спектра молекулярных графов [3]. Для преодоления этой проблемы в методе CoMFA линейная регрессия строится на главных компонентах МД матрицы (метод PLS) [4]. В работе используется аналогичный подход при построении функциональной зависимости в виде деревьев решений. Главные компоненты используются при выборе метрики в алгоритме кластер-анализа, а также при построении собственно линейной регрессии на каждом найденном кластере.

Метод решения

Признаками молекулярного графа являются инварианты различных типов. В основе их построения лежит понятие алфавита примитивов описания графов. Мощность алфавита, формируемого по данной обучающей выборке, зависит от определения отношения эквивалентности на элементах примитивов — особых точках (ОТ). Инвариантом первого уровня является число повторений примитивов в графах обучающей выборки; второго уровня — число повторений пар примитивов, находящихся в графе на данном интервале расстояния; третьего (четвертого) уровня — число повторений троек (четверок) примитивов, расположенных в графе на заданных интервалах расстояний.

Рис. 1. Схема построения дерева решений (k — число кластеров).

МД матрица, построенная описанным выше способом, получается очень «широкой», то есть $M \gg N$, где N — количество молекул обучающей выборки, M — мощность алфавита. Метод главных компонент позволяет формировать существенные для прогноза столбцы-факторы матрицы. За основу базовой модели взята линейная множественная регрессия, которая строится на главных компонентах. В силу неоднородности обрабатываемой выборки, ищем зависимость значения активности от значений дескрипторов в виде дерева решений. Мы разбиваем обучающую выборку на кластеры-классы, внутри которых строится функциональная зависимость. Для выделения классов применяем метод кластерного анализа [5].

Перед применением кластерного анализа выделяем главные компоненты, на которых задаем евклидову метрику и формируем матрицу расстояний между молекулами. На полученной матрице расстояний запускаем кластерный анализ и на каждом из «содержательных» кластеров строим линейную регрессию на «кластерных» главных компонентах. Далее, вычисляем коэффициент корреляции скользящего контроля, позволяющий оценить качество прогностической устойчивости полученного дерева решений. Алгоритм (без построения кластеров) был применен к выборкам амбровых одорантов (низкомолекулярных соединений, обладающих амбровым запахом), состоящих из 50 и 129 молекул. Для каждого соединения выборки формировалось 3D-описание соответствующего молекулярного графа — перечислены вершины графа (атомы) с дополнительными атрибутами: символом химического элемента, трехмерными координатами в ангстремах и электрическим зарядом. Матрица МД содержала 703 дескриптора.

Результаты вычислений следующие: на «большой» выборке на 3, 4 и 5 факторах Q^2 (квадрат коэффициента множественной регрессии на скользящем контроле) получился равным соответственно 0.705, 0.74 и 0.66; на

«маленькой» выборке на 3, 4 и 5 факторах Q^2 получился равным соответственно 0.74, 0.76 и 0.74. Таким образом, оптимальным является использование всего четырех факторов.

Работа выполнена при поддержке РФФИ, проект № 07-07-00282.

Литература

[1] *Стьюпер Э., Брюгер У., Джурс П.* Машинный анализ связи химической структуры и биологической активности // М.: Мир, 1982.

[2] *Cramer III R. D., Patterson D. E., Bunce J. D.* Comparative molecular fields analysis (CoMFA) // Effect of shape on binding of steroids to carrier proteins J. Am. Chem. Soc. 110 (1988) 5959-5967. — С. 109–112.

[3] *Кумсков М. И., Смоленский Е. А., Пономарева Л. А., Митюшев Д. Ф., Зефиров Н. С.* Системы структурных дескрипторов для решения задач «структура-свойство» // М.: Доклады Академии Наук, 1994. — С. 336.

[4] *Clark M., Cramer III R. D.,Jones D. M., Patterson D. E., Simeroth P. E.* Comparative Molecular Field Analysis(CoMFA) Toward Its Use with 3D-Structural Databases // Tetrahedron Comput. Methodol. , 1990. — С. 3, 47–59.

[5] *5. Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М.* Многомерный статистический анализ в экономике // Учеб. Пособие для ВУЗов под ред. проф. Тамашевича, М.: ЮНИТИ-ДАНА, 1999.