Секция «Вычислительная математика и кибернетика»

Kernel-методы решения задачи «структура-свойство»

Беккер Александра Владимировна
Студент
Московский государственный университет имени М.В. Ломоносова,
Механико-математический факультет, Москва, Россия
E-mail: bekker_ aleksandra@mail.ru

Поиск количественных корреляций "структура-свойство"химических соединений (QSAR-задача, задача «структура-свойство») [1] – одно из наиболее интенсивно развивающихся в настоящее время направлений теоретической химии. Задача «структура-свойство» заключается в поиске зависимостей между структурой химических соединений и их физико-химическими свойствами и биологической активностью. Решение задачи «структура-свойство» позволяет производить синтез химических соединений с определенными нужными свойствами без лишних затрат времени и средств.

Целью работы является разработка эффективных алгоритмов для анализа баз данных химических соединений и прогнозирования их биологической активности, а также применение и оценка прогностической способности построенных алгоритмов на конкретных выборках химических соединений.

Традиционно решение задачи «структура-свойство» разбивается на несколько этапов, первый из которых заключается в выборе признакового пространства и представлении молекулярных структур в виде векторов значений дескрипторов [2], на втором этапе осуществляется поиск функциональной зависимости между значениями дескрипторов и известной активностью соединений. В результате проведения первого этапа обучающая выборка представляется матрицей «молекула-дескриптор», и таким образом задача «структура-свойство» сводится к классической задаче распознавания образов. Однако в зависимости от выбора признакового пространства размерность матрицы «молекула-дескриптор» может оказаться слишком большой, что затрудняет применение алгоритмов распознавания образов. Для решения этой задачи обычно применяют различные методы отбора информативных дескрипторов.

В работе предложен альтернативный подход к решению задачи, основанный на использовании kernel-методов [3], позволяющих описывать не сами объекты обучающей выборки, а взаимосвязь объектов между собой или меру их сходства. Существенным преимуществом kernel-методов является возможность перехода к признаковым пространствам большей размерности, в которых функциональная зависимость между структурой и «внешним» свойством объекта может быть выражена линейно. При этом переход к пространству большой размерности не приводит к увеличению вычислительной сложности построения модели. Следует выделить два подхода kernel-методов: один из них связан с работой в признаковом пространстве (т.е. после представления объектов в виде векторов значений дескрипторов) и при помощи различных kernel-функций позволяет установить взаимосвязи между объектами; второй подход основан на беспризнаковом сравнении молекулярных структур. Вне зависимости от подхода вся информация об обучающей выборке, необходимая для дальнейшего построения модели, содержится в симметричной квадратной матрице с размерностью, равной размерности обучающей

выборки. Далее к полученной матрице могут применяться kernel-модификации различных алгоритмов анализа данных, таких как метод главных компонент, метод опорных векторов, алгоритмы кластеризации и другие.

В работе исследуется применимость kernel-методов к описанию химических соединений в целях поиска зависимости «структура-свойство», рассматриваются как признаковые, так и беспризнаковые подходы к построению моделей. Проводятся вычислительные эксперименты на выборках химических соединений, в частности, производных молекул бетулина, проявляющих в эксперименте противоопухолевую активность. Структурные формулы и данные о наличии или отсутствии активности веществ извлечены из Базы данных по противоопухолевым веществам Российского Онкологического научного центра им. Н.Н. Блохина РАМН. С помощью построенных алгоритмов удается выделить соединения, предположительно обладающие активностью и рекомендованные для синтеза.

Литература

1. Karelson M. Molecular Descriptors in QSAR/QSPR. Wiley-interscience, 2000
2. Kumskov M.I., Zyryanov I.L., Svitan'ko I.V. A New Method for Representing Spatial Electronic Structures of Molecules in the Problem of Structure-Biological Activity Relationship. Pattern Recognition and Image Analysis, 1995
3. Shawe-Taylor J., Cristianini N. Kernel methods for pattern analysis. Cambridge University Press, 2004.

Слова благодарности

Автор выражает глубокую признательность своему научному руководителю Михаилу Ивановичу Кумскову за постановку задач и постоянное содействие в работе, а также Апрышко Галине Николаевне (Российский онкологический научный центр имени Н.Н. Блохина) и Свитанько Игорю Валентиновичу (Высший Химический Колледж РАН) за предоставленные материалы и плодотворное сотрудничество.