Филиал Московского государственного университета
имени М.В. Ломоносова в г.Ташкенте

Мирвалиев Руслан Маратович

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему «Использование эволюционных алгоритмов для построения кусочно-линейных классификаторов свойств молекул»
на соискание степени бакалавра
по направлению 010500 - «Прикладная математика и информатика»

ВКР рассмотрена и рекомендована к защите	Научный руководитель: д. ф.-м. н.,профессор
_____	_____ Кумсков М.И.
«___»_____2010 год	«___»_____2010 год

Ташкент-2010

В работе решена задача «структура-свойство» для молекулярных графов в применении к производным бетулиновой кислоты, химическим соединениям, обладающими противоопухолевой активностью (данные получены из Российского онкологического научного центра имени Н. Н. Блохина).

В качестве признаков молекул выбраны топологические и пространственные дескрипторы. Первые используются для разбиения обучающей выборки на кластеры, формирования правила отказа от прогноза и правила принадлежности к кластеру. Вторые – для построения кусочно-линейной функции прогнозирования активности веществ с помощью эволюционного алгоритма МГУА.

Все этапы построенной модели реализованы в системе Matlab. При этом использовались как основной пакет программ, так и дополнительный пакет (Toolbox) Statistics, предлагающий широкий выбор инструментов для статистического исследования.

Проведены вычислительные эксперименты на полученной выборке из 52 молекул с заданными значениями активностей.

Содержание

Глава 4. Особенности программной реализации, проведение расчетов

4.1. Архитектура программы

4.2. Бетулин, фармакологическая активность его производных

4.3. Проведение вычислительных экспериментов

Выводы

Заключение

Список литературы

Введение.

Задача поиска функциональной зависимости «структура-свойство» (Quantitative Structure-Activity Relationship, QSAR-задача) или задача прогнозирования физико-химической или биологической активности вещества является актуальной проблемой математической химии[1]. Математические модели «структура-свойство» широко используются на практике, как для предсказания активности веществ, так и для поиска новых соединений с заданными химико-биологическими свойствами. Данные модели позволяют значительно сократить расходы и время, необходимое для исследований, при синтезе новых соединений с заданными свойствами.

Целью работы является разработка алгоритма построения кусочно-линейных классификаторов и области их определения для заданной обучающей выборки молекулярных графов (М-графов).

Новизной работы является использование различных типов дескрипторов на этапе выбора кластера и на этапе построения модели в кластере, формирование правил принадлежности к кластеру и правил отказа от прогноза для идентификации выбросов.

В работе М-графы представлены структурными дескрипторами[2], характеризующими наличие и взаимное расположение заданных элементов описания структурных фрагментов. В качестве таких элементов описания для исследования выбраны цепочки атомов в М-графе.

С использованием топологических структурных фрагментов обучающая выборка разбивается на кластеры. Осуществляется построение линейной функции прогнозирования на каждом кластере. При появлении нового М-графа осуществляется поиск ближайшего кластера. В случае нахождения прогнозируем активность соединения по функции, построенной на данном кластере, иначе осуществляем отказ от прогнозирования.

Особенностью МД-матрицы является очень большое число дескрипторов, значительно превышающее число молекул. Поэтому в работе используется метод группового учета аргументов[3]. Все этапы построения функции-классификатора по обучающей выборке реализованы в среде MATLAB. Проведены расчеты на реальных соединениях, обладающих противоопухолевой активностью (данные Российского онкологического научного центра имени Н. Н. Блохина).

В первой главе работы приведена постановка общей задачи «структура-свойство» как частного случая из области распознавания образов, рассмотрены основные методы распознавания образов, а также описаны основные этапы решения этой задачи. Вторая глава содержит базовые определения, используемые в работе, схему формирования МД-матрицы на основе кодированных цепочек атомов. В третьей главе описан Метод Группового Учета Аргументов, используемый для построения линейной прогнозирующей функции. В четвертой главе описана архитектура написанной программы, ее основные модули, а также алгоритмы, используемые в них, дан краткий обзор химическим соединениям выборки, исследуемой в работе.

Глава 1. Общая постановка задачи «структура-свойство»

1.1. Постановка задачи, основные определения

Меченый молекулярный граф G = {E, V} – граф, вершины которого интерпретируются как атомы молекулы, а ребра – как валентные связи между парами атомов[4]. 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.). Атом водорода не считается вершиной М-графа - "водород стерт".

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 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:

$$\varphi(F) = 1 - \frac{\sum_{i=1}^{N} \varepsilon_i}{N}, \text{ где } \varepsilon_i = \begin{cases} 0, \textit{если } F(G_i) = C_i \\ 1, \textit{в противном случае} \end{cases},$$

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

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

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

1.2.Сведение задачи «структура-свойство» к классической задаче распознавания образов

Характерной особенностью QSAR-задачи является представление анализируемых объектов в виде М-графов, а не в виде заранее фиксированного вектора признаков. Это приводит к постановке задачи формирования такого пространства признаков, в котором может быть получено наилучшее решение. Признаки, выбираемые химиком-экспертом, ориентированы на анализ конкретного свойства в заданном химическом ряду соединений. Выбор описания молекул в виде вектора признаков является ключевым моментом проведения QSAR-моделирования, поскольку этот выбор сводит задачу построения QSAR уравнения к классической задаче распознавания образов со стандартной информацией [5].

Задана обучающая выборка, состоящая из N молекулярных графов (М-графов) и представляющая собой список пар $\{(X_1,C_1),(X_2,C_2),...,(X_N,C_N)\}$, где i-й М-граф представлен в виде вектора-строки $X_i=(X_{i1},X_{i2},...,X_{iM})$ значений выбранных признаков; C_i - внешний признак, задающий значение активности. Требуется построить классифицирующую функцию *F* такую, что на объектах обучающей выборки ее значения приближают заданную внешнюю классификацию С, минимизируя сумму квадратов ошибок:

$$F(X_i) = C_i + e_i; \quad \Sigma e_i^2 \rightarrow \min ; \quad i=1,\ldots, N; \qquad (1)$$

Уравнение (1) используется для классификации новых М-графов, принадлежащих области определения функции F.

1.2.1. Основные методы распознавания образов

Опишем основные методы распознавания исследуемых объектов после их представления в виде векторов признаков.

Метод k ближайших соседей - классический статистический метод[6]. При классификации неизвестного объекта находится k геометрически ближайших к нему в пространстве признаков других объектов (ближайших соседей) с уже известной принадлежностью к распознаваемым классам. Решение об отнесении неизвестного объекта к тому или иному классу принимается, например, с помощью простого подсчета голосов.

Алгоритмы вычисления оценок. Принцип действия алгоритмов вычисления оценок состоит в вычислении приоритетов (оценок сходства), характеризующих “близость” распознаваемого и эталонных объектов по системе признаков. Оптимальные параметры решающего правила и процедуры вычисления оценок находятся из решения задачи оптимизации модели

распознавания: находятся такие значения параметров, при которых точность распознавания является максимальной.[7,8]

Метод опорных векторов (SVM — support vector machines) — набор схожих алгоритмов вида «обучение с учителем», использующихся для задач классификации и регрессионного анализа. Этот метод принадлежит к семейству линейных классификаторов. Основная идея метода опорных векторов — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей наши классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей[9].

Деревья принятия решений обычно используются для решения задач классификации данных или, иначе говоря, для задачи аппроксимации заданной булевой функции. Ситуация, в которой стоит применять деревья принятия решений, обычно выглядит так: есть много случаев, каждый из которых описывается некоторым конечным набором дискретных атрибутов, и в каждом из случаев дано значение некоторой (неизвестной) булевой функции, зависящей от этих атрибутов. Задача — создать достаточно экономичную конструкцию, которая бы описывала эту функцию и позволяла классифицировать новые, поступающие извне данные. Дерево принятия решений — это дерево, на ребрах которого записаны атрибуты, от которых зависит целевая функция, в листьях записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение[10].

Искусственные нейронные сети (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей. При обучении сети предлагаются различные образцы образов с указанием того, к какому классу они относятся. Образец, как правило, представляется как вектор значений признаков. При этом совокупность всех признаков должна однозначно определять класс, к которому относится образец. По окончании обучения сети ей можно предъявлять неизвестные ранее образы и получать ответ о принадлежности к определённому классу[11,12].

Генетические алгоритмы - это алгоритмы поиска, используемые для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию[13]. Они являются разновидностью эволюционных вычислений, работают с совокупностью "особей" - популяцией, каждая из которых представляет возможное решение данной проблемы. Каждая особь оценивается мерой ее "приспособленности" согласно тому, насколько "хорошо" соответствующее ей решение задачи. Наиболее приспособленные особи получают возможность "воспроизводить" потомство с помощью "перекрестного скрещивания" с другими особями популяции. Это приводит к появлению новых особей, которые сочетают в себе некоторые характеристики, наследуемые ими от родителей. Наименее приспособленные особи с меньшей вероятностью смогут воспроизвести потомков, так что те свойства, которыми они обладали, будут постепенно исчезать из популяции в процессе эволюции. Иногда происходят мутации, или спонтанные изменения в генах.

Таким образом, из поколения в поколение, хорошие характеристики распространяются по всей популяции. Скрещивание наиболее

приспособленных особей приводит к тому, что исследуются наиболее перспективные участки пространства поиска. В конечном итоге популяция будет сходиться к оптимальному решению задачи. Преимущество данного алгоритма состоит в том, что он находит приблизительные оптимальные решения за относительно короткое время.

Коллективы решающих правил. Заканчивая обзор основных методов распознавания образов, остановимся еще на одном подходе. Это так называемые коллективы решающих правил. В таком правиле применяется двухуровневая схема распознавания. На первом уровне работают частные алгоритмы распознавания, результаты которых объединяются на втором уровне в блоке синтеза. Наиболее распространенные способы такого объединения основаны на выделении областей компетентности того или иного частного алгоритма. Простейший способ нахождения областей компетентности заключается в априорном разбиении пространства признаков исходя из профессиональных соображений конкретной науки (например, расслоение выборки по некоторому признаку). Тогда для каждой из выделенных областей строится собственный распознающий алгоритм. Самый общий подход к построению блока синтеза рассматривает результирующие показатели частных алгоритмов как исходные признаки для построения нового обобщенного решающего правила[14,15].

1.3. Этапы решения QSAR-задачи

Традиционно QSAR-задача разбивается на две подзадачи:

1) Этап описания:

Преобразование информации о М-графе к вектору численных признаков (дескрипторов), формирование матрицы молекула-дескриптор.

2) Этап поиска функциональной зависимости:

Анализ полученной МД-матрицы и построение модели функциональной зависимости - классифицирующей функции F с наилучшей прогностической способностью.

1.3.1.Этап описания

В QSAR-задаче предполагается, что молекулы со сходными локальными элементами структуры обладают сходными биологическими (химическими) свойствами и наоборот. Нахождение количественной связи между структурой и свойствами и является одной из главных целей задачи. Выразить в числовом виде свойство достаточно просто – активность серии веществ можно измерить количественно. Гораздо сложнее численно выразить структуры химических соединений. Для такого выражения в QSAR-задаче используются молекулярные дескрипторы.

Дескриптор - какое-либо свойство, численное значение которого может быть вычислено для произвольного молекулярного графа G. Дескриптор – параметр, характеризующий структуру органического соединения, причём так, что подмечаются какие-то определенные особенности этой структуры. В принципе дескриптором может являться любое число, которое можно рассчитать из структурной формулы химического соединения – молекулярный вес, число атомов определенного типа (гибридизации), связей или групп, молекулярный объём, частичные заряды на атомах и т.д. Приведем основные молекулярные дескрипторы, используемые при решении задачи «структура-свойство»:

Физико-химические дескрипторы[16] — числовые характеристики, получаемые в результате моделирования физико-химических свойств химических соединений, либо величины, имеющие четкую физико-химическую интерпретацию. Наиболее часто используются в качестве дескрипторов: липофильность (LogP), молярная рефракция (MR), молекулярный вес (MW), дескрипторы водородной связи[17], молекулярные объемы и площади поверхностей.

Квантово-химические дескрипторы[18] — числовые величины, получаемые в результате квантово-химических расчетов. Наиболее часто в качестве дескрипторов используются: энергии граничных молекулярных орбиталей (ВЗМО и НСМО), частичные заряды на атомах и частичные порядки связей, индексы реакционной способности Фукуи (индекс свободной валентности, нуклеофильная и электрофильная суперделокализуемость), энергии катионной, анионной и радикальной локализации, дипольный и высшие мультипольные моменты распределения электростатического потенциала.

Константы заместителей впервые были введены Л. П. Гамметом в рамках уравнения, получившего его имя, которое связывает константы скорости реакции с константами равновесия для некоторых классов органических реакций[19]. Константы заместителей вошли в практику QSAR после появления уравнения Ганча-Фуджиты, связывающего биологическую активность с константами заместителей и значением липофильности. В настоящее время известно несколько десятков констант заместителей.

Фармакофорные дескрипторы показывают, могут ли простейшие фармакофоры, состоящие из пар или троек фармакофорных центров со специфицированным расстоянием между ними, содержатся внутри анализируемой молекулы[20].

Топологические индексы привлекают внимание химиков как инструмент для компактного и эффективного описания структурных формул химических молекул, позволяющий также изучать и прогнозировать взаимосвязь строения и свойств органических соединений[21]. Топологическим индексом называют инвариант М-графа.

Инвариантом графа G называется число (функция, определенная на множестве всех графов), вычисляемое по G и не зависящее от способа нумерации его вершин. Примерами инвариантов графа могут служить число вершин или ребер графа, число простых цепочек заданной длины.

Преимуществом данного подхода является простота, сравнительно малые затраты компьютерного времени, а также хорошая корреляция с широким набором физико-химических параметров. Однако такой подход имеет и недостаток: он не позволяет различать разные конфигурации молекул и не учитывает их конформационные особенности. Существует более двадцати топологических индексов, различающихся способами преобразования М-графа в число. Приведем некоторые из них:

Индекс Винера - число связей, существующих между всеми парами атомов в М-графе G. Определяется по следующей формуле:

$$W(G) = 0.5\ \Sigma d_{ij},$$

где d_{ij} – число ребер, соединяющих i-ю и j-ю вершины графа G наикратчайшим путем. Суммирование проводится по всем вершинам графа[22].

Индекс Хосойи задается уравнением:

$$Z(G) = \Sigma p(G,k),$$

где p(G,k) – число способов, с помощью которых k ребер графа G могут быть выбраны так, что никакие два не будут смежными. Суммирование проводится по k=0..[n/2]. n- количество вершин графа[23].

Индекс Платта F(G) равен сумме степеней каждого ребра в графе G:

$$F(G) = \Sigma \deg e_f,$$

где deg e_f – число ребер, смежных с ребром f. Суммирование проводится по f=1..m, где m - полное число ребер в графе[24].

Индекс связности 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[25].

Помимо перечисленных также широко используются семейства информационных топологических индексов[26] и топологических индексов спектрального типа[27].

Фрагментные дескрипторы характеризуют наличие, количество и взаимное расположение в молекуле определенных структурных фрагментов (атомов, связей и т.д.) [2,28]. Бывают двух типов:

- 2D-дескрипторы. В данном случае не учитываются значения валентных углов и евклидовых расстояний между атомами, не учитывается трехмерная структура фрагмента, важны только связи между атомами. Структурные 2D-фрагменты обычно имеют вид цепочек связанных атомов с определенными метками вершин и ребер, образующих данную цепочку.

- 3D-дескрипторы. Дескрипторы этого типа учитывают трехмерную структуру фрагмента и обычно представляют собой множество вершин с заданными условиями на расстояния между ними и на их метки.

1.3.2.Этап поиска функциональной зависимости

На этом этапе необходимо по полученной МД-матрице построить классифицирующую функцию F, обеспечивающую лучшее значение

функционала качества. Все прогнозирующие функции, используемые в задаче «структура-свойство», можно условно разделить на два больших класса:

- линейные или кусочно-линейные функции,
- нелинейные зависимости.

Линейные или кусочно-линейные функции исторически наиболее широко применялись для поиска зависимостей «структура-свойство». Для построения линейных зависимостей используются стандартные статистические методы поиска зависимостей, например, линейная или гребневая регрессия. При использовании кусочно-линейных функций для разделения выборки на подклассы используются методы кластерного анализа [2929].

Наиболее распространенными методами поиска нелинейных зависимостей являются построение и обучение искусственной нейронной сети [11] , а также классификация методом k ближайших соседей [6].

Особенностью задачи «структура-свойство» является большое количество дескрипторов при использовании стандартных методов описания молекул. Так, количество разработанных различных топологических дескрипторов превышает несколько сотен, а при использовании структурных дескрипторов число дескрипторов быстро растет с усложнением описания (увеличением количества символьных меток, использованием более сложных структурных фрагментов).

В таких случаях использование стандартных регрессионных методов затруднительно; по этой причине, используется широкий спектр эволюционных алгоритмов с тем, чтобы выбрать из всех дескрипторов информативные (т.е., те которые наиболее сильно влияют на активность молекул) и свести задачу классификации к более простой.

В качестве таких эволюционных алгоритмов используются метод группового учета аргументов (МГУА) [3], метод частичных наименьших квадратов [30], генетические алгоритмы для поиска информативных

дескрипторов [13**Error! Reference source not found.**], обучение нейронных сетей [12], совмещение методов нейронных сетей и нечеткой классификации [3131].

Важной при использовании нелинейных методов классификации является также задача содержательной химико-биологической интерпретации полученной модели «структура-свойство», т.е. задача выделения элементов молекулы, ответственных за активность. Как правило, нелинейные модели (в частности, нейронные сети) имеют вид «черного ящика» с молекулой на входе и предсказанием ее активности на выходе, что делает интерпретацию модели затруднительной.

В результате, при выборе метода машинного обучения в работе отдается предпочтение МГУА – эволюционному методу, строящему линейные модели. Данный выбор обосновывается следующими причинами:

- линейные модели исторически широко распространены среди ученых-химиков;

- линейные модели легко интерпретируемы: в линейной комбинации, представляющей прогнозирующую функцию, положительный коэффициент перед дескриптором показывает, что данные дескриптор усиливает биологическое свойство, отрицательный – ослабляет его;

-эволюционные алгоритмы решают проблему большого количества дескрипторов.

Выводы.

Задача «структура-свойство» или задача прогнозирования физико-химической или биологической активности вещества является актуальной проблемой. Математические модели «структура-свойство» широко

используются на практике. С их помощью осуществляется поиск новых соединений с заданными химико-биологическими свойствами, предсказываются активности веществ. Данные модели позволяют значительно сократить расходы и время, необходимое для исследований, при синтезе новых соединений с заданными свойствами.

В главе проведен обзор основных молекулярных дескрипторов, используемых при описании молекул, рассмотрены их достоинства и основные недостатки, такие как невозможность учесть трехмерную структуру молекулы, сложность содержательной интерпретации дескрипторов, некорректная обработка молекул, имеющих множество пространственных конфигураций (для топологических дескрипторов); большое число дескрипторов, необходимость оптимизации параметров алгоритма (для структурных дескрипторов). Описаны основные методы распознавания образов, используемые при решении задачи, а также преимущества выбора в работе эволюционного алгоритма МГУА, строящего линейные модели для прогнозирования свойств молекул.

Глава 2. Описание обучающей выборки

1.1. Выбор элементов описания

В качестве молекулярных дескрипторов в работе используются фрагментные дескрипторы. Этот подход заключается в выделении фрагментов и сопоставлении каждому из них структурных дескрипторов, значение которых соответствует наличию или отсутствию данного фрагмента в М-графе или количеству его повторений. В первом случае дескриптор принимает логические значения, во втором - целые положительные. Уникальная роль фрагментных дескрипторов заключается в том, что они образуют базис дескрипторного пространства, то есть любой молекулярный дескриптор может быть однозначно разложен по этому базису[32].

Этот метод описания обучающей выборки в общем виде заключается в следующем:

- проводится дополнительная классификация атомов на основе их некоторых свойств (заряда, эксцентриситета вершины, каких-либо топологических свойств). В результате этого метка каждой вершины заменяется на другую, содержащую информацию о локальных свойствах.

- в молекулах выбираются структурные фрагменты (атомы, цепочки связанных атомов, группы атомов). Каждому структурному фрагменту сопоставляется символьное имя.

- множества фрагментов для всех молекулярных графов выборки объединяются. Для каждого молекулярного графа и каждого фрагмента находим значение соответствующего структурного дескриптора. В итоге по данной обучающей выборке получаем матрицу «молекула-признак».

Алфавит дескрипторов - множество всех дескрипторов, используемых для анализа обучающей выборки, обозначенных различными символьными метками.

Пусть алфавит дескрипторов состоит из M элементов. **Вектором признаков** молекулярного графа G будем называть вектор $X=(x_1,...,x_M)$, где x_j-значение j-ого дескриптора, вычисленное для G.

Матрицей «молекула-дескриптор» (матрицей признаков) для рассматриваемой обучающей выборки из N молекул будем называть матрицу размера N x M, в i-ой строке которой стоит вектор признаков X i-ого соединения.

В работе для разбиения на кластеры используются 2D-дескрипторы. Ими являются k-фрагменты (для k=2,3,4). **k-фрагментом** называется ациклический фрагмент, состоящий из k атомов. Значениями структурных дескрипторов, соответствующих цепочкам атомов, являются числа, характеризующие количество повторений этих цепочек в M-графе.

Рассмотрим шаги описания более подробно на примере. Пусть обучающая выборка состоит из следующих молекул:

1. H_3C———NH———O———Na
2. H_3C———S———O———CH_3
3. H_3C———NH———O———S———CH_3

В качестве меток атомов возьмем их имена в таблице Менделеева. В качестве фрагментов возьмем цепочки из трех атомов. В каждой молекуле осуществляем поиск выбранных фрагментов. Кодируем цепочки по меткам атомов в них. В результате получим:

1. C N O, N O Na
2. C S O, S O C
3. C N O, N O S, O S C

Объединяем получившиеся коды фрагментов в общий список, убирая повторяющиеся:

1, "C N O "
2, "C S O "
3, "N O Na"
4, "N O S "
5, "O S C "
6, "S O C "

Формируем МД-матрицу по молекулам выборки:

	1	2	3	4	5	6
1.	**1**	**0**	**1**	**0**	**0**	**0**
2.	**0**	**1**	**0**	**0**	**0**	**1**
3.	**1**	**0**	**0**	**1**	**1**	**0**

В этой матрице количество столбцов равно количеству элементов выбранного алфавита дескрипторов, то есть шести. Количество строк совпадает с количеством молекул. В данном примере элементы матрицы соответствуют количеству повторений фрагмента в М-графе.

1.2. Кодирование фрагментов

Каждому фрагменту сопоставляется код, символьная строка, составленная из меток его атомов, записанных в порядке обхода вершин фрагмента. Коды фрагментов совпадают тогда и только тогда, если фрагменты изоморфны. Это позволяет представлять список фрагментов в виде списка символьных кодов (строк) и работать в дальнейшем только с этими строками. При составлении метки атома в работе используются три маркера:

d-маркер ("degree") - односимвольный маркер-цифра, указывающий число соседних атомов (кроме атомов водорода). Также маркер можно рассматривать как степень вершины М-графа со стертым водородом.

b-маркер ("bond") - односимвольный маркер, описывающий кратные химические связи атома:

"s" (single) - все связи атома одинарные;

"d" (double) - у атома есть двойная связь;

b = "t" (triple) - у атома есть тройная связь;

"a" (aromatic) - у атома есть ароматическая связь;

"w" - у атома есть две двойных связи.

r-маркер ("ring") - односимвольный маркер, определяющий положение атома в кольцах:

c (chain) - атом ациклический (цепной);

r = r (ring) - атом "чисто кольцевой";

s (substitute) - атом "кольцевой с заместителем".

При формировании r-маркера используются следующие определения. Ребро называется цепным, если при удалении данного ребра из графа число его связных компонент увеличивается на 1. Ребро называется кольцевым, если при удалении данного ребра из графа число его связных компонент не меняется. Атом назовем кольцевым, если все ребра, инцидентные ему кольцевые. Соответственно если у атома все ребра цепные, то атом будем называть цепным. Если у атома есть и кольцевые ребра (два и более) и цепное - то такой атом называется кольцевым с заместителем.

В качестве метки атома в работе используется строка символов "NNdbr", полученная конкатенацией имени атома по таблице Менделеева (NN) и значений трех выше описанных маркеров. Если при описании молекул какой-либо маркер не используется, то в метках атомов на его месте ставится символ "*".

1.3. Формирование МД-матрицы

Для получения МД-матрицы необходимо для каждой молекулы выборки найти все цепочки заданной длины, закодировать их, объединить в единый

список, убрав повторяющиеся коды, и занумеровать его. Количество столбцов полученной матрицы будет совпадать с количеством разных кодов цепочек. Значение (i,j)-го элемента МД-матрицы будет соответствовать количеству повторений j-го кода дескриптора в i-й молекуле.

Общую схему формирования МД-матрицы можно представить в виде следующего алгоритма:

- выбирается параметр k (сложность фрагментов, которые будут перечисляться в М-графах), символьные маркеры, на основе которых проводится дополнительная классификация вершин М-графов и определяются символьные метки атомов.

- проводится маркирование атомов

- проводится полное перечисления в каждом графе всех его k-фрагментов и строится соответствующий список кодов найденных фрагментов.

- формируется список кодов всех k-фрагментов, встречающихся в графах, каждый фрагмент получает свой порядковый номер.

- для каждого графа обучающей выборки формируется структурный спектр относительно общего списка.

Так как в случае описания выборки 2D-дескрипторами важны только связи между атомами и не учитываются значения валентных углов и евклидовых расстояний между атомами, то с целью улучшения качества прогноза функции-классификатора в работе используются 3D-дескрипторы.

Этап описания обучающей выборки 3D-дескрипторами заключается в следующем:

- каждому структурному фрагменту сопоставляются координаты (планарные или пространственные) и символьное имя - тип фрагмента;

- для всех структурных фрагментов строится матрица евклидовых расстояний между ними внутри молекулы. Выбирается разбиение расстояний на интервалы;

- для каждой молекулы перечисляются пары фрагментов "$(T_1, T_2, P), N$", где T_1 и T_2 - типы (имена) структурных фрагментов, входящих в пару; P - номер интервала расстояния между ними, N - число повторений фрагмента "(T_1, T_2, P)" в молекуле. Все такие пары объединяются в единый список, строится МД-матрица.

Выводы

В работе использованы два вида описания обучающей выборки: с помощью топологических и пространственных дескрипторов. В главе расписан алгоритм их кодирования, включающий в себя формирование меток атомов с помощью маркеров, описывающих топологию молекулы. Приведены этапы формирования МД-матриц, описывающих обучающую выборку.

Глава 3. Построение функции прогнозирования

3.1. Метод Группового Учета Аргументов

3.1.1. Описание алгоритма, основные достоинства

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. В частности, в конце работы предлагается использовать модификации МГУА для логических и вероятностных функций.

Направление МГУА может быть полезным по следующим причинам:

- находится оптимальная сложность структуры модели, адекватная уровню помех в выборке данных. (Для решения реальных проблем с зашумленными или короткими данными, упрощенные прогнозирующие модели оказываются более точными.)

- количество слоев и нейронов в скрытых слоях, структура модели и другие оптимальные параметры нейросетей находятся автоматически.

- гарантируется нахождение наиболее точной или несмещенной модели - метод не пропускает наилучшего решения во время перебора всех вариантов (в заданном классе функций)

- любые нелинейные функции или воздействия, которые могут иметь влияние на выходную переменную, используются как входные параметры

- автоматически находятся интерпретируемые взаимосвязи в данных и выбираются эффективные входные переменные

- переборные алгоритмы МГУА довольно просто запрограммировать

- метод использует информацию непосредственно из выборки данных и минимизирует влияние априорных предположений автора о результатах моделирования

- подход МГУА используется для повышения точности других алгоритмов моделирования

- МГУА дает возможность отыскания несмещенной физической модели объекта (закона или кластеризации) - одной и той же для всех будущих выборок.

3.1.2. Общая схема МГУА

Пусть задана матрица X “объект-признак” размерности (NxM). В заданном классе функций необходимо построить функцию $F(X_{j1},X_{j2},...,X_{jk})$ от k столбцов-переменных, где переменные $X_{j1},X_{j2},...,X_{jk}$ отбираются из большого набора дескрипторов $X_1,X_2,...,X_M$ (т.е. среди столбцов матрицы X, M>> k). При построении функции F необходимо минимизировать заданный критерий Ф, например, критерий наименьших квадратов невязки:

$$E = (C - F(X_{j1},X_{j2},...,X_{jk})); \qquad \Phi: (E,E) \rightarrow \min,$$

где C – заданный вектор свойств объектов длины N. Построение функции F проводится по шагам, которые называются селекциями.

На первой селекции строятся все уравнения от двух переменных:

$$S(1)=F(X_{j1},X_{j2})$$

Они упорядочиваются по возрастанию критерия Ф и выбираются Q_1 лучших уравнений, которые будут участвовать во второй селекции. Проведем замену переменных и сформируем новые Q_1 вектор-столбцов $S_i(1)=F_i(X_{j1},X_{j2})$, где $i=1..Q_1$.

На (k+1)-й селекции с помощью Q_k столбцов-признаков, полученных на предыдущей селекции строятся новые Q_k*M уравнений от двух переменных:

$$S(k+1)=F(X_j,S_i(k)),\ i=1..Q_k,\ j=1..M.$$

Из полученных уравнений отбираются Q_{k+1} лучших уравнений. Формируются Q_{k+1} вектор-столбцов $S_i(k+1)=F_i(X_j,S_i(k))$ для исследования в следующей селекции.

Работа алгоритма останавливается при достижении заданного количества селекций или предельного уровня функционала качества.

3.2. Построение линейной функции с помощью МГУА

Пусть необходимо провести прогнозирование свойства C и построить линейную функцию от k переменных:

$$F(X_{j1},X_{j2},...,X_{jk}) = b_0 + b_1*X_1 + b_2*X_2 +...+ b_k*X_k;$$

$$E = (C - F(X_{j1},X_{j2},...,X_{jk})); \quad \Phi: (E,E) \rightarrow \min.$$

где переменные $X_{j1},X_{j2},...,X_{jk}$ отбираются на основе критерия Ф из большого набора дескрипторов $X_1,X_2,...,X_M$ (т.е. среди столбцов матрицы X). Будем строить модель F шаг за шагом следующим образом:

Первая селекция. Строятся всевозможные регрессионные уравнения с двумя переменными:

$$S[1] = F(X_i, X_j) = b_0 + b_1*X_i + b_2*X_j ; \quad 1 \le i < j \le M.$$

Общее число таких уравнений равно w1=M*(M-1)/2. Согласно оптимизационному критерию Ф, отбираются наилучшие Q_1 уравнений, которые будут принимать участие во втором селекции ($Q_1 << w_1$)**.** Для каждого q-го уравнения ($q=1..Q_1$) сохраняются следующие значения (i, j, b_0, b_1, b_2).

(k+1)-ая селекция. Строятся $M*Q_k$ уравнений:

$$S[k+1]=F(X_i,S[k]_q)=b_0+b_1*X_i+b_2*S[k]_q; \quad i= 1,...,M; \quad q=1,...,Q_k.$$

где $S[k]_q$ - q-ый столбец, отобранный на k-й селекции. Согласно критерию Ф, отбираются наилучшие Q_{k+1} уравнений и для каждого из них сохраняются параметры (i, q, b_0, b_1, b_2).

После проведения заданного числа селекций можно построить семейство целевых уравнений, используя сохраняемые на каждой селекции коэффициента линейной регрессии и номера столбцов.

$$F(X_{j1},X_{j2},...,X_{jk})_q = (b_0 + b_1*X_1 + b_2*X_2 +...+ b_k*X_k)_q; \; q=1..Q_k$$

3.2. Линейная регрессия

В МГУА для линейной функции прогнозирования на каждой шаге решаются линейные регрессионные уравнения. 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 – результирующий вектор активности. По этим данным требуется построить прогнозирующую функцию F(x), где x – строка матрицы X, соответствующая одному объекту. 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.

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

$$\varepsilon_i = y_i - F(x_i) = y_i - z_i \beta; \qquad i=1..N$$

были как можно меньше, а именно, чтобы сумма их квадратов была минимальной.

Выводы

В главе дан обзор Метода Группового Учета Аргументов, его основные преимущества, описана схема работы алгоритма, как в общем случае, так и в случае линейной прогнозирующей функции. Рассмотрен метод построения линейной прогностической модели с помощью линейной регрессии.

Глава 4. Особенности программной реализации, проведение расчетов

4.1. Архитектура программы

Предложенное в работе построение кусочно-линейных классификаторов свойств молекул реализовано в среде Matlab, так как в используемых алгоритмах работа в основном проводится над матрицами и векторами и используются стандартные методы многомерного статистического анализа, имеющие готовые реализации в пакете Matlab Statistics. Программная реализация представлена в виде отдельных модулей – функций.

Опишем основные модули и используемые в них алгоритмы. По каждой молекуле выборки составляется топологическая матрица и вектор строка с именами атомов молекулы. С помощью них осуществляется маркирование атомов. При подсчете r-маркера необходимо для каждого ребра графа выяснить является оно кольцевым или цепным. Для этого в работе используется волновой алгоритм, который по графу и двум его вершинам определяет, существует ли путь, их соединяющий и выдает в качестве ответа либо количество ребер кратчайшего пути, либо 0, если вершины находятся в различных компонентах связности. Алгоритм работает следующим образом:

Дано: непустой граф G=(V,E). Требуется найти путь между вершинами s и t графа (s не совпадает с t), содержащий минимальное количество промежуточных вершин (ребер).

1. Каждой вершине v_i приписывается целое число $T(v_i)$ - волновая метка (начальное значение $T(v_i)=-1$);

2. Заводятся два списка OldFront и NewFront (старый и новый "фронт волны"), а также переменная T (текущее время);

3. OldFront:={s}; NewFront:={}; T(s):=0; T:=0;

4. Для каждой из вершин, входящих в OldFront, просматриваются инцидентные (смежные) ей вершины u_j, и если $T(u_j) = -1$, то $T(u_j):=T+1$, NewFront:=NewFront + $\{u_j\}$;

5. Если NewFront = {}, то ВЫХОД("нет решения");

6. Если t принадлежит NewFront (т.е. одна из вершин u_j совпадает t), то найден кратчайший путь между s и t с T(t)=T+1 промежуточными ребрами; ВЫХОД("решение найдено");

7. OldFront:=NewFront; NewFront:={}; T:=T+1; goto (4).

Поиск всех k-цепочек молекулы в работе осуществляется по топологической матрице М-графа. Сложность поиска всех фрагментов обучающей выборки равна $O(N*n^k)$, где N – число молекул выборки, n – среднее число атомов в одной молекуле, k – длина цепочек атомов, используемых при описании. Подсчитаем максимальное количество k-цепочек молекулы из n атомов. Учитывая особенности М-графа, предполагаем, что максимальная степень вершин равна 4.

Утверждение: В графе из n вершин, максимальная степень которых равна 4, не более чем $2*n*3^{k-2}$ цепочек, состоящих из k вершин. k=2,3,4.

Доказательство: Пусть G – граф из n вершин, максимальная степень которых равна 4. Докажем отдельно для каждого значения k:

k=2. Из теоремы, утверждающей, что сумма степеней всех вершин графа равна удвоенному числу ребер графа следует, что число ребер (цепочек из двух вершин) графа G не больше 2*n.

k=3. Рассмотрим все цепочки из двух вершин. Их количество <= 2*n. Рассмотрим одну цепочку (v1,v2). Так как степени вершин не больше 4, то из

ребра (v1,v2) можно присоединением одного ребра получить цепочку из трех вершин максимум шестью способами (добавляем 3 ребра, входящие в вершину v1, и 3 ребра, выходящие из v2). Во избежание повторения, для определенности, будем образовывать цепочки из трех вершин добавлением только ребер, выходящих из v2. Таким образом, максимальное число цепочек из трех вершин равно $3*2*n=2*n*3^{3-2}$.

k=4. Повторив рассуждения предыдущего пункта, получим что максимальное число цепочек из 4 вершин равно $3*2*n*3^{1}=2*n*3^{4-2}$. Утверждение доказано.

Из утверждения следует, что максимальное число k-цепочек во всей выборки равно $N*2*n*3^{k-2}$. После кодировки всех цепочек и отбрасывания повторяющихся кодов, формируется новый список и строится МД-матрица.

Разбиение на кластеры осуществляется по МД-матрице, составленной по k-фрагментам с помощью стандартной функции среды MATLAB fcm. Формируются правила отказа от прогноза для идентификации выбросов, а также правило принадлежности к кластеру для новых молекул, активность которых неизвестна. Для новой молекулы, обладающей «непохожей» структурой, активность не будет спрогнозирована и будет осуществлен отказ.

Для построения функции прогнозирования формирование МД-матрицы осуществляется с помощью 3D-дескрипторов, которые позволяют учитывать трехмерную структуру молекул. Функция строится с помощью модификации схемы работы МГУА. Она заключается в следующем:

1. На первой селекции строятся уравнения от одной переменной, и все они отбираются для следующих селекций (Q_1 = M, где M – количество получившихся различных кодов 3D-фрагментов обучающей выборки). Количество отбираемых моделей на всех селекциях, начиная со второй, одинаково и равно Q_2. Общее количество селекций [N/5]+1.

Первая селекция:

$$S[1]_i = F(X_i) = b_0 + b_1 * X_i ; \quad 1 \le i \le M.$$

На следующих селекциях вместо столбцов исходной матрицы X используются столбцы $S_i[1]$, полученные на первой селекции.

(k+1)-ая селекция. Строятся $Q_1 * Q_2$ уравнений:

$$S[k+1]=F(S[1]_i,S[k]_q)=b_0+b_1 * S[1]_i +b_2 * S[k]_q; \quad i= 1,...,Q_1; \quad q=1,...,Q_2.$$

где $S[k]_q$ - q-ый столбец, отобранный на k-й селекции.

2. Так как селекции МГУА основаны на переборе всех полученных прогнозирующих моделей и всех имеющихся переменных, проведение расчетов занимает много времени. С этой целью на каждом шаге МГУА проводится фильтрация прогнозирующих моделей: отбрасываются почти константные модели, дисперсия которых меньшего некоторого заданного порога, близкого к 1, для каждой пары моделей вычисляется коэффициент корреляции, и если он оказывается меньше заданного порога, то из этой пары отбрасывается модель, хуже прогнозирующая свойство.

Для поиска необходимых коэффициентов линейной функции прогнозирования по предложенной модификации необходимо:

- не более чем M + A + Q1* Q2*([N/5]-1) раз решить уравнение линейной регрессии, такое же количество раз необходимо вычислить дисперсию моделей.

- вычислить не более чем M*(M-1)/2 + A*(A-1)/2 + ½* Q1*Q2* *(Q1*Q2-1)*([N/5]-1) коэффициентов корреляции, где A=½* Q1*(Q1-1).

4.2. Бетулин, фармакологическая активность его производных

В обучающей выборке, рассматриваемой в работе, 52 молекулы производных бетулиновой кислоты. 14 из них обладают цитотоксической

активностью. Дадим краткий обзор исследуемым в работе химическим соединениям.

Бетулин - это тритерпеновый спирт ряда лупана, имеющий химическую формулу $C_{30}H_{50}O_2$ и химическое название бетуленол. Он содержится в большом количестве растений (орешник, календула, солодка и пр), в промышленных масштабах его получают экстракцией из бересты - наружного слоя коры березы белой (betula alba). Несмотря на то, что бетулин известен своими целебными свойствами давно (он был открыт Т.Е. Ловицем - преемником М.В. Ломоносова - в 1778г), в последние годы в мировой фармакологии наблюдается небывалый всплеск интереса к нему. Исследованиями специалистов разных стран доказано, что сам бетулин (бетуленол) и его производные - бетулиновая кислота и другие дериваты обладают выраженной фармакологической активностью.

Бетулиновая кислота - растительный пентациклический тритерпеноид, обладающий избирательным цитотоксическим действием в отношении различных опухолевых клеток (цитотоксичность - способность вызывать патологические изменения в клетках живого организма). Противоопухолевая активность была показана на опухолевых клетках человека (на клеточных линиях меланом, лимфом, нейробластом). В настоящее время бетулиновая кислота проходит клинические исследования в качестве средства для лечения злокачественной меланомы. Бетулиновая кислота блокирует рост меланомы без вреда для нормальных клеток.

4.3. Проведение вычислительных экспериментов

Расчеты проводились на обучающей выборке, состоящей из 52 молекул. Для всех вариантов k=2,3,4 и всех видов маркеров с помощью топологических дескрипторов построены МД-матрицы, осуществлено разбиение на кластеры и построена кусочно-линейная функция. Наилучший полученный процент

правильного прогноза активности на всей выборке составляет 92.31%, а при разбиении на два кластера – 91.43% и 88.24%.

Выводы

В работе был разработан новой подход к решению задачи «структура-свойство». Он заключается в различном описании выборки на этапах кластеризации и прогнозирования активности веществ. Описание, основанное на топологических дескрипторах, позволяет эффективно разбить на кластеры и при построении правила принадлежности к кластеру. А описание с помощью 3D-дескрипторов, учитывающих трехмерную структуру молекул, используется для построения функции-классификатора, обладающей лучшей прогностической способностью по сравнению с описанием топологическими дескрипторами. Были исследованы варианты использования эволюционного алгоритма на различных параметрах, построено правило отказа от прогноза для идентификации выбросов. Решение задачи реализовано в среде MATLAB в виде отдельных модулей-функций.

Заключение

В работе решена задача «структура-свойство» для выборки производных бетулиновой кислоты. Новизной работы является использование различных видов описания обучающей выборки для разбиения на кластеры и построения функции. С помощью фрагментного подхода по молекулам формируется вектор признаков, с помощью которых описывается выборка. Каждая молекула выборки представляется в виде значений данных признаков.

С целью улучшения качества прогноза обучающая выборка, построенная на топологических дескрипторах, разбивается на кластеры, на каждом из которых строится своя модель прогнозирования. По новой молекуле определяется ближайший из кластеров и по его функции прогнозирования находится активность. Для идентификации выбросов строится правило отказа от прогноза.

По матрице «молекула-признак», полученной по пространственным дескрипторам, строится линейная прогнозирующая функция. Для этого используется эволюционный алгоритм Метод Группового Учета Аргументов. В работе предложена модификация этого алгоритма, позволяющая сократить объем вычислений. Она состоит в фильтрации векторов каждой селекции: отбрасываются почти константные и сильно коррелирующие. Все этапы построения функции-классификатора по обучающей выборке реализованы в среде MATLAB. Написано около 15 отдельных модулей-функций.

Проведены вычислительные эксперименты на различных видах описания выборки, самый лучший процент правильного прогноза активности 92.31%.

Список литературы

1. Karelson M. Molecular Descriptors in QSAR/QSPR. Wiley-interscience, 2000
2. 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.
3. Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. Принятие решений на основе самоорганизации. М.: Сов. Радио, 1976
4. Rouvray D.H. (Ed.) Computational Chemical Graph Theory. / Nova Publ., New York, 1989
5. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и распознавание изображений. / В сб. Распознавание. Классификация. Прогноз. Математические методы и их применение. Вып.2, М.: Наука, 1989, с.5-72.
6. Дуда Р., Харт П. Распознавание образов и анализ сцен. – М.:Мир, 1976.
7. 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.
8. Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания и классификации. - М.: Наука, 1978, вып. 33.
9. Christopher J.C. Burges A Tutorial on Support Vector machines for Pattern Recognition. - Data Mining and Knowledge Discovery 2, 1998.
10. Ананий В. Левитин Глава 10. Ограничения мощи алгоритмов: Деревья принятия решения // Алгоритмы: введение в разработку и анализ = Introduction to The Design and Analysis of Aigorithms. — М.: «Вильямс», 2006. — С. 409-417

11. Барцев С. И., Охонин В. А. Адаптивные сети обработки информации. Красноярск : Ин-т физики СО АН СССР, 1986. Препринт N 59Б. — 20 с.
12. Aoyama T., Suzuki Y., Ichikawa H. Neural networks applied to quantitative structure-activity relationship analysis. J.Med.Chem., vol.33, pp.2583-2590, 1990.
13. J. H. Holland. Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, 1975.
14. Kuncheva L.I. Combining pattern classifiers. Methods and Algorithms. – Wiley, 2004.
15. Vetrov D.P. On the Stability of the Pattern Recognition Algorithms. – Pattern Recognition and Image Analysis, 2003, vol. 13, no. 3.
16. О. А. Раевский (1999). "Дескрипторы молекулярной структуры в компьютерном дизайне биологически активных веществ". Успехи химии 68 (6): 555-575
17. О. А. Раевский (2006). "Дескрипторы водородной связи в компьютерном молекулярном дизайне". Рос. хим. ж. (Ж. Рос. хим. об-ва им. Д.И.Менделеева) L (2): 97-107.
18. M. Karelson, V. S. Lobanov, A. R. Katritzky (1996). «Quantum-Chemical Descriptors in QSAR/QSPR Studies». Chem. Rev. 96 (3): 1027-1044
19. Пальм, В. А. Основы количественной теории органических реакций. — 2-е, пер. и доп.. — Л.: Химия, 1977. — 360 с.
20. F. Bonachera, B. Parent, F. Barbosa, N. Froloff, D. Horvath (2006). «Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes». J. Chem. Inf. Model. 46 (6): 2457-2477
21. Randic M. On Characterization of Molecular Branching. Journal of the American Chemical Society, 1975, vo.97, pp.6609-6615

22. Wiener H., J. Am. Chem. Soc. 1947, v. 69, p. 17, 2636; J. Chem. Phys., 1947, v.15, p.766.
23. Hosoya H., Int. J. Quant. Chem., 1972, v.6, p. 801.
24. Platt J., J. Chem, Phys., 1947, v.15, p.419
25. Randic M., J. Am. Chem. Soc., 1975, v.97, p.6609
26. В. Магнусон, Д. Харрис, С. Бейсак В кн. Химические приложения топологии и теории графов. (Ред. Р. Кинг) - М.:Мир, 1987.
27. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Успехи химии, 57.
28. I. Baskin, A. Varnek. «Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening». In: Chemoinformatic Approaches to Virtual Screening, A. Varnek, A. Tropsha, eds., RCS Publishing, 2008
29. Cho S.J., Tropsha A. Cross-Validated R2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. J.Med.Chem., 1995, v.38, pp.1060-1066.
30. Wold S., Ruhe A., Wold H., Dunn W.J. The collinearity problem in linear regression: The partial least squares approach to generalized inverses. SIAM J Sci. Stat. Comput. Vol.5, pp.735-743, 1984.
31. Loukas Y. Adaptive neuro-fuzzy inference system: an instant and architecture-free predictor for improved QSAR studies. J. Med. Chem., vol.44, pp. 2772-2783, 2001.
32. И. И. Баскин, М. И. Скворцова, И. В. Станкевич, Н. С. Зефиров (1994). «О базисе инвариантов помеченных молекулярных графов». Докл. РАН 339 (3): 346-350.