Филиал Московского Государственного Университета им. М. В. Ломоносова в г. Ташкенте

Механико-математический факультет

Ли Александр

Группа М1-06

.

Дипломная работа

Использование топологических индексов на основе матрицы расстояний в решении задач «структура-свойство».

Научный руководитель

д. ф.-м. н.

Кумсков М. И.

Ташкент 2010

Аннотация

В данной работе рассматривается метод решения задачи «структура-свойство» с использованием топологических индексов на основе матрицы расстояний.

В работе предложен, проанализирован и реализован метод решения задачи «структура-свойство» (QSAR-задачи) для молекулярных графов в применении к конкретной выборке молекул бетулинов с целью определения их химической активности/неактивности.

Оглавление

Введение

Одним из наиболее интенсивно развивающихся в настоящее время направлений молекулярного моделирования является поиск взаимозависимостей между структурами химических соединений и их свойствами посредством построения математических моделей. Данная методология получила название QSAR, что означает Quantitative Structure Activity Relationships. Полученные модели используются для скрининга молекулярных баз данных, поиска новых потенциально активных веществ. В развитых странах работы в области QSAR ведутся постоянно возрастающими темпами, так как применение методов QSAR при создании новых соединений с заданными свойствами позволяет значительно сократить затраты на скрининг и осуществлять более целенаправленный синтез соединений, обладающих заданным набором свойств. Так QSAR-задача обеспечивает подход к решению основной задачи химии - синтезу химических соединений с определенными нужными свойствами без лишних затрат времени и денег.

В нашем случае в основу поиска функциональной зависимости между структурой молекулы и ее химической активностью был положен Метод Группового Учета Аргументов (МГУА).

В первой главе данной работы приведена постановка общей задачи «структура-свойство» (QSAR-задачи) как частного случая из области распознавания образов. Вторая глава содержит постановку конкретной задачи, определяет этапы ее решения. И наконец, в третьей главе раскрывается этап построения классификаторов, рассматриваются применяемые методы и схемы реализованных алгоритмов.

Глава 1. Общая постановка задачи «структура-свойство» (QSAR-задачи) для молекулярных графов.

1.1. Методы распознавания образов.

Прежде чем рассматривать саму задачу «структура-свойство», которую предстоит решать, коснемся совокупности математических методов под названием *распознавание образов*. Задача *Quantitative Structure Activity Relationships* является всего лишь одной из частных проблем данной области. Поэтому необходимо рассмотреть, какие существуют методы теории распознавания.

1.1.1. Постановка задачи распознавания образов.

Исходной информацией являются описания объектов, ситуаций, предметов, явлений или процессов S в виде векторов значений признаков $S = (x_1(S), x_2(S), ..., x_n(S))$, где признаки x_i, $i = 1, ..., n$, характеризуют различные стороны-свойства S. У объектов S существует "основное свойство" $y(S)$, которое для части объектов $S_1, S_2, ..., S_m$ предполагается известным, а для части объектов нет. Задача распознавания (прогноза, идентификации, "классификации с учителем") состоит в определении значения свойства $y(S)$ по информации $S_1, S_2, ..., S_m, y(S_1), y(S_2), ..., y(S_m)$ (*обучающей или эталонной выборке*).

Признаки могут быть числовыми (задающими степень выраженности какого-либо свойства), бинарными ("есть" или "нет" свойство), номинальными (обозначающими наличие различных свойств без числовой оценки - пол, цвет, и т.д.).

1.1.2.Основные методы теории распознавания.

1) Алгоритмы распознавания, основанные на вычислении оценок. Распознавание осуществляется на основе сравнения распознаваемого объекта с

эталонными по подмножествам признаков различной мощности, и использовании процедур голосования. Оптимальные параметры решающего правила и процедуры голосования находятся из решения задачи оптимизации модели распознавания: находятся такие значения параметров, при которых точность распознавания является максимальной [4, 5, 6].

2) Алгоритмы голосования по тупиковым тестам. Сравнение распознаваемого объекта с эталонными осуществляется по различным "информативным" подмножествам признаков. В качестве подобных подсистем признаков используются тупиковые тесты (аналоги тупиковых тестов для вещественнозначных признаков) различных случайных подтаблиц исходной таблицы эталонов [4, 5, 6, 7]

3) Алгоритмы голосования по логическим закономерностям. По обучающей выборке вычисляются множества логических закономерностей каждого класса - наборы признаков и интервалы их значений, свойственные каждому классу. При распознавании нового объекта вычисляется число логических закономерностей каждого класса, выполняющихся на распознаваемом объекте. Каждое отдельное "выполнение" считается "голосом" в пользу соответствующего класса. Объект относится в тот класс, нормированная сумма "голосов" за который является максимальной. Настоящий метод позволяет оценивать веса признаков, логические корреляции признаков, строить логические описания классов, находить минимальные признаковые подпространства [8, 9, 10].

4) Алгоритмы статистического взвешенного голосования. По данным обучающей выборки находятся статистически обоснованные логические закономерности классов. При распознавании новых объектов вычисляется оценка вероятности принадлежности объекта к каждому из классов, которая является взвешенной суммой "голосов" [6, 11, 12, 13, 14].

5) Линейная машина. Для каждого класса находится некоторая линейная функция. Распознаваемый объект относится в тот класс, функция которого принимает максимальное значение на данном объекте. Оптимальные линейные функции классов находятся в результате решения задачи поиска максимальной совместной подсистемы системы линейных неравенств, которая формируется по обучающей выборке - находится специальная кусочно-линейная поверхность, правильно разделяющая максимальное число элементов обучающей выборки [15, 16].

6) Линейный дискриминант Фишера. Классический статистический метод построения кусочно-линейных поверхностей разделяющих классы. Благоприятными условиями применимости линейного дискриминанта Фишера являются выполнение следующих факторов: линейная отделимость классов, дихотомия, "простая структура" классов, невырожденность матриц ковариаций, отсутствие выбросов [15].

7) Метод *k* ближайших соседей. Классический статистический метод. Распознаваемый объект относится в тот класс, из которого он имеет максимальное число соседей. Оптимальное число соседей и априорные вероятности классов оцениваются по обучающей выборке [15].

8) Нейросетевая модель распознавания с обратным распространением. Модификация метода обучения нейронной сети распознаванию образов (метод обратного распространения ошибки). В качестве критерия качества текущих параметров нейронной сети используется гибридный критерий, учитывающий как сумму квадратов отклонений значений выходных сигналов от требуемых, так и количество ошибочных классификаций на обучающей выборке [17].

9) Метод опорных векторов (support vector machine). Метод построения нелинейной разделяющей поверхности с помощью опорных векторов. В новом признаковом пространстве (спрямляющем пространстве) строится линейная разделяющая поверхность. Построение данной поверхности сводится к решению задачи квадратичного программирования [18].

10) Алгоритмы решения задач распознавания коллективами различных распознающих алгоритмов. Задача распознавания решается в два этапа. Сначала применяются независимо различные алгоритмы теории распознавания образов. Далее находится автоматически оптимальное коллективное решение с помощью специальных методов-"корректоров" [4, 5, 19, 20].

11) Методы кластерного анализа.

- алгоритмы иерархической группировки;
- кластеризация по методу минимизации суммы квадратов отклонений;
- метод *k*-средних.

Возможно решение задачи классификации как при заданном, так и неизвестном числе классов [3,15].

12) Алгоритм построения коллективных решений задачи классификации. Задача классификации решается в два этапа. Сначала находится набор различных решений (в виде покрытий или разбиений) при фиксированном числе классов с помощью различных алгоритмов теории распознавания образов. Далее находится оптимальная коллективная классификация в результате решения специальной дискретной оптимизационной задачи [21].

1.2. Quantitative Structure Analysis Relationships.

1.2.1. Общая постановка QSAR-задачи, основные определения и термины.

Дадим определение так называемого QSAR/QSPR-анализа (QSAR/QSPR - Quantitative Structure Analysis/Property Relationships, задача «структура-свойство»). Он является самым распространенным методом установления количественных соотношений между структурой и активностью соединений и представляет собой статистический подход к проблеме.

Определение: **Меченый молекулярный граф** $G = \{E, V\}$ – помеченный граф, вершины которого интерпретируются как атомы молекулы, а ребра – как валентные связи между парами атомов. Метки вершин и ребер (числа или символы) кодируют атомы и связи различной химической природы. В качестве меток вершин могут быть использованы любые характеристики соответствующих атомов (например, трехмерные координаты, символ химического элемента, заряд ядра, поляризуемость, атомный вес, атомный радиус и др.), а в качестве меток ребер – любые характеристики соответствующих связей (кратность, длины, порядки связей, полученные из квантово-химических расчетов, и т.д. [22]).

Определение: (**задача «структура-свойство»**): Пусть задана *обучающая (или эталонная) выборка* - база данных из N химических соединений, где:

1) i-ое соединение представлено меченым молекулярным графом G_i, имеющим укладку в трехмерном пространстве (т.е., для каждой вершины в качестве меток заданы ее трехмерные координаты);
2) либо i-ое соединение отнесено к C_i - одному из K классов активности (например, «активных», «слабоактивных», «неактивных» веществ) согласно

исследуемому свойству, либо для него задано численное значение исследуемого свойства A_i.

Необходимо построить классифицирующую функцию F, получающую в качестве аргумента произвольный молекулярный граф с метками того же типа, и «наилучшим образом» относящую это соединение к одному из классов активности, либо «наилучшим образом» предсказывающую численное значение исследуемого свойства.

Какая из классифицирующих функций «лучше», позволяет определить *функционал качества φ(F)*. Например, в качестве функционала качества можно использовать процент верно классифицированных функцией F молекул из обучающей выборки:

$$\varphi(F) = 1 - \frac{\sum_{i=1}^{N} \varepsilon_i}{N}, \text{ где } \varepsilon_i = \begin{cases} 0, \text{ åñëè } F(G_i) = C_i \\ 1, \text{ â ïðîòèâíîì ñëó÷àå} \end{cases} \quad (1)$$

или, в случае, когда функция должна предсказывать численное значение свойства,

$$\varphi(F) = 1 - \frac{\sum_{i=1}^{N} (F(G_i) - A_i)^2}{\sum_{i=1}^{N} A_i^2}. \quad (2)$$

Поставленную таким образом задачу поиска классифицирующей функции будем называть **задачей «структура-свойство»** или **QSAR-задачей.**

Определение: **Дескриптором** будем называть какое-либо свойство, численное значение которого может быть вычислено для произвольного молекулярного графа G.

Определение: **Алфавитом дескрипторов** будем называть множество всех дескрипторов, используемых для анализа обучающей выборки, обозначенных различными символьными метками.

Определение: Пусть алфавит дескрипторов состоит из M элементов. **Вектором признаков** молекулярного графа G будем называть вектор $\bar{x} = (x_1,..,x_M) \in R^M$, где x_j-значение j-ого дескриптора, вычисленное для G.

Определение: **Матрицей «молекула-признак» (матрицей признаков)** для рассматриваемой обучающей выборки будем называть матрицу размера N x M, в i-ой строке которой стоит вектор признаков $\bar{x}_i = (x_{i1},..,x_{iM})$ i-ого соединения.

1.2.2. Основные этапы решения QSAR-задачи.

В вышеописанных терминах задача «структура-свойство» разбивается на две части:

1) этап описания**:**

Исходя из формата молекулярных графов (типа меток вершин и ребер) выбирается алфавит дескрипторов *A*. На основе этого алфавита строится отображение из множества молекулярных графов в признаковое пространство R^M и формируется матрица «молекула-признак» для обучающей выборки.

2) этап поиска модели функциональной зависимости:

В результате анализа матрицы «структура-свойство» на признаковом пространстве строится модель функциональной зависимости - классифицирующая функция *F* с наилучшей прогностической способностью, т.е. с наибольшим значением функционала качества.

1.2.3. Этап описания молекулярного графа.

1) Проводится дополнительная классификация атомов (вершин молекулярного графа) на основе их локальных свойств (заряда, эксцентриситета вершины, каких-либо топологических свойств). В результате этого метка каждой вершины заменяется на другую, содержащую информацию о локальных свойствах.
2) В молекулах выбираются структурные фрагменты (атомы, цепочки связанных атомов, группы атомов).
3) Каждому структурному фрагменту сопоставляется символьное имя - тип фрагмента (например, если рассматриваются цепочки атомов, то «именем» цепочки может служить объединение символьных меток входящих в нее атомов).
4) Множества фрагментов для всех молекулярных графов выборки объединяются.
5) Для каждого молекулярного графа и каждого фрагмента находим значение соответствующего структурного дескриптора (либо количество повторений, либо наличие/отсутствие в молекулярном графе).

В итоге, получаем матрицу «молекула-признак», состоящую из «структурных спектров» молекул.

1.2.4. Этап поиска функциональной зависимости.

Напомним, что после формирования матрицы «структура-свойство» для обучающей выборки необходимо построить классифицирующую функцию $F(x_1, x_2, ..., x_M)$, где $(x_1, x_2, ..., x_M)$ – вектор признаков молекулярного графа. Причем, построенная функция должна обеспечивать лучшее значение функционала качества.

Обычно вид классифицирующей функции F заранее задается (например, функция может быть линейной, квадратичной и др.) и зависит от ряда параметров, которые определяются по обучающей выборке соединений. Чаще всего в качестве F используется линейная функция. Получаемое уравнение называют **линейной регрессионной моделью**.

Заметим, что полученная формулировка задачи на этапе поиска функциональной зависимости полностью совпадает с формулировкой задачи распознавания. Поэтому для нахождения классифицирующей функции F можно использовать любые методы распознавания и классификации, описанные в начале главы.

Следует отметить, что в задаче «структура-свойство» число дескрипторов M, как правило, значительно превышает число молекул в обучающей выборке ($M >> N$), что затрудняет анализ матрицы «молекула-признак». Для того чтобы сократить число дескрипторов, необходимо рассматривать лишь наиболее **информативные** из них, т.е. те, которые потенциально будут значимы при построении классифицирующей функции на признаковом пространстве. Это можно проделать как на этапе описания (например, при эволюционном формировании дескрипторов), так и на этапе анализа матрицы признаков. В данной работе предлагается отбирать наиболее информативные дескрипторы путем взаимодействия этих двух этапов – использования результатов этапа анализа на этапе описания.

1.3. Выводы.

QSAR-задача (задача «структура-свойство») представляет собой одну из задач области распознавания образов. Поэтому для ее решения могут быть применены все методы решения задач теории распознавания.

Глава 2. Постановка конкретной задачи для выборки молекул .

2.1. Определение и виды топологических индексов

Определение: **топологическая матрица** $A = (a_{ij})$ меченого молекулярного графа строится следующим образом: в ней элемент a_{ii} – это метка i-ой вершины, a_{ij} – метка ребра (i, j).

Очевидно, что при изменении нумерации вершин графа получается, вообще говоря, другая матрица. Это является существенным недостатком описания химических структур в терминах матриц. Эта проблема привела к термину «инвариант графа».

Определение: **инвариант графа** – это такое число (или функция, если метка графа - символы), вычисляемое по матрице графа А, которое не зависит от нумерации вершин графа.

Определение: молекулярный граф называется **простым** в следующем случае:

- метки вершин равны нулю
- если атомы связаны химической связью, то им сопоставляется ребро с меткой «1», в противном случае метка ребра равна нулю,

т.е. простой граф отображает только наличие связей между вершинами.

Определение: **топологическими индексами (ТИ)** называют инварианты простых графов. Часто это определение переносится и на инварианты меченых графов, которые могут отражать не только топологию молекулы, но и элементы электронного и пространственного строения.

Топологические индексы и широко используются как дескрипторы при решении задачи «структура-свойство». Популярность данного подхода к описанию молекулярной структуры связана с простотой и быстротой вычисления ТИ, возможностью учитывать при их построении элементы электронного и пространственного строения, я также наличием огромного количества удачных корреляций вида «ТИ – свойство». Однако такой подход имеет и очевидный недостаток: он не позволяет различать разные конфигурации молекул и не учитывает их конформационные особенности.

Ниже приведены примеры некоторых наиболее популярных ТИ.

Индекс Рандича χ определяется по следующей формуле:

$$\chi = \Sigma(v_i v_j)^{-1/2},$$

где v_i - степень i-ой вершины графа; суммирование проводится по всем ребрам графа.

Индекс χ был обобщен Киром и Холлом, которые ввели определение так называемого *индекса связности m-ого порядка (индекс Кира-Холла)*

$$^{m}\chi^{v} = \Sigma(\delta_i\delta_j\delta_k\ldots)^{-1/2},$$

$m \geq 1$, $\delta = (Z_i^v - h_i) \ / \ (Z_i - Z_i^v - 1)$, где Z_i – общее число, а Z_i^v - чтсло валентных электронов i-ого атома; h_i – число атомов водорода, связанных с i-ым атомом; суммирование происходит по всем цепочкам, состоящим из m ребер графа; i, j, k – номера вершин, образующих соответствующую цепочку.

Индекс Винера W определяется по следующей формуле:

$$W = 0.5\ \Sigma d_{ij},$$

где d_{ij} – кратчайшее расстояние между i-ой и j-ой вершинами. Матрица $D = (d_{ij})$ называется **матрицей топологических расстояний**.

Индекс Хосойя Z определяется формулой:

$$Z = \Sigma p_k,$$

где p_k – число способов выбрать в графе k ребер ($k \geq 1$) так, что никакие два из них не являются смежными.

Индекс Балабана (для ациклических графов (деревьев)) определяется по следующей формуле:

$$B = \Sigma r_i^2,$$

где r_i - число вершин дерева, отсекаемых на i-ом шаге процедуры обрезки деревьев. Данная процедура состоит из последовательного удаления вершин степени 1 исходного дерева и инцедентных им ребер.

2.2. Описание входных данных.

В нашем случае перед группой студентов была поставлена QSAR-задача (задача «структура-свойство»).

Была предоставлена выборка, состоящая из 50 молекул. Для каждого соединения выборки было дано описание соответствующего молекулярного графа в отдельном файле с расширением *.mol.* В данном файле перечислены вершины графа (атомы) с дополнительными атрибутами: символом химического элемента, трехмерные координаты в ангстремах и электрический заряд. В файле с расширением *.sdf* кроме уже описанной информации были представлены сведения о биологической активности соединений

2.3.Поставленная задача.

Как и в любой QSAR-задаче, в данном случае было необходимо найти способ описания молекулярных графов, а потом по построенной матрице «структура-свойство» построить некоторую модель классификации (классифицирующую функцию) с лучшим качеством прогноза.

Этап описания молекулярных графов разделяется на 2 крупных шага:

- этап формирования особых точек
- этап формирования алфавита дескрипторов и построения матрицы признаков

2.4. Этап формирования особых точек.

В данной работе дескрипторы формируются не на основе исходных молекулярных графов, а на основе производных из них графов, в вершинах которых находятся так называемые ***особые точки.***

Построение особых точек

1. На основе исходных молекулярных графов для каждой молекулы строится матрица топологических расстояний.
2. Пары вершин разбиваются на группы, в зависимости от расстояния между ними.
3. Таким образом в качестве особых точек получаем фрагмент вида (A,B,d), где A- первая вершина, B-вторая вершина, а d-группа по расстоянию

2.5. Этап формирования дескрипторов и построение матрицы «структура –свойство».

Полученные выше фрагменты формируют молекулярные дескрипторы. Далее формируем матрицу «структура-свойство» A следующим образом:

1. Строками матрицы A являются молекулы из выборки.
2. Столбцы матрицы A представляют собой дескрипторы.
3. Элемент a_{ij} показывает сколько раз j–й дескриптор встречается в i-ой молекуле.

2.6. Выводы.

Таким образом, на входе анализа была предоставлена матрица «структура-свойство» и вектор значений химической активности. Анализируя матрицу и результирующий вектор необходимо построить прогнозирующую модель в виде линейной прогнозирующей функции

Глава 3. Анализ матрицы «структура-свойство» и построение прогнозирующей функции.

3.1. Постановка задачи.

Имея в наличии матрицу из значений сформированных дескрипторов для каждой молекулы выборки и вектор значений биологической активности соединений, перейдем к решению основной задачи - нахождению метода распознавания новых поступающих на вход молекулярных графов и прогнозированию его свойств. Необходимо найти некоторую зависимость, функцию $\varphi = \varphi(x_1, x_2, \ldots, x_m)$ (где $x_1, x_2, \ldots, x_m$ – значения отобранных дескрипторов), которая приближала бы свойства объектов, представленных молекулярными графами.

Чаще всего для нахождения таких зависимостей используются линейные функции φ. Поэтому в данной работе за основу взят **линейный вид классификаторов**. В результате для решения задачи построения прогнозирующей функции необходимо решить систему линейных уравнений

$$X \times \bar{a} = \bar{y},$$

где X – матрица N x M (N – количество объектов обучающей выборки, M – количество выявленных признаков объектов), a – искомый вектор коэффициентов линейной функции, y – вектор истинных свойств объектов (биологической активности). Как показано в предыдущей главе, матрица получается очень «широкой», то есть $M >> N$. Можно, конечно, решать эту систему и находить зависимости для таких матриц полным перечислением, но при этом возникает несколько трудностей:

- Технически сложно держать и работать с таким количеством памяти.
- В стандартных пакетах анализа данных модулей для обработки таких «широких» таблиц отсутствует.

Для преодоления этих трудностей зачастую используют **генетический (эволюционный) алгоритм Метода Группового Учета Аргументов (МГУА)**, позволяющий динамически подбирать самые существенные для прогноза столбцы матрицы, не рассматривая остальные столбцы. Именно его мы применяем для нахождения линейной функциональной зависимости.

3.2. Метод Группового Учета Аргументов.

3.2.1. Обзор метода, его преимущества.

Метод Группового Учета Аргументов применяется в самых различных областях для анализа данных и отыскания знаний, прогнозирования и моделирования систем, оптимизации и распознавания образов. Индуктивные алгоритмы МГУА дают уникальную возможность автоматически находить взаимозависимости в данных, выбрать оптимальную структуру модели или сети, увеличить точность существующих алгоритмов.

Этот подход самоорганизации моделей принципиально отличается от обычно используемых дедуктивных методов. Он основан на индуктивных принципах - нахождение лучшего решения основано на переборе всевозможных вариантов.

При помощи перебора различных решений подход индуктивного моделирования пытается минимизировать роль предпосылок автора в результатах моделирования. Алгоритм сам определяет структуру модели и законы, действующие в объекте. Он может быть использован как советчик для отыскания новых решений в проблемах искусственного интеллекта.

Метод Группового Учета Аргументов состоит из нескольких алгоритмов для решения разных задач. В него входят как параметрические, так и алгоритмы кластеризации, комплексирования аналогов, ребинаризации и вероятностные алгоритмы. Этот подход самоорганизации основан на переборе постепенно усложняющихся моделей и выборе наилучшего решения согласно минимуму внешнего критерия. В качестве базисных моделей используются не только линейные функции, как в нашем случае, но и полиномы, а также нелинейные, вероятностные функции или кластеризации.

3.2.2. Общая схема алгоритма МГУА.

Пусть заданы:

- матрица «объект-признак» X, состоящая из N строк и M столбцов;
- *результирующий вектор* - вектор y, который необходимо «предсказать» по матрице X;
- класс функций C для построения классифицирующей функции;

- *функционал качества* $\varphi(F, X, y)$, $F \in C$;
- количество отбираемых на каждом этапе наилучших функций Q;
- условие остановки алгоритма (максимальное число используемых переменных в классификаторе или предельный уровень функционала качества).

Тогда эволюционный алгоритм МГУА состоит из следующих этапов:

1) Перебором строим всевозможные функции $F \in C$ от одной переменной (столбцов матрицы X).

2) Оцениваем множество всех построенных на данный момент функций при помощи функционала качества φ и выбираем из них Q наилучших (данный набор функций, отобранных на определенном этапе, называется *селекцией*), причем для каждого шага число Q может иметь отдельное значение.

3) Перебором присоединяем к каждой из отобранных функций новую переменную (в рамках класса функций C), или функцию от одной переменной из числа отобранных (зависит от реализации), определенным методом формируем классифицирующие функции новой селекции.

4) Проверяем, достигнуто ли условие остановки алгоритма; если оно не достигнуто, переходим к п. 2)

Таким образом, алгоритмы МГУА воспроизводят схему массовой селекции. На каждой селекции классифицирующие модели усложняются и отбираются лучших из них. Прогнозирующая функция (*прогнозирующая модель*) $\varphi = \varphi(x_1,x_2,x_3,...,x_m)$, где φ — некоторая элементарная функция, в нашем случае линейная, заменяется несколькими рядами "частных" описаний:

1-ый ряд селекции: $y_1^{(1)} = f(x_1, x_2)$, $y_2^{(1)} = f(x_1, x_3)$,..., $y_s^{(1)} = f(x_{m-1}, x_m)$,

2-ой ряд селекции: $y_1^{(2)} = f(x_1, y_1^{(1)})$, $y_2^{(2)} = f(x_1, y_2^{(1)})$, ..., $y_p^{(2)} = f(x_m, y_s^{(1)})$, где $s=m*(m-1)/2$, $p=s*m$ и т.д.

3.2.3. Случай линейных классифицирующих функций.

В нашем случае C представляет собой класс линейных функций от нескольких переменных. Тогда на каждом шаге селекции мы строим линейные прогностические модели на основе уже построенных добавлением одной переменной, а затем из них с помощью функционала качества отбираем лучшие. Когда же условие остановки

алгоритма будет выполняться, рассматривается лучшая по функционалу качества прогностическая модель, в которой, переходя обратно от последних селекций к первой, раскрывая скобки, можно получить прогностическую линейную функцию φ от большого числа переменных. Для полного определения алгоритма необходимо предоставить правило построения новой функции от двух переменных (см. шаг 3) и функционал качества.

Алгоритм построения линейной функции от двух переменных.

В качества данного алгоритма используется **линейная регрессия**.

Линейная регрессия представляет собой самый простой метод построения линейной прогностической модели и работает следующим образом.

Пусть X – матрица переменных, которые влияют на значения y - зависимых переменных. В нашем случае X – один или два столбца матрицы признаков, y – тот же результирующий вектор активности. По этим данным требуется построить прогнозирующую функцию $F(x)$, где x – строка матрицы X, соответствующая одному объекту (наблюдению). Прогнозирующую функцию ищем в виде:

$F(x) = a + xb = z\beta$, где a – свободный член, b – вектор-столбец коэффициентов при переменных прогнозирующей функции, β – вектор-столбец, полученный конкатенацией векторов a и b, z – вектор-строка, полученный из x путем добавления 1 в начало вектора. Из векторов z сформируем матрицу Z, порожденную матрицей X:

$Z = [I_{N,1}\ X]$,

где $I_{N,1}$ – вектор размерностью N x 1, состоящий из единиц, а квадратные скобки обозначают конкатенацию матриц.

Для нахождения значений вектора β используется метод наименьших квадратов. Находим такие коэффициенты прогноза, чтобы величины ошибок прогноза

$\varepsilon_i = y_i - F(x_i) = y_i - z_i \beta$,

называемые *остатками*, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - z_i \beta)^2 = \min \quad \beta$$

Легко показать, что решение данной задачи имеет вид:

$\beta = (Z^T Z)^{-1} Z^T y$, где Z – матрица, порожденная X.

Откуда вектор остатков (ошибок прогноза) равен:

$\varepsilon = y - Z\beta = y - Z(Z^T Z)^{-1} Z^T y = (E - Z(Z^T Z)^{-1} Z^T)\, y$,

где E – единичная матрица размерности N x N.

Функционал качества.

Обычно в случае линейного классификатора роль функционала качества играет величина **среднеквадратической ошибки** – нормированная сумма квадратов отклонений полученного прогноза от исходного результирующего вектора.

Отклонение прогноза *i*-ого явления равно

$\varepsilon_i = | y_i - a - x_i \times b |$,

где a - свободный член, b – вектор-столбец линейной прогнозирующей модели, x_i – *i*-ая строка матрицы X.

Среднеквадратическая ошибка вычисляется по формуле:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \varepsilon_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2},$$

где $\bar{y}$ - среднее значение элементов вектора $y = (y_1, y_2, \ldots, y_N)$.

В результате проведения селекций значение R^2 увеличивается. С одной стороны, необходимо получить значение R^2, близкое к единице (когда ошибка минимальна), с другой – как можно более простую прогнозирующую модель (т. е. на более раннем этапе селекции).

3.3. Алгоритм решения задачи.

В итоге для нахождения прогнозирующей функции был предложен алгоритм, состоящий из следующих шагов:

1) Обработка обучающей выборки и формирование дескрипторов.
2) Построение матрицы «молекула-свойство».
3) Запуск МГУА на полученной матрице.

3.4. Выводы

На основе стандартных методов распознавания образов был предложен алгоритм построения линейной прогнозирующей функции.

Заключение

В работе разработан и описан метод поиска функциональной зависимости химической активности от значений дескрипторов в виде линейной прогнозирующей функции.

Построение прогнозирующей функции производилось на основе матрицы «молекула-свойство» с помощью Метода Группового Учета Коэффициентов.

Можно выделить следующие основные направления развития данной работы:

1) Использование матрицы геометрических расстояний вместо матрицы геометрических расстояний и сравнение результатов.
2) Использование фактор-анализа вместо МГУА для построения прогнозирующей функции.
3) Использование дескрипторов более высокого порядка для построения МД матрицы и дальнейшего анализа.

Список литературы.

1. Журавлев Ю.И., Рязанов В.В., Сенько О.В. *«Распознавание». Математические методы. Программная система. Практические применения.* - М.: ФАЗИС, 2006.

2. Скворцова М.И., Станкевич И.В., Палюлин В.А., Зефиров Н.С. *Концепция молекулярного подобия и ее использование для прогнозирования свойств химических соединений*. (в публикации)

3. Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М. *Многомерный статистический анализ в экономике.*

4. Журавлев Ю. И. *Избранные научные труды.* - М.: Магистр, 1998.

5. Журавлев Ю. И. *Об алгебраическом подходе к решению задач распознавания и классификации.* - М.: Наука, 1978, вып. 33.

6. Ryazanov V.V., Sen'ko O.V., Zhuravlev Yu.I. *Mathematical Methods for Pattern Recognition: Logical, Optimization, Algebraic Approaches.* – Proceedings of the 14th International Conference on Pattern Recognition. Brisbane, Australia, August 1998.

7. Дмитриев А. Н., Журавлев Ю. И., Рязанов В.В., Чернявский Г.М. *Разработка системы оперативного прогнозирования сельскохозяйственного урожая на территории РФ.* – В кн.: Доклады 10-ой Всероссийской конференции «Математические методы распознавания образов (ММРО-10)» - Москва, 2001.

8. Ryazanov V.V. *Recognition Algorithms Based on Local Optimality Criteria.* – Pattern Recognition and Image Analysis, 1994, vol. 4, no. 2.

9. Ryazanov V.V. *About some approach for automatic knowledge extraction from precedent data.* - Proceedings of the 7th International Conference «Pattern Recognition and Image Processing», Minsk, May 21-23, 2003, vol. 2.

10. Ryazanov V.V., Vorotnichkin V.A. *Discrete Approach for Automatic Knowledge Extraction from Precedent Large-scale Data, and Classification.* – Proceedings of the 16th International Conference on Pattern Recognition. Quebec, Canada, 11-15 August 2002.

11. Кузнецов В.А., Сенько О.В., Кузнецова А.В. и др. *Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической нормы и хронической патологии.* – Химическая физика, 1996, 15(1).

12. Сенько О.В. *Использование процедуры взвешенного голосования по системе базовых множеств в задачах прогнозирования.* – Журнал вычислительной математики и математической физики, 1995, 35(10).

13. Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabotina T.N., Korotkova O.V. *The prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges.* – Journal Theoretical Medicine, 2000, vol. 2.

14. Ryazanov V.V., Sen'ko O.V., Zhuravlev Yu.I. *Methods of Recognition and Prediction Based on Voting Procedures.* – Pattern Recognition and Image Analysis, 1999, vol. 9, no. 4.

15. Дуда Р., Харт П. *Распознавание образов и анализ сцен.* – М.:Мир, 1976.

16. Обухов А.С., Рязанов В.В. *Применение релаксационных алгоритмовпри оптимизации линейных решающих правил.* – В кн.: Доклады 10-ой Всероссийской конференции «Математические методы распознавания образов (ММРО-10)» - Москва, 2001.

17. Уоссермен Ф. *Нейрокомпьютерная техника.* – М.:Мир, 1992.

18. Christopher J.C. Burges *A Tutorial on Support Vector machines for Pattern Recognition.* - Data Mining and Knowledge Discovery 2, 1998.

19. Kuncheva L.I. *Combining pattern classifiers^ Methods and Algorithms.* – Wiley, 2004.

20. Vetrov D.P. On *the Stability of the Pattern Recognition Algorithms.* – Pattern Recognition and Image Analysis, 2003, vol. 13, no. 3.

21. Рязанов В.В. *О построении оптимальных алгоритмов распознавания и таксономии (классификации) при решении прикладных задач.* – В кн.: Распознавание, классификация, прогноз: Математические методы и их применение. – М.: Наука, 1998, вып. 1.

22. O.Ivanciuc, S.L.Taraviras, D.Cabrol-Bass - Journal of Chemical Information and Computer Sciences, 40

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

24. В. Магнусон, Д. Харрис, С. Бейсак В кн. *Химические приложения топологии и теории графов.* (Ред. Р. Кинг) - М.:Мир, 1987.

25. Станкевич М.И., Станкевич И.В., Зефиров Н.С. *Успехи химии,* 57.

26. Makeev G.M., Kumskov M.I., Svitan'ko I.V., Zyryanov I.L. *Recognition of Spatial Molecular Shapes of Biologically Active Substances for Classification of Their Properties.* Pattern Recognition and Image Analysis, 1996, v.6, n.4.