Прохоров Евгений

Использование нечёткой функции классификации при прогнозировании свойств химических соединений

В работе предложен новый подход к анализу матрицы «молекула - дескриптор» в задаче «структура - свойство», основанный на нечёткой кластерной структуре обучающей выборки. Описаны методы построения «быстрых» правил отказа от прогноза и поиска выбросов в обучающей выборке. Для этого введено специальное пространство дескрипторов, лёгких с вычислительной точки зрения. Предложены методы улучшения качества прогноза, основанные на выборе формы кластеров и области допустимых М-графов.

Рассматривается оптимизация классифицирующей функции по параметрам нечёткой кластеризации.

Построены прогнозирующие модели с высоким качеством прогноза, основанные на предложенном подходе. Проведено сравнение моделей, показывающее эффективность описанных методов.

Оглавление

Введение

В настоящее время уровень развития вычислительной техники и ресурсы, предоставляемые современными вычислительными системами достигли того уровня, на котором задача применения средств информатики в различных областях естественных и гуманитарных наук становится основой развития новых подходов и получения новых результатов. Колоссальный рост мощности и распространённости компьютерной техники позволили обрабатывать накопленные за долгие годы экспериментальные данные, проверять и анализировать различные эмпирические знания и методы, широко распространённые в естественных науках.

Информатизация проникает во все сферы человеческой деятельность и в первую очередь в науку, позволяя применять численные методы и алгоритмы вычислительной математики к большим объемам научной информации. Автоматизация процессов обработки информации основывается на моделировании материальных объектов и процессов, включая и процесс человеческого мышления — один из сложнейших видов человеческой деятельности.

Широчайшие возможности для применения во многих областях науки получила теория распознавания образов, как научная дисциплина, изучающая методы отнесение исходных данных к определенному классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы несущественных данных.

Вообще, так как задача распознавания формулируется очень общо, а сама теория использует множество методов заимствованных из различных областей науки, начиная генетикой и заканчивая статистикой, распознавание образов находит применение всё в новых и новых областях.

В химии информатика находит более, чем широкое применение. К возможным задачам информатики в области химии можно отнести в общем случае дизайн, создание, организацию, управление, поиск, анализ, распространение, визуализацию и использование химической информации. В 1998 году Ф. К. Брауном даже был введен в употребление специальный термин «хемоинформатика». Он определял хемоинформатику как совместное использование информационных ресурсов для преобразования данных в информацию и информации в знания для быстрейшего принятия наилучших решений при поиске соединений-лидеров в разработке лекарств и их оптимизации. Сейчас мы понимаем под хемоинформатикой вообще любое применение методов информатики для решения химических проблем.

Методы хемоинформатики начинают активно внедряться во все области химии, и, прежде всего, в органическую химию.

Данная работа посвящена задаче «структура – свойство». Задача «структура – свойство» (QSAR - Quantitative Structure Activity Relationship) – актуальная задача распознавания образов – состоит в том, чтобы по структуре химического соединения предсказать его активность (химическую или биологическую) [1, 2, 3]. Полученные модели используются для скрининга молекулярных баз данных, поиска новых потенциально активных веществ. 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.

Особенностью QSAR-задачи является необходимость описать структуру химического соединения в виде дескрипторов - любых свойств молекулы, выраженных численно. Дескрипторы выступают в роли признаков объекта распознавания. Поэтому решение задачи разбивается на два основных этапа: этап построения описания обучающей выборки, на котором формируется

матрица «молекула – дескриптор», и этап поиска функциональной зависимости.

Задача построения описания обучающей выборки была подробно описана в [3, 4]. Решение задачи поиска функциональной зависимости разбивается на несколько этапов: кластеризация и обработка выбросов; отбор значимых дескрипторов; построение модели; прогноз. Особое внимание в этой работе уделено кластеризации, обработке выбросов, а также методам построения моделей. Возможные реализации других этапов можно посмотреть в [1, 2].

Кластеризация позволяет строить модель локально, внутри небольшой группы сходных соединений, что часто оказывается очень полезным. Задача кластеризации чрезвычайно важна также для ускорения вычислений, ведь построение моделей внутри каждого кластера может идти параллельно, сам алгоритм кластеризации также может быть распараллелен. Эффективность проводимых вычислений важна, так как необходимо обрабатывать большие массивы молекулярных структур, собранных во многих организациях, например, таких, как National Cancer Institute [www.cancer.gov], где выборки содержат сотни молекул, для каждой из которых обрабатываются тысячи дескрипторов.

Метод, разработанный в этой работе, развивает данное направление, однако ориентирован на избавление от ряда недостатков, которыми обладают классические подходы:

Новое соединение предсказывается моделью, безотносительно к тому принадлежит ли оно вообще классу веществ, с помощью которых строилась модель. Таким образом, ошибка при прогнозировании больших групп соединений неоправданно велика. Выбросы ищутся уже в процессе построения модели, что замедляет работу алгоритма.

Чёткая (стандартная) кластеризация во многом неудовлетворительна, существующие алгоритмы не позволяют одному и тому же объекту

(соединению) принадлежать одновременно нескольким кластерам, что создаёт сложности для равноудалённых от двух кластеров объектов. Также при построении модели молекулы из центра кластера вносят такой же вклад в прогнозирование новых соединений, что и удалённые от центра, что не логично с точки зрения сделанных предположений.

Таким образом, мы выдвигаем следующие требования к разрабатываемому методу:

– разработать быстрые с вычислительной точки зрения правила отказа от прогноза, которые осуществляли бы отказ в случае, когда предсказание свойства нового соединения нецелесообразно;

– разработать метод построения классифицирующей функции, основанной на кластерной структуре обучающей выборки и прогнозирующей значение активности нового соединения исходя из активности наиболее схожих с ним соединений из обучающего множества;

– метод должен быть устойчив к наличию в обучающей выборке выбросов, и не учитывать их при построении прогнозирующих моделей.

В данной работе мы используем кластерную структуру обучающей выборки для построения правил отказа от прогноза и поиска выбросов. Выброс – химическое соединение в выборке, признаки которого существенно отличаются от признаков остальных соединений. Такие молекулы могут попасть в выборки из-за ошибки составителей, но могут сами по себе являться содержательными с точки зрения химии. Присутствие выбросов существенно ухудшает качество прогноза. Поэтому при построении модели мы их не учитываем.

Целью данной работы являлась разработка и исследование метода построения классифицирующей функции, основанного на нечеткой кластерной структуре обучающей выборки, и быстрых правил отказа от

прогноза, когда классификация невозможна, в задаче обнаружения функциональной зависимости «структура-свойство».

Научная новизна работы состоит в следующем:

1. Предложен новый метод построения классифицирующей функции в задаче «структура - свойство», основанный на нечёткой кластерной структуре обучающей выборки. Описаны методы построения «быстрых» правил отказа от прогноза и поиска выбросов в обучающей выборке.

2. В рамках предложенных методов разработаны алгоритмы анализа матрицы «молекула - дескриптор» в задаче «структура-свойство» и проведена оценка вычислительной сложности алгоритмов.

3. Подтверждена практическая значимость подхода в серии вычислительных экспериментов по прогнозированию биологической активности органических соединений.

Практическая значимость работы состоит в том, что разработанные алгоритмы решения QSAR-задачи могут быть использованы для решения прикладных задач предсказания физико-химической или биологической активности веществ по их структуре. Это может позволить отказаться от дорогостоящих и длительных исследований внеэкспериментальным скринингом на больших наборах химических соединений.

Материалы работы докладывались и обсуждались на Международной научной конференции "Компьютерные науки и информационные технологии" (2009 г.), 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009 (2009 г.), XVII Международной конференции студентов, аспирантов и молодых учёных «Ломоносов» (2010 г.). Полученные результаты также обсуждались на научных семинарах механико-математического факультета МГУ им. М.В. Ломоносова и Института Органической Химии им. Н.Д.Зелинского РАН. По материалам

работы опубликованы 4 научные работы. Основные результаты содержатся в следующих статьях:

1. Прохоров Е.И., Перевозников А.В., Воропаев И.Д., Кумсков М.И., Пономарёва Л.А. Поиск представления молекул и методы прогнозирования активности в задаче «структура–свойство» // Доклады 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. – М: МАКС Пресс. – 2009. – С. 589-591.

2. Деветьяров Д.А., Кумсков М.И., Апрышко Г.Н., Носеевич Ф.М., Прохоров Е.И., Перевозников А.В., Пермяков Е.А. Сравнительный анализ применения нечетких дескрипторов при решении задачи «структура-свойство» // Доклады 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. – М: МАКС Пресс. – 2009. – С. 511-514.

3. Прохоров Е.И., Перевозников А.В., Пономарева Л.А. Кумсков М.И. Нейронная сеть как инструмент реализации кусочно-линейного классификатора при массовом скрининге молекул в задаче «структура-свойство» // Нейрокомпьютеры: разработка, применение. – № 3. – С. 39-45.

4. E. I. Prokhorov, L. A. Ponomareva, E. A. Permyakov and M. I. Kumskov Fuzzy classification and fast rules for refusal in the QSAR problem // Pattern Recognition and Image Analysis, 2011, Volume 21, Number 3, Pages 542-544 DOI: 10.1134/S105466181102091X

Работа поддержана Российским Фондом Фундаментальных Исследований (РФФИ) по гранту №07-07-00282.

Работа состоит из введения, четырёх глав основного текста, заключения и списка литературы.

В первой главе дана общая постановка задачи «структура-свойство» и приведен краткий обзор наиболее популярных существующих методов ее решения. Вторая глава содержит постановку конкретной задачи, определяет этапы ее решения и описывает их. Предлагается метод построения классифицирующей функции, учитывающий нечёткую кластерную структуру выборки. Описано построение быстрых правил отказа от прогноза. В третьей главе раскрывается этап анализа МД-матрицы, рассматриваются применяемые методы и схемы реализованных алгоритмов. В четвертой главе описаны особенности программной реализации проекта, и приведены результаты работы алгоритмов на обучающих выборках биологически активных веществ. Проведено сравнение результатов работы алгоритма с аналогом без использования принципов нечеткой логики. Показана перспективность применения нечеткой классификации. Заключение содержит основные выводы работы и перспективы дальнейших исследований.

Благодарность

Автор выражает глубокую признательность своему научному руководителю Кумскову Михаилу Ивановичу за постановку задач, постоянное внимание к работе и многочисленные плодотворные обсуждения. Автор также выражает благодарность своим коллегам – студентам и аспирантам кафедры вычислительной математики московского университета за помощь в подготовке работы и плодотворную совместную деятельность, а также к.х.н. Свитанько Игорю Валентиновичу (Высший Химический Колледж РАН) и к.б.н. Апрышко Галине Николаевне (Российский онкологический научный центр имени Н.Н. Блохина) за предоставление выборок химических соединений для анализа и тестирования разработанных методов.

Глава 1. Задача «структура-свойство» и методы её решения

В главе приведена краткая постановка задачи «структура - свойство» и дан краткий обзор наиболее распространенных существующих методов ее решения.

1.1. Краткая постановка QSAR-задачи

Задача «структура-свойство» для молекул (QSAR-задача) – это задача поиска численной зависимости между структурой молекулы и ее физико-химическими или биологическими свойствами. Методы решения этой задачи основаны на предположении о том, что свойства молекулы определяются отдельными элементами ее структуры. Построенные модели «структура-свойство» (QSAR-модели) позволяют прогнозировать биологическую активность и свойства химических соединений, что дает возможность осуществлять синтез веществ с заведомо заданными свойствами и избегать дорогостоящего и длительного внеэкспериментрального скрининга соединений.

Более формально, задача "структура - свойство" – это задача распознавания образов, где объектами являются молекулы, векторное описание которых заранее не задано.

Пусть задана обучающая выборка из N химических соединений. Для каждой молекулы из обучающей выборки известно её описание в виде вектора признаков $x_i = (x_{i1}, \dots x_{iM}),\ i = 1, \dots N$, а также значение её активности (целевое свойство) $y_i, i = 1, \dots N$. Задача распознавания состоит в определении активности нового соединения молекулы $\tilde{x}$ по её описанию $\tilde{x} = (\tilde{x}_1, \dots \tilde{x}_M)$ и информации об обучающей выборке $(x_1, y_1), \dots (x_N, y_N)$.

В настоящее время разработано большое число методов поиска количественных корреляций «структура-свойство». Между методами

существуют значительные различия: так, разные подходы могут использовать или не использовать информацию о трехмерной структуре молекул, использовать топологические или структурные дескрипторы и т.п. В то же время, любой метод решения QSAR-задачи сводится к применению методов машинного обучения для поиска зависимости вида $P_i = f(D_1,...,D_n)$, где P_i – исследуемое свойство, $D_1,...,D_n$ – вычисленные (или полученные экспериментально) признаки (дескрипторы) веществ.

Таким образом, QSAR-задачу можно разбить на 2 подзадачи: задачу представления информации о структуре молекулы в виде векторов признаков (*этапом описания*) и задачу поиска функциональной зависимости f между значениями признаков и значением активности/свойства (*этапом анализа*).

На первом этапе формируется так называемая МД-матрица или матрица «молекула - дескриптор», содержащая по строкам описание молекул в виде вектора дескрипторов. Второй же этап посвящён анализу этой матрицы методами распознавания образов и классификации. Далее опишем методы решения каждой из 2-ух подзадач.

Рис 1. Этапы решения задачи «структура-свойство»

1.2. Методы решения

Важным моментом в задаче «структура-свойство» является задача выбора представления информации о структуре молекулы (т.е., набора признаков-дескрипторов).

Подходы к описанию молекул можно разделить на несколько больших групп:

– описание топологическими дескрипторами, основанными на свойствах молекулярного графа (теоретико-графовыми индексами);

– описание дескрипторами, соответствующими базовым геометрическим свойствам молекул (таким, как площадь поверхности, объем, диаметр и т.д.);

– описание структурными дескрипторами, характеризующими наличие, количество, и взаимное расположение в молекуле определенных структурных фрагментов;

– описание дескрипторами, характеризующими локальные физико-химические свойства молекулы в трехмерном пространстве, заполненном некоторой регулярной сеткой.

Для построение описания обучающей будем применять структурные дескрипторы [3, 7] – пары и тройки особых точек (ОТ), определенных на триангулизированной молекулярной поверхности химического соединения. Структурный символьный спектр молекулярного графа представляет собой число повторений молекулярных фрагментов в молекулярном графе путем полного перечисления всех пар, троек, четверок особых точек [3, 7].

1.2.1. Описания молекулярного графа.

Топологические дескрипторы

К топологическим дескрипторам можно отнести дескрипторы, основанные либо на свойствах молекулярного графа молекулы (теоретико-графовые индексы) и дескрипторы, описывающие общую топологию молекулы (т.е., ее форму, размер и т.д.). В QSAR-задаче теоретико-графовые индексы были впервые использованы Рандичем [019] в задаче моделирования температуры кипения ациклических насыщенных углеводородов (алканов). Впоследствии, полученные результаты были развиты и обобщены в [20, 21].

Дескрипторы, описывающие общую топологию молекулы, как правило, строятся на основе ван-дер-ваальсовой или молекулярной поверхности молекулы. Например, в [22] в качестве дескриптора предложено использовать объем, ограниченный ван-дер-ваальсовой поверхностью, а в [023] – площадь молекулярной поверхности.

Приведем ряд примеров топологических дескрипторов.

Индекс Винера [24]. Пусть $G = \{V, E\}$ – молекулярный граф с *n* вершинами (атомами) и $D(G) = (d_{ij})$ – его матрица расстояний, так что d_{ij} – длина кратчайшего пути между *i*-ой и *j*-ой вершинами в графе *G*. Определим индекс Винера как $W(G) = \frac{1}{2}\sum_{i,j=1}^{n} d_{ij}$. В [0] показано, что индекс Винера сильно коррелирует с температурой кипения алканов.

Индекс связности Рандича [025]. Пусть $G = \{V, E\}$ – молекулярный граф с *n* вершинами (атомами) и d_i – степень *i*-ой вершины в графе *G*. Определим индекс Рандича как $\chi(G) = \sum_{e \in E} \frac{1}{\sqrt{d_i d_j}}$, где суммирование проводится по всем ребрам графа *G*, а индексы *i* и *j* относятся к номерам атомов, соединенных данным ребром. В [0] показано, что индекс Рандича может быть использован для предсказания ряда физико-химических свойств соединений.

В качестве дескриптора молекулы с графом *G* можно использовать площадь ее молекулярной поверхности $M_r(G)$ [026, 23].

В настоящее время разработано и используется более 1000 различных топологических дескрипторов. Регрессионный анализ на основе топологических дескрипторов дает относительно неплохие результаты при моделировании физико-химических свойств «простых» молекул (углеводородов и полимеров) [27]. Благодаря своей простоте, топологические дескрипторы могут быть вычислены с минимальными затратами времени.

В то же время, топологические дескрипторы практически не применяются в прогнозировании биологической активности, поскольку не учитывают трехмерную структуру исследуемых молекул. Большинство разработанных топологических дескрипторов используют лишь информацию о молекулярном графе, игнорируя его укладку в пространстве.

В частности, топологические дескрипторы неприменимы в случае анализа гибких молекул, которые могут принимать несколько различных пространственных конфигураций. Поскольку молекулярные графы одной

молекулы для всех таких конфигураций будут одинаковы, то для них будут совпадать и топологические дескрипторы; в то же время, активность молекулы в различных конфигурациях может отличаться, что делает построение предсказывающей модели на основе топологических дескрипторов невозможным.

Еще одним важным недостатком является сложность содержательной интерпретации модели, основанной на значениях топологических дескрипторов.

Структурные дескрипторы

Структурные дескрипторы характеризуют наличие или отсутствие определенных структурных фрагментов (химических функциональных групп) в молекуле. Использование данного класса дескрипторов основано на понятии «фармакофор», что есть набор структурных признаков в молекуле, которые отвечают за биологическую активность молекулы. Данное понятие было развито и использовано в [28].

Метод выделяет структурные фрагменты в структуре молекулы и находит функциональную зависимость между наличием тех или иных групп или цепочек и биологической активностью. В качестве таких фрагментов могут выступать отдельные атомы, группы атомов и цепочки атомов, соединенных связями. После выделения фрагментов каждому фрагменту сопоставляется структурный дескриптор, значение которого соответствует числу повторений фрагмента в молекуле [029, 7].

Структурные дескрипторы были впервые использованы в [30]. В работе рассматривались структурные фрагменты – пары атомов, отстоящих друг от друга на определенное расстояние. Таким образом, структурный фрагмент записывался строкой (V_i, V_j, d), где V_i, V_j – метки атомов, а d – длина минимального пути между атомами. Значением соответствующего дескриптора было число повторений фрагмента вида (V_i, V_j, d) в молекуле.

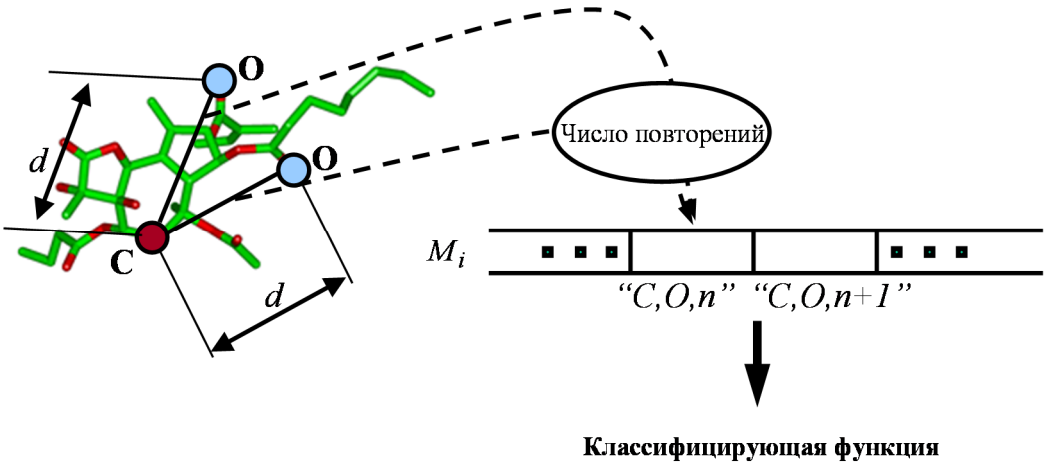


Рис 2. Построение матрицы «молекула-признак» со структурными дескрипторами – парами атомов

В дальнейшем описанный подход развивался. Было предложено использовать в качестве меток (маркеров) атомов не только символы химических элементов, но и более сложные строковые выражения, включающие в себя информацию о топологии и физико-химических свойствах отдельных атомов в молекуле [029]; увеличивалась глубина описания молекул структурными дескрипторами: вместо пар точек использовались тройки, четверки с соответствующим описанием их взаимного расположения. Наконец, было предложено использовать в качестве концов пар не атомы, а некоторые точки в пространстве, окружающем молекулы, не связанные непосредственно с молекулярным графом [31, 32].

Показано, что при достаточном уровне усложнения описания качество классификации на основе структурных дескрипторов и сходных методов сравнимо с качеством классификации, достигаемым более сложными алгоритмами, такими как 3D-QSAR [25].

К недостаткам метода структурных дескрипторов можно отнести очень большое число дескрипторов, которое, как правило, значительно превышает число молекул. Это затрудняет анализ матрицы «молекула-признак» стандартными статистическими методами и заставляет прибегать к эволюционным алгоритмам с тем, чтобы выбрать и использовать в

предсказывающей модели наиболее «информативные» дескрипторы (т.е. те, которые наиболее сильно влияют на предсказываемое свойство).

Необходимо также отметить, что матрица «молекула-признак», а значит, и качество QSAR-модели, сильно зависит как от выбора дополнительной классификации атомов, так и от выбора разбиений на интервалы. Очевидно, что даже небольшой сдвиг заданных интервалов разбиения отрезка расстояний может привести к тому, что структурные фрагменты будут отнесены к другому дескриптору, сильно изменив описание молекулы. В связи с этим встает задача оптимизации выбора как классификации атомов, так и выбора интервалов разбиения.

3D-QSAR анализ трехмерных молекул

Построение зависимостей «структура-свойство» с использованием пространственного представления молекул обучающей выборки получило название 3D-QSAR. Основные этапы 3D-QSAR моделирования можно представить следующим образом:

1. Строятся пространственные представления молекул обучающей выборки, и проводится их «выравнивание» (alignment) согласно заданным правилам выбора ориентаций.
2. Для всех молекул вычисляется набор пространственно зависимых признаков.
3. Строится функция, выражающая зависимость вычисленных признаков и исследуемого значения химико-биологической активности.
4. Определяется устойчивость и предсказательную способность найденной функциональной зависимости;
5. При необходимости модель модифицируется, повторив пункты 1-4.

Наиболее распространенным методом 3D-QSAR является метод сравнительного анализа молекулярных полей CoMFA (Comparative Molecular Field Analysis), впервые предложенный Р. Крамером с соавторами в [020]. За последние годы метод CoMFA нашел широкое применение при построении моделей «структура-свойство» для биологической активности [34, 35, 36].

При классическом CoMFA моделировании после выполнения процедуры пространственного выравнивания каждая молекула помещается в трехмерную регулярную решетку с заданным достаточно маленьким шагом. В узлах решетки вычисляются значения стерического (ван-дер-ваальсова) и электронного взаимодействия между данной молекулой и «пробным» атомом, имеющим свойства атома углерода в состоянии с зарядом «+1». Получаются две трехмерные матрицы, каждая из которых является дискретным представлением стерического и электронного поля исследуемой молекулы. Значения элементов этих матриц являются признаками молекул обучающего множества:

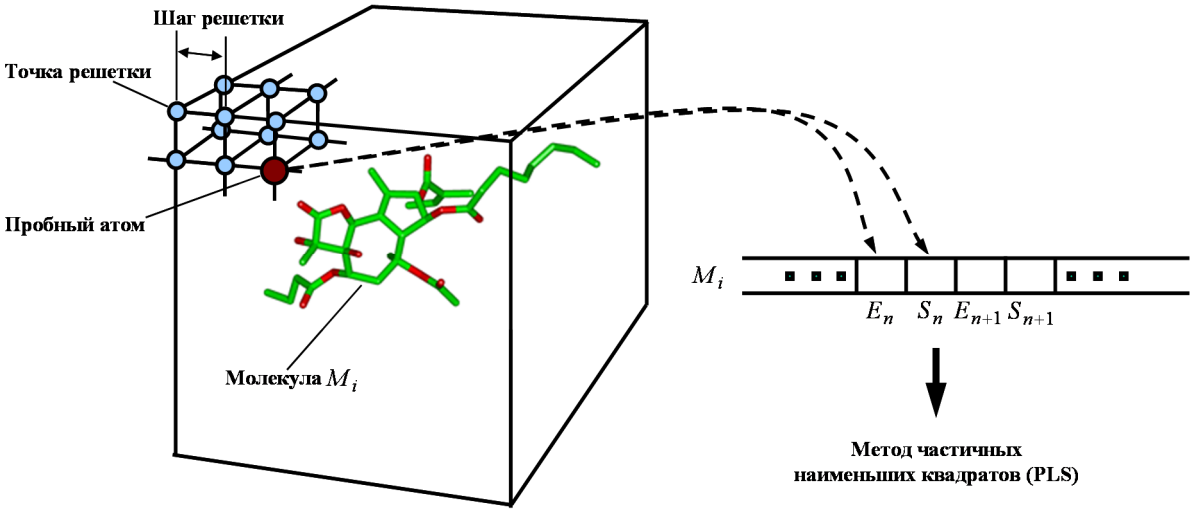


Рис 3. Построение матрицы «молекула-признак» в алгоритме CoMFA.

В результате формируется очень широкая ($M >> N$) таблица «молекула-признак», которая затем анализируется одним из регрессионных методов. В [33] предложено использовать метод частичных наименьших квадратов (PLS, Partial Least Square). Фактически, PLS строит регрессионное уравнение «структура-свойство» на главных компонентах (латентных переменных) таблицы «молекула-признак» [037], а оптимальное число латентных переменных, включаемых в уравнение, определяется на основе процедуры скользящего контроля (cross-validation) [38]. Эта процедура обеспечивает формирование прогностически устойчивых моделей.

Одним из важных достоинств метода является возможность (после возврата к исходным признакам) определения пространственных участков вокруг молекулы, изменения стерического и/или электронного поля в которых приводит к существенным изменениям в биологической активности, т.е. содержательной интерпретации полученной модели

Развитием CoMFA является метод сравнительного анализа индексов молекулярного сходства CoMSIA (Comparative Molecular Similarity Index Analysis) [039, 40]. В этом методе в узлах решетки вычисляется широкий набор численных свойств (а именно, значения электростатического, стерического взаимодействия, липофильности, а также свойства донора-акцептора водородной связи). Этот набор свойств затем с помощью заданной функции специального вида трансформируется в единый индекс молекулярного сходства, который и помещается в матрицу «молекула-признак».

При правильном выборе метода выравнивания молекул в регулярной решетке методы CoMFA и CoMSIA могут показывать хорошие результаты классификации [041], что сделало такие методы стандартом для прикладных исследований. На основе CoMFA были разработаны еще несколько сходных методов 3D-QSAR [042, 43].

Наиболее сложным моментом в использовании CoMFA и подобных ему методов является необходимость «пространственной нормализации» молекул обучающей выборки, т.е. выбор их расположения (после взаимного пространственного выравнивания) относительно системы координат регулярной решетки. Так, было обнаружено, что даже изменение положения системы координат (например, ее простой поворот) может привести к падению прогностической способности модели в два раза [44].

Кроме этого, при увеличении сложности и гибкости молекул обучающего множества, методы 3D-QSAR, как правило, становятся неприменимы. При усложнении молекул резко возрастает число возможных способов их связывания с рецептором, что сильно затрудняет (а в ряде случаев делает

невозможным) выбор однозначного выравнивания молекул в регулярной решетке.

1.2.2. Анализ МД-матрицы.

Пусть имеем матрицу «молекула – дескриптор», где по строкам находятся значения всех дескрипторов для данной молекулы, а по столбцам значения данного дескриптора для всех молекул. Кроме того дан вектор активности, компоненты которого – значения активности для каждой молекулы.

Необходимо по значениям дескрипторов для данной молекулы установить значение её активности. Используя обучающую выборку, будем строить модели, предсказывающие активность молекул. Для оценки прогностической способности моделей можно использовать коэффициент так называемого «скользящего контроля» (cross validation) [8], [9]. Возможны и другие подходы к оценке качества моделей. Например, когда выборка разбивается на 2 части, на первой из которых идёт построение моделей, а на второй вычисляется процент успешно спрогнозированных молекул.

Решение задачи поиска функциональной зависимости разобьём на несколько шагов:

- кластеризация и обработка выбросов;
- отбор значимых дескрипторов;
- построение модели;
- прогноз.

Для каждого из этапов в литературе предложено множество реализаций. Кратко охарактеризуем каждый этап.

Кластеризация (кластерный анализ) — задача разбиения заданной выборки объектов на подмножества, называемые кластерами, так, чтобы каждый

кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. В задаче «структура-свойство» кластеризация используется, во-первых, для предварительного анализа выборки, разделения её на классы сходных веществ и построения классифицирующих моделей локально – внутри каждого кластера, а во-вторых для осуществления отказа от прогноза в случае, когда новая молекула не принадлежит ни одному из кластеров. Выделяют следующие основные методы кластерного анализа: K-средних (K-means), графовые алгоритмы кластеризации, статистические алгоритмы кластеризации, иерархическая кластеризация и другие. Далее в главе 2 мы опишем подробно некоторые алгоритмы при построении правил отказа от прогноза.

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.

Предложено большое число алгоритмов, которые решают данную проблему. Отметим, например, факторный анализ (метод главных компонент) [9, 12] или различные варианты перебора сочетаний дескрипторов. В данной работе в качестве метода отбора значимых дескрипторов будет использоваться метод группового учёта аргументов (МГУА) [mgua]. Метод основан на рекурсивном селективном отборе моделей, на основе которых строятся более сложные модели. Точность моделирования на каждом следующем шаге рекурсии увеличивается за счет усложнения модели.

Наконец, осталось сказать несколько слов о построении моделей. Здесь фактически можно использовать весь аппарат теории распознавания образов. Например, алгоритмы распознавания, основанные на вычислении оценок или алгоритмы голосования по тупиковым тестам, метод *k* ближайших соседей или нейросетевая модель распознавания с обратным распространением, метод опорных векторов (support vector machine) или алгоритмы статистического взвешенного голосования. Наиболее популярным и простым остаётся регрессионный анализ и построение модели на основе линейной регрессии. Также в этой работе будет использоваться модель на основе так называемого дерева решений.

Подробное описание алгоритмов и параметров построенных моделей можно посмотреть в главе 3.

Выводы

Исходной информацией QSAR-задачи являются описания объектов в виде векторов значений признаков $x_i = (x_{i1}, \ldots x_{iM}),$ где признаки x_i, $i = 1, \ldots N$, характеризуют различные стороны - свойства объекта. У объектов существует «целевое свойство», которое для части объектов предполагается известным, а для части объектов нет. В QSAR-задаче - Как и в задаче распознавания (прогноза, идентификации, "классификации с учителем") - определение значений свойств происходит по информации, заложенной в обучающей или эталонной выборке.

Таким образом, 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. Построения нечёткой классифицирующей функции и правил отказа от прогноза в задаче «структура - свойство»

Глава содержит строгие постановки решаемых задач, а также теоретическое описание методов, предложенных в рамках подхода. Описаны методы построения быстрых правил отказа от прогноза, обсуждается выбор формы области допустимых аргументов классифицирующей функции, как базового элемента правил отказа от прогноза. Затем предлагается обобщение модели с помощью нечёткой кластеризации, описано построение нечёткого классификатора и оптимизация прогнозирующей модели по параметрам нечёткой классификации.

2.1. Определения и постановка задач

М-граф (меченый молекулярный граф $G = \{E, V\}$) – это помеченный граф, вершины которого $\{E\}$ интерпретируются как атомы молекулы, а ребра $\{V\}$ – 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, 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.) Предполагаем, что различным с точки зрения химии соединениям соответствуют различные М-графы.

Обучающая выборка $\{(G_i, C_i)\}_{i=1}^{N}$ - совокупность из N химических соединений, где:

– i-ое соединение представлено меченым молекулярным графом G_i;

– *i*-ое соединение отнесено к C_i - одному из H классов активности (например, «активных», «слабоактивных», «неактивных» веществ) согласно исследуемому свойству.

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

Алфавитом дескрипторов будем называть 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 $x = (x_1, \dots x_M) \in R^M$, где x_i - значение *i*-ого дескриптора, вычисленное для *G*. *Описывающим отображением* $D : \{G\} \to R^M$ назовём отображение, ставящее в соответствие *M*-графу его вектор признаков. Пространство R^M в данном случае будем называть *пространством дескрипторов*.

МД-матрицей или матрицей «молекула – дескриптор» (матрицей признаков) для рассматриваемой обучающей выборки $\{(G_i, C_i)\}_{i=1}^N$ будем называть матрицу *X* размера *N x M*, в *i*-ой строке которой стоит вектор признаков $\bar{x}_i = (x_{i1}, .., x_{iM})$ *i*-ого соединения.

Назовём *классифицирующей* функцию $F : R^M \to \{C_i\}_{i=1}^H$, получающую в качестве аргумента вектор признаков $x = (x_1, \dots x_M) \in R^M$ произвольного молекулярного графа *G*, и относящую соответствующее этому *M*-графу соединение к одному из классов активности C_i. Иногда удобно, чтобы классифицирующая функция была задана на множестве молекулярных графов. Когда в качестве аргумента *F* указан *M*-граф, надо понимать, что *F* вычисляется на соответствующем векторе признаков. Положим по определению $F(G_i) := F(D(G_i))$, где D - описывающее отображение из

множества *M*-графов в пространство дескрипторов R^M. Предполагаем, что функция $F: R^M \to \{C_i\}_{i=1}^H$ принадлежит некоторому заранее заданному классу функций Ф.

Введём функционал качества φ(F) на Ф:

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

- процент верно классифицированных функцией *F* молекул из обучающей выборки.

Задача «структура – свойство» заключается в следующем:

Задана обучающая выборка $\{(G_i, C_i)\}_{i=1}^N$, задача «структура – свойство» включает задачу построения описания обучающей выборки - выбор алфавита дескрипторов, построение описывающего отображения $D: \{G\} \to R^M$ и формирование матрицы «молекула–дескриптор» *X* для обучающей выборки, а также задачу поиска функциональной зависимости – построение классифицирующей функции $F: R^M \to \{C_i\}_{i=1}^H$, максимизирующей значение функционала качества *φ(F)* на Ф.

Пусть зафиксирован алгоритм *Alg* построения классифицирующей функции *F* по обучающей выборке $\{(G_i, C_i)\}_{i=1}^N$, *прогнозирующей моделью* назовём совокупность обучающей выборки $\{(G_i, C_i)\}_{i=1}^N$ и алгоритма *Alg* построения классифицирующей функции, $F = Alg(\{(G_i, C_i)\}_{i=1}^N)$.

Используя обучающую выборку, строятся прогнозирующие модели, предсказывающие активность молекул. Для оценки прогностической способности моделей используется коэффициент скользящего контроля R_{cv}^2 (cross validation) [8, 9].

$R^2_{cv}(\{(G_i, C_i)\}_{i=1}^N, F = Alg(\{(G_i, C_i)\}_{i=1}^N))$ - вычисляется по фиксированной выборке и алгоритму построения классифицирующей функции в ходе следующей процедуры:

– в цикле по числу молекул из выборки удаляется текущая молекула;

– по оставшимся в обучающей выборке молекулам строится модель;

– с помощью модели предсказывается значение активности удалённой молекулы.

Процент успешных прогнозов в этом случае и есть коэффициент R^2_{cv}.

Для построения качественной (в смысле коэффициента R^2_{cv}) классифицирующей функции F чрезвычайно важно ограничить область её допустимых аргументов (в данном случае молекулярных графов).

Сформулируем теперь задачу построения правил отказа от прогноза:

Правилом отказа назовём одну или несколько функций $g : R^M \to \{0,1\}$ со следующей интерпретацией: $g(G_i) = 1$ будет означать отказ от прогноза активности данного молекулярного графа, в противном случае прогноз может быть осуществлён.

Как и ранее полагаем $g(G_i) := g(D(G_i))$, где D - описывающее отображение из множества *M*-графов в пространство дескрипторов R^M.

Назовём молекулярный граф G_i *допустимым*, если согласно принятым правилам отказа, он принадлежит области допустимых аргументом классифицирующей функции F. То есть $g(G_i) = 0$.

Пусть $O = \{(G_i, C_i)\}_{i=1}^N$ - обучающая выборка, обозначим

$\tilde{O} = \{(G_i, C_i)\}_{i=1}^{N} \setminus \{(G_j, C_j) \mid g(G_j) = 1, j = 1, \ldots N\}$ - выборка, составленная только из допустимых М-графов обучающей выборки *O*.

Назовём правило отказа $g : R^M \to \{0,1\}$ *сильным*, если для него выполнено неравенство $R^2_{cv}(O, Alg) < R^2_{cv}(\tilde{O}, Alg)$.

2.2. Построение правил отказа

Идея состоит в том, чтобы использовать кластерную структуру исходной обучающей выборки для построения правил отказа от прогноза для данного химического соединения. Здесь речь идёт не только о выбросах, которые просто не попадают ни в один кластер, но и о соединениях, предсказывать активность которых не следует по более тонким соображениям кластерной структуры. Например, молекулы, принадлежащие в равной степени двум кластерам, модели которых предсказывают её активность по-разному.

Кроме того, важным моментом является необходимость определять допустимость молекулярного графа с минимальными вычислительными затратами. Поэтому предлагается вычислять правила отказа на специальном пространстве дескрипторов, гораздо меньшей размерности, чем исходное, например, только топологических.

Таким образом, строится 2 пространства дескрипторов, одно – для построения правил отказа, другое – для классификации и прогноза активности. Здесь естественным образом возникает *сокращённая (специальная) матрица «молекула - дескриптор»,* строки которой являются векторами в специальном пространстве дескрипторов.

Итак, пусть имеется сокращённая матрица «молекула - дескриптор» *X* и интерпретация её строк $x_i = (x_{i1}, \ldots x_{iM})$ как векторов пространства R^M. Пусть также в пространстве R^M зафиксированна метрика ρ. Для

определения числа кластеров предлагается использовать алгоритм построения минимального покрывающего дерева.

Пусть для сокращённой матрицы «молекула - дескриптор» вычисленны расстояния между любыми её двумя строками как точками в пространстве R^M в метрике ρ. Выберем точку и начнем нарашивать “минимальное покрывающее дерево” — сначала добавим к ней ближайшую точку из оставшихся. Затем добавим точку, ближайшую к ним обоим. Теперь точки понимаются как вершины графа, а при добавлении новой вершины проводится ребро от неё до ближайшей вершины графа. В каждый момент имеем дерево, частично покрывающее наши точки, и ищем следующую вершину, которая менее всего удалена от этого дерева.

Поле работы алгоритма имеем матрицу $\Xi = [\xi_{ij} \rho(x_i, x_j)]$, где

$$\xi_{ij} = \begin{cases} 1, & \text{если точки } x_i \text{ и } x_j \text{ связанны ребром,} \\ 0, & \text{иначе.} \end{cases}$$

$\rho(x_i, x_j)$ — расстояние между точками x_i и x_j.

Для разбиения множества точек на кластеры выберем *пороговое значение R*. Теперь удалим из дерева все рёбра, длины которых больше *R.* В результате множество точек (наше дерево) разобьётся на компоненты связности. Компоненты, состоящие лишь из одной точки, будем считать выбросами, остальные являют собой искомые кластеры. Выбрать пороговое значение можно проанализировав, например, гистограмму длин рёбер минимального покрывающего дерева или положив значение *R* равным средней длине ребра (удаление только самых длинных рёбер не приводит, как правило, к выявлению кластеров, а лишь отсекает выбросы).

Теперь, когда число кластеров известно и выбросы отсутствуют, можно провести более тонкую разбивку на кластеры. Для этого предлагается

использовать алгоритм k-средних с ядрами. Это модификация широко известного алгоритма k-средних [6]. Отличие заключено в том, что центром кластера считаются ядро, состоящее из q точек обучающей выборки, а не одна точка пространства R^M. Число точек в ядре q является параметром для данной модели построения правил отказа.

Пусть найдена кластерная структура исходной обучающей выборки с учётом удаления выбросов. Пусть, как и ранее, число кластеров k, и известна *матрица чёткого разбиения*:

$$S = [\nu_{ij}], \ \nu_{ij} \in \{0,1\}, \ i \in \{1,\dots N\}, \ j \in \{1,\dots k\},$$

$$\sum_{j=1}^{k} \nu_{ij} = 1, i \in \{1,\dots N\}, \quad 0 < \sum_{i=1}^{N} \nu_{ij} < N, j \in \{1,\dots k\},$$

в которой *i*-ая строчка содержит информацию о принадлежности объекта $(x_{i1},\dots x_{iM})$ к одному из кластеров $S_1,\dots S_k$.

Алгоритм заключается в следующем:

В каждом из кластеров запускаем алгоритм k-средних с k = q – число ядер. Обозначим $Z_i = \{c_{i1},\dots c_{iq}\}$ - центр *i*-го кластера S_i, состоящий из q ядер.

Определим расстояние от *l*-ой точки до *i*-го кластера как

$$\rho(x_l, S_i) = \rho(x_l, Z_i) = \rho(x_l, \{c_{i1},\dots c_{iq}\}) = \min_{j\in\{1,\dots q\}} \rho(x_l, c_{ij})$$

Разбиваем заново множество точек на кластеры. Если

$\rho(x_l, Z_p) = \min_{i\in\{1,\dots k\}} \rho(x_l, Z_i)$, то $x_l \in S_i$ - включаем точку x_l в кластер S_i.

3) Проверяем условия:

- $\| S - S^* \| = 0$, где S^* - матрица чёткого разбиения на предыдущей итерации алгоритма.

- число итераций больше, чем максимально допустимое.

Если хоть одно из них верно, то завершаем алгоритм, иначе переходим к шагу 1.

Для каждого кластера определим его радиус $r_i = \max_{x_j \in S_i} \rho(x_j, Z_i)$.

Сформулируем правила отказа. Пусть $\tilde{x}$ - вектор признаков молекулярного графа G химического соединения, активность которого необходимо предсказать.

$g_1(\tilde{x}) = (\min_{i \in \{1,\dots N\}} \rho(\tilde{x}, x_i) > R)$, то есть если $\min_{i \in \{1,\dots N\}} \rho(\tilde{x}, x_i) > R$, где R – пороговое значение, то отказываемся от прогноза (новая молекула *чужая* для всей выборки).

$g_2(\tilde{x}) = (\rho(\tilde{x}, Z_i) > r_i)$, то есть если $\rho(\tilde{x}, Z_i) > r_i$ - отказ от прогноза в i-ом кластере (молекула *чужая* для данного кластера).

Окончательно имеем:

$$g(\tilde{x}) = (\rho(\tilde{x}, Z_i) > r_i) \vee (\min_{i \in \{1,\dots N\}} \rho(\tilde{x}, x_i) > R)$$

.

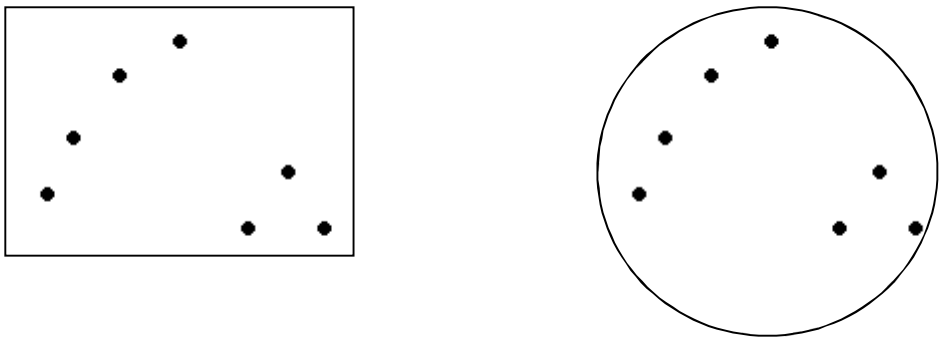
Таким образом, соединения, заведомо отличные от используемых при построении модели, не будут с помощью неё предсказываться, что существенно улучшит качество прогноза.

Конечно, утверждать, что для построенного правила отказа выполнено $R^2_{cv}(O, Alg) < R^2_{cv}(\tilde{O}, Alg)$ мы не можем. Так как указанное неравенство существенно зависит от алгоритма построения классифицирующей функции. Но возможны более слабые оценки.

2.3. Форма области допустимых аргументов классифицирующей функции

Проиллюстрируем вышесказанное. При построении классифицирующей функции F на всём пространстве R^M, очень сложно добиться от неё точности в области, где не лежат точки обучающей выборки. Таким образом, возникает необходимость ограничить область допустимых аргументов классифицирующей функции.

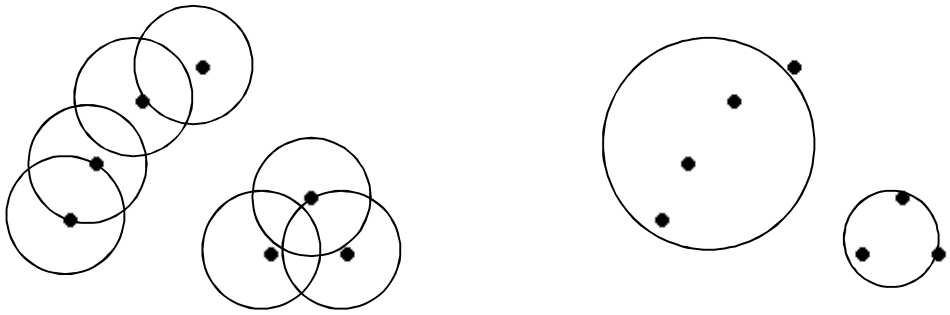
Простейшими ограничениями могут быть многомерный параллелепипед и сфера, ограничивающие обучающую выборку как множество точек в R^M. Для построения сферы аналогично тому, как мы это делали на кластерах, можно определить центр и радиус всей выборки.



а) б)

Рис 4. Ограничение параллелепипедом и сферой

Если в выборке присутствуют выбросы или же она занимает слишком большой объём в пространстве, простейшие ограничения неудобны и не дают результата. Тогда предлагается использовать несложные алгоритмы кластеризации. Отметим, что подход к описанию кластера – его центром (одной или несколькими точками R^M) и радиусом, заведомо определяет форму области допустимых аргументов классифицирующей функции как объединение конечного числа сфер в R^M.



а) б)

Рис 5. Результат работы минимального покрывающего дерева и k-средних с k=2

Наконец в этой работе рассмотрен подход, выбирающей в качестве области допустимых аргументов классифицирующей функции F пересечение областей 5.а) и 6.

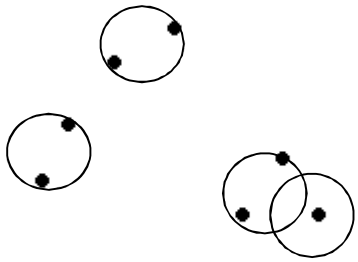


Рис 6. Результат работы алгоритма k-средних с 2-мя ядрами и k=2

Не всегда полезно так сильно ограничивать исходное пространство объектов. Дальнейшим развитием данного подхода является переход к нечётким функциям принадлежности точки кластеру, что позволяет увеличить радиус кластера, не рискуя получить заведомо ложный прогноз. Интерес представляет сравнение качества классифицирующих моделей при различных ограничениях на область допустимых аргументов классифицирующей функции.

2.4. Построение нечёткого классификатора

Для прогнозирования активности данного химического соединения предложен следующий подход. Пусть $x_i = (x_{i1}, \ldots x_{iM})$ - описание i-го объекта, y_i - значение его активности. Тогда имеем матрицу $X = [x_{ij}]$, называемую в соответствии с терминологией МД-матрицей или матрицей «молекула – дескриптор», а также вектор активности $(y_1, \ldots y_N)$.

Имея описание обучающей выборки в виде МД-матрицы *X*, а также интерпретация строк данной матрицы $x_i = (x_{i1}, \ldots x_{iM})$ как векторов пространства R^M. Пусть также в пространстве R^M зафиксированна метрика ρ. Применяем некоторый алгоритм нечёткой кластеризации (c-means fuzzy или любой другой) [6]. Нечеткие методы кластеризации, в отличие от четких методов, позволяют одному и тому же объекту принадлежать одновременно нескольким кластерам, но с различной степенью. Нечеткая кластеризация во многих ситуациях более «естественна», чем четкая, например, для объектов, расположенных на границе кластеров.

Нечёткие кластеры опишем следующей *матрицей нечёткого разбиения*

$$S = [\mu_{ij}], \ \mu_{ij} \in [0,1], \ i \in \{1, \ldots N\}, \ j \in \{1, \ldots k\},$$

в которой *i*-ая строчка содержит степени принадлежности объекта $(x_{i1}, \ldots x_{iM})$ к кластерам $S_1, \ldots S_k$. Единственным отличием матрицы нечёткого разбиения от соответствующей матрицы чёткого разбиения то, что при нечетком разбиении степень принадлежности объекта к кластеру принимает значения из интервала [0, 1], а при четком - из двухэлементного множества {0, 1}.

Матрица *S* должна соответствовать следующим условиям [6]:

$$\sum_{j=1}^{k} \mu_{ij} = 1, i \in \{1, \dots N\}, \quad (3)$$

$$0 < \sum_{i=1}^{N} \mu_{ij} < N, j \in \{1, \dots k\}. \quad (4)$$

Нечёткое разбиение позволяет просто решить проблему объектов, расположенных на границе двух кластеров - им назначают степени принадлежностей равные 0.5. Недостаток нечеткого разбиения проявляется при работе с объектами, удаленными от центров всех кластеров. Удаленные объекты имеют мало общего с любым из кластеров, поэтому интуитивно хочется назначить для них малые степени принадлежности. Однако, по условию (3) сумма их степеней принадлежностей такая же, как и для объектов, близких к центрам кластеров, т.е. равна единице. Для устранения этого недостатка можно использовать нечёткое разбиение, которое требует, только чтобы произвольный объект принадлежал хотя бы одному кластеру. Нечёткое разбиение получается следующим ослаблением условия (3): $\exists j, \quad \mu_{ij} > 0 \quad \forall i$.

Для оценки качества разбиения используется критерий разброса, показывающий сумму расстояний от объектов до центра своего кластера. Для евклидового пространства этот критерий записывается так [6]:

$$\sum_{j=1}^{k} \sum_{i=1}^{N} (\mu_{ij})^m \| x_i - v_j \|^2, \quad (5)$$

$$v_j = (\sum_{i=1}^{N} (\mu_{ij})^m) \cdot x_i) / (\sum_{i=1}^{N} (\mu_{ij})^m)), \quad x_i = (x_{i1}, \dots x_{iM}),$$

$m \in [1, +\infty]$ – экспоненциальный вес, определяющий нечеткость, размазаность кластеров.

Предложено множество алгоритмов нечеткой кластеризации, основанных на минимизации критерия (5). Нахождение матрицы нечеткого разбиения с минимальным значение критерия (5) представляет собой задачу нелинейной оптимизации, которая может быть решена разными методами. Наиболее известный и часто применяемый метод решения этой задачи алгоритм нечетких с-средних [6], в основу которого положен метод неопределенных множителей Лагранжа. Он позволяет найти локальный оптимум, поэтому выполнение алгоритма из различных начальных точек может привести к разным результатам.

Теперь, имея разбивку исходного пространства на нечёткие кластеры, внутри каждого строим локальную классифицирующую модель (далее предполагается, что это линейная регрессия, но может быть использован и любой другой алгоритм) [1, 2]. Пусть для простоты имеем 2 возможных значения активности: активна/неактивна, обозначим их соответственно числовыми значениями 1 и -1.

Для нового соединения $\tilde{x} = (\tilde{x}_1, \ldots \tilde{x}_M)$ у нас есть k предсказаний активности в соответствии с числом кластеров (моделей). Пусть i-ая модель дала предсказание R_i, тогда можно вычислить результирующее предсказание по формуле:

$$\tilde{y} = \frac{\sum_{i=1}^{k} R_i \mu_i}{k},$$

где μ_i - коэффициент принадлежности данной молекулы к i-ому кластеру.

Можно задать рамки нормировки ответа $\tilde{y}$, например, $\tilde{y} < -0,5 \Rightarrow \tilde{y} = -1$, $\tilde{y} > 0,5 \Rightarrow \tilde{y} = 1$, иначе $\tilde{y} = 0$ - отказ от прогноза.

2.5. Параметры нечёткой классификации

Рассмотрим теперь задачу оптимизации нечёткой классифицирующей функции по параметрам нечёткой кластеризации.

В работе [5] подробно описаны некоторые методы построения кластерной структуры обучающей выборки. Нас будет интересовать «нечёткое» обобщение этой кластерной структуры для применения описанного подхода.

Пусть найдена кластерная структура исходной обучающей выборки с учётом удаления выбросов. Пусть, как и ранее, число кластеров k, и известна *матрица чёткого разбиения*:

$$S = [\nu_{ij}],\ \nu_{ij} \in \{0,1\},\ i \in \{1,\dots N\},\ j \in \{1,\dots k\},$$

$$\sum_{j=1}^{k} \nu_{ij} = 1, i \in \{1,\dots N\},\ 0 < \sum_{i=1}^{N} \nu_{ij} < N, j \in \{1,\dots k\},$$

в которой *i*-ая строчка содержит информацию о принадлежности объекта $(x_{i1},\dots x_{iM})$ к одному из кластеров $S_1,\dots S_k$.

Пусть также каждый кластер задан своим центром $Z_i = \{c_{i1},\dots c_{iq}\}$ - некоторым подмножеством множества точек кластера S_i, точки центра будем называть ядрами данного кластера, и радиусом $r_i = \max_{x_j \in S_i} \rho(x_j, Z_i)$.

Построим матрицу нечёткого разбиения $\tilde{S} = [\mu_{ij}]$, в которой *i*-ая строчка содержит степени принадлежности объекта $(x_{i1},\dots x_{iM})$ к кластерам $\tilde{S}_1,\dots \tilde{S}_k$. Параметрами оптимизации будут являться $\lambda_1, \lambda_2 \in R$, $\lambda_1 \le 1, \lambda_2 \ge 1$.

Определим малый и большой радиус кластера $\tilde{S}_i$, как $r_i^1 = \lambda_1 r_i$ и $r_i^2 = \lambda_2 r_i$ соответственно. Тогда элементы матрицы $\tilde{S} = [\mu_{ij}]$ вычислим по формуле:

$$\mu_{ij} = \begin{cases} 1, & \text{если } \rho(x_i, Z_j) < r_i^1, \\ 0, & \text{если } \rho(x_i, Z_j) > r_i^2, \\ \dfrac{r_j^2 - \rho(x_i, Z_j)}{r_j^2 - r_j^1}, & \text{иначе.} \end{cases}$$

Функция принадлежности точки кластеру может быть также нелинейной и содержать дополнительные параметры оптимизации. На рисунке графики различных функций принадлежности.

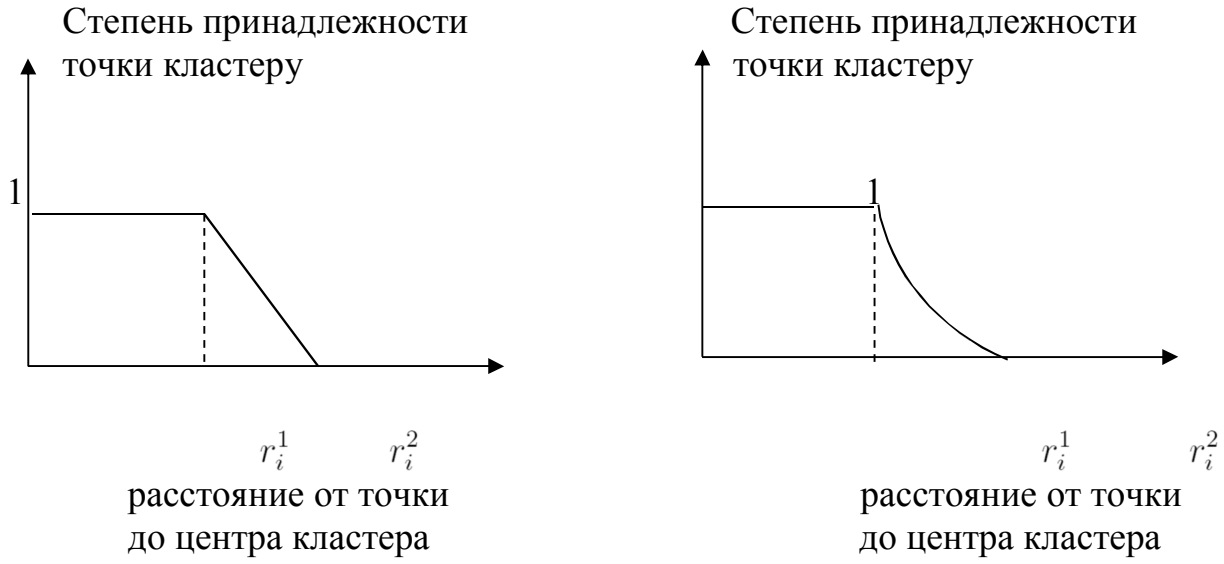


Рис 7. Линейная и нелинейная функции принадлежности точки кластеру

Так как чёткая кластеризация является в нашем случае частным случаем нечёткой при значениях параметров $\lambda_1 = \lambda_2 = 1$, то переход к этим функциям не ухудшит качество прогноза. Данный подход позволяет нам содержательно

использовать кластерную структуру выборки, не ограничиваясь при этом пределами лишь одного кластера. На рисунке приведен пример нечёткой кластеризации полученной из заданного разбиения множества точек на чёткие кластеры.

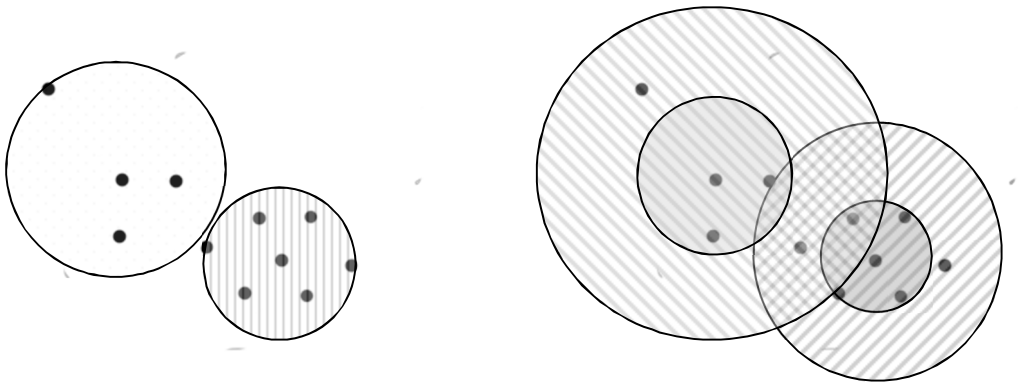


Рис 8. Чёткая и нечёткая кластерные структуры

«Нечёткость» классифицирующей функции на границах кластеров позволяет прогнозировать активность одного и того же соединения в разных локальных моделях, и получать ответ взвешенным голосованием. Наблюдается значительное улучшение качества прогноза за счёт кусочной линейности строящихся моделей, а также оптимизации моделей по параметрам $\lambda_1, \lambda_2 \in R$.

Выводы

Кластерный анализ обучающей выборки фактически определяет построение классифицирующей функции. Для скрининга больших баз соединений чрезвычайно важно построить правила отказа от прогноза, к тому же необходимо, чтобы они были быстрыми с вычислительной точки зрения. Нечёткая кластеризация позволяет устранить основные недостатки классических методов и предлагает возможность выбора классифицирующей функции из более широких, параметризованных классов.

Глава 3. Методы и алгоритмы решения задачи

В данной главе раскрывается этап анализа МД-матрицы, изложена схема проводимых вычислений, рассматриваются применяемые методы и реализация алгоритмов. Описан метод эволюционного отбора дескрипторов, методы построение локальной линейной модели и решающего дерева для построения прогнозирующих моделей. Проводится оценка вычислительной сложности алгоритмов. Обсуждается среда разработки и архитектура программ.

3.1. Общая схема решения

3.1.1. Этап формирования дескрипторов и построения матрицы «молекула - дескриптор»

Для построения матрицы «молекула - дескриптор» использовались структурные 3D-дескрипторы [14]. Для вычисления их численных значений введено понятие соответствия структурного дескриптора D и химической функциональной группы G [14]. Значением структурного дескриптора является количество фрагментов (химических функциональных групп) молекулярного графа, соответствующих данному дескриптору.

Кроме структурных 3D-дескрипторов, при формировании матрицы также использовался ряд скалярных дескрипторов – общих химико-физических характеристик молекул. Среди них - молярный вес, объем, рефракция, поверхностное натяжение, плотность, диэлектрическая постоянная, поляризуемость и пр.

3.1.2. Этап поиска функциональной зависимости

На входе описанные ниже методы имеют вычисленную матрицу «молекула - дескриптор» и записанный отдельно столбец активности соединений (целевое свойство) в текстовом формате. Функционалом качества в этом случае является коэффициент скользящего контроля [8, 9] –

в цикле по числу молекул из выборки удаляется одно соединение, по оставшимся строится классифицирующая (прогнозирующая) функция, используя которую прогнозируют активность удалённого на этом шаге соединения.

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

В отдельном случае выборка может быть дана как набор из двух матриц – обучающее множество и проверочное множество. Тогда классифицирующая функция строится по обучающему множеству, а указанный функционал качества вычисляется на проверочном множестве.

В данной работе, предложен следующий подход к поиску функциональной зависимости (построению классифицирующей функции) при анализе матрицы «молекула - дескриптор».

Сначала с помощью алгоритмов кластерного анализа (предлагаются алгоритмы k-means [6] и иерархическая кластеризация [13]) выборка разбивается на чёткие кластеры. Вводятся два параметра – малый радиус кластера и большой радиус кластера, варьируя которые получаем семейство нечётких кластерных структур на базе исходного чёткого разбиения. Сокращая число дескрипторов эволюционным отбором, в дальнейшем называемым МГУА [mgua], в каждом кластере строим локальную прогнозирующую модель. Предсказываем активность нового соединения взвешенным (коэффициентами принадлежности к нечётким кластерам) предсказанием локальных моделей. В качестве локальных моделей используем линейную регрессию и решающее дерево.

На рисунке представлена схема предлагаемого подхода. Далее охарактеризуем каждый этап подробнее. В главе 4 приведены результаты

применения этих алгоритмов, а также сравнение их с алгоритмами, опускающими какой-либо компонент этой схемы.

Рис 9. Схема этапа поиска функциональной зависимости

3.2. Анализ кластерной структуры выборки

Для анализа кластерной структуры использовались алгоритмы К-средних и его модификации, уже описанные нами в 2.2, а также встроенные средства Matlab, такие как иерархический кластерный анализ [13].

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие. Для вычисления расстояния между объектами

используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний.

В методе, который использовался, реализуется иерархический агломеративный алгоритм, смысл которого заключается в следующем. Перед началом кластеризации все объекты считаются отдельными кластерами, в ходе алгоритма они объединяются. Вначале выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате количество кластеров становится равным N-1. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования зависит от способов вычисления расстояния между объектами и определения близости между кластерами.

3.3. Эволюционный отбор дескрипторов

Мы называем эволюционный отбор дескрипторов методом группового учёта аргументов (МГУА) вне зависимости от алгоритма, с помощью которого строятся классифицирующие функции. Под МГУА понимается следующий алгоритм.

Фиксируется некоторое число Q - количество отбираемых на каждом этапе дескрипторов (может разниться от этапа к этапу). А также выбирается критерий остановки алгоритма – им может являться ограничение числа селекций.

Затем с помощью зафиксированного алгоритма строятся всевозможные прогнозирующие функции на M матрицах «молекула - дескриптор», каждая из которых состоит лишь из одного столбца-дескриптора.

Оцениваем множество всех построенных на данный момент функций при помощи функционала качества φ и выбираем из них Q наилучших. Соответственно отбираются Q столбцов-дескрипторов (данный набор

дескрипторов, отобранных на определенном этапе, называется *селекцией*), причем для каждого шага число *Q* может иметь отдельное значение.

Затем к каждой из отобранных Q матриц «молекула - дескриптор» добавляется новый столбец из исходной матрицы. Имеем Q*M матриц (наборов дескрипторов или прогнозирующих функций). Отбираем вновь лучшие с точки зрения функционала качества функции.

На каждом этапе мощность отобранных наборов дескрипторов увеличивается на 1. Алгоритм работает пока не достигнуто условие остановки алгоритмы.

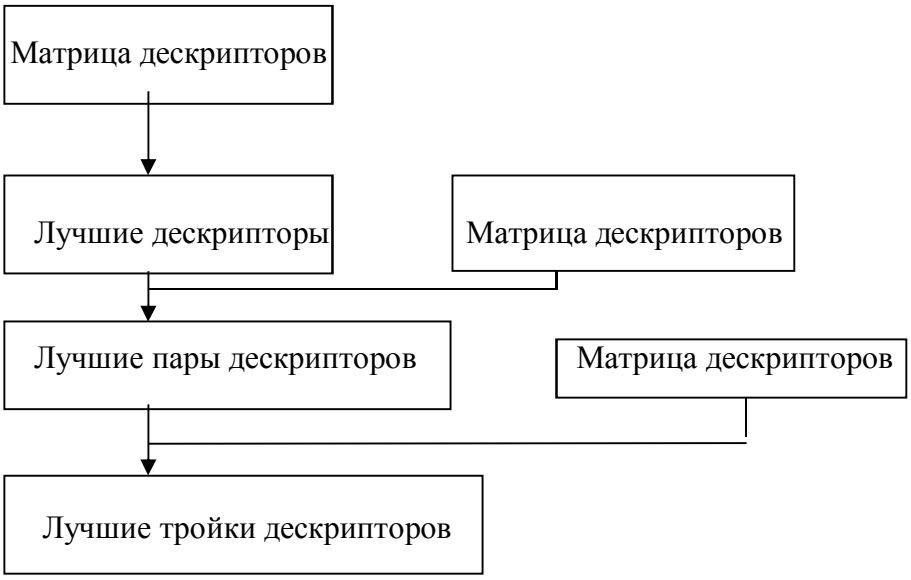


Рис. 10 Схема этапа эволюционного отбора дескрипторов.

Вычислительная сложность этого метода значительно меньше, чем у полного перебора - это имеет большое значение из-за большого числа дескрипторов (> 1000). Гарантируется нахождение наиболее точной модели - метод не пропускает наилучшего решения во время перебора всех вариантов (в заданном классе функций). 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.

3.4. Локальные прогнозирующие модели

В главе 4 рассматриваются результаты, полученные с помощью двух алгоритмов построения локальных классифицирующих моделей. Это линейная регрессия [15] и решающее дерево [16]. Эти алгоритмы широко известны и реализованы в среде Matlab. Отметим, что вместо них мог быть использован любой алгоритм распознавания образов, например, метод *k* ближайших соседей или нейросетевая модель распознавания [6]. Охарактеризуем кратко использованные алгоритмы.

3.4.1. Линейная регрессия

На практике линия регрессии чаще всего ищется в виде линейной функции $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_N X_N$ (линейная регрессия), наилучшим образом приближающей искомую кривую. Делается это с помощью метода наименьших квадратов, когда минимизируется сумма квадратов отклонений реально наблюдаемых Y от их оценок $\hat{Y}$ (имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость):

$$\sum_{k=1}^{M} (Y_k - \hat{Y_k})^2 \to min$$

(M — объём выборки). Этот подход основан на том известном факте, что фигурирующая в приведенном выражении сумма принимает минимальное значение именно для того случая, когда $Y = y(X_1, X_2 ... X_N)$.

Для решения задачи регрессионного анализа методом наименьших квадратов вводится понятие функции невязки:

$$\sigma(\bar{b}) = \frac{1}{2}\sum_{k=1}^{M}(Y_k - \hat{Y}_k)^2$$

Условие минимума функции невязки:

$$\begin{cases} \frac{d\sigma(\bar{b})}{db_i} = 0 \\ i = 0...N \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{M} My_i = \sum_{i=1}^{M}\sum_{j=1}^{N} b_j x_{i,j} + b_0 M \\ \sum_{i=1}^{M} y_i x_{i,k} = \sum_{i=1}^{M}\sum_{j=1}^{N} b_j x_{i,j} x_{i,k} + Mb_0 \sum_{i=1}^{M} x_{i,k} \\ k = 1...N \end{cases}$$

Полученная система является системой N+1 линейных уравнений с N+1 неизвестными $b_0, b_1 ... b_N$, которая решается методом Гаусса.

3.5. Решающее дерево

Пусть задано обучающее множество O, содержащее объекты (примеры), каждый из которых характеризуется *M* атрибутами (признаками), причем один из них указывает на принадлежность объекта к определенному классу.

Пусть через $C_1, C_2, \ldots C_k$ обозначены классы (значения метки класса), тогда существуют 3 ситуации:

- множество *O* содержит один или более примеров, относящихся к одному классу C_k. Тогда дерево решений для *O* – это лист, определяющий класс C_k;
- множество *O* не содержит ни одного примера, т.е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества отличного от *O*, скажем, из множества, ассоциированного с родителем;
- множество *O* содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество *O* на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значений $X_1, X_2 ... X_M$. *O* разбивается на подмножества $O_1, O_2 ... O_t$, где каждое подмножество O_i содержит все

примеры, имеющие значение X_i для выбранного признака. Это процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Вышеописанная процедура лежит в основе многих современных алгоритмов построения деревьев решений, этот метод известен еще под названием разделения и захвата (divide and conquer). Очевидно, что при использовании данной методики, построение дерева решений будет происходить сверху вниз.

Поскольку все объекты были заранее отнесены к известным нам классам, такой процесс построения дерева решений называется обучением с учителем (supervised learning). Процесс обучения также называют индуктивным обучением или индукцией деревьев (tree induction).

3.5. Среда разработки

Matlab – это высокопроизводительный язык для технических расчетов. Он включает в себя вычисления, визуализацию и программирование в удобной среде, где задачи и решения выражаются в форме, близкой к математической. Типичное использование Matlab – это:

Математические вычисления

Создание алгоритмов

Моделирование

Анализ данных, исследования и визуализация

Научная и инженерная графика

Разработка приложений, включая создание графического интерфейса

Matlab – это интерактивная система, в которой основным элементом данных является двухмерный массив, или матрица. Matlab развивался в течении нескольких лет, ориентируясь на различных пользователей. В университетской среде, он представлял собой стандартный инструмент для работы в различных областях математики, машиностроения и науки.

В Matlab важная роль отводится специализированным группам программ, называемых Toolboxes. Они очень важны для большинства пользователей Matlab, так как позволяют изучать и применять специализированные методы. Toolboxes – это всесторонняя коллекция функций Matlab (*.m*-файлов), которые позволяют решать частные классы задач. Toolboxes применяются для обработки сигналов, систем контроля, нейронных систем, нечеткой логики, моделирования и т.д.

Тексты программных модулей на языке Matlab сохраняются в файлах с расширением *.m*. Они представляют собой текстовые файлы, содержащие внешние определения команд и функций системы. Модули реализуются в виде процедур и функций.

Предложенный в работе алгоритм построения нечётких деревьев решений – нелинейных классификаторов реализован в среде Matlab 7.0. Программная реализация представлена в виде вызывающих друг друга процедур со вложенными функциями. Написано порядка 10 модулей-процедур, которые изменялись по мере совершенствования алгоритма.

При написании программ активно использовались дополнительные пакеты программ Matlab Fuzzy Logic Toolbox и Statistics Toolbox, предлагающий широкий выбор инструментов для статистического исследования и работы с нечёткой логикой.

Структура данных, их форматы, подробное описание работы модулей, а также сам текст программ приведены в Приложении к текущей работе.

Выводы

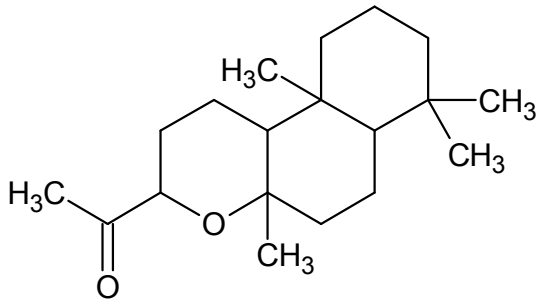
Предложены конкретные алгоритмы для реализации метода построения быстрых правил отказа от прогноза и нечёткой классифицирующей функции, описанного в главе 2. Приведена общая схема проводимых вычислений, по которой были получены результаты главы 4. Описаны алгоритмы эволюционного отбора дескрипторов, построения локальной линейной модели и решающего дерева. Для реализации проекта была выбрана среда разработки MATLAB.

Глава 4. Результаты расчётов

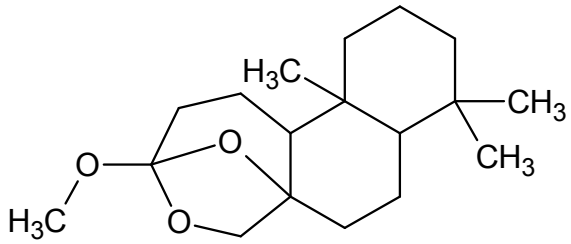
В данной главе приводятся результаты работы описанных в главе 2 методов и предложенных в главе 3 алгоритмов. Все расчёты проводились на реальных выборках химических соединений и имеют практическую значимость в соответствующих областях. Проведено необходимое сравнение предлагаемых методов с классическими. Даётся анализ и интерпретация результатов.

4.1. Выборка амбровых одорантов.

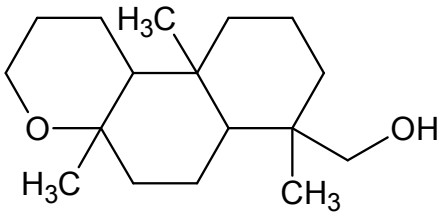
Для начала приведём результаты расчётов на выборке амбровых одорантов (низкомолекулярных соединений, обладающих амбровым запахом), состоящей из 129 молекул. Для каждого соединения выборки было дано 3D-описание соответствующего молекулярного графа в отдельном файле с расширением .mol. В данном файле перечислены 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.

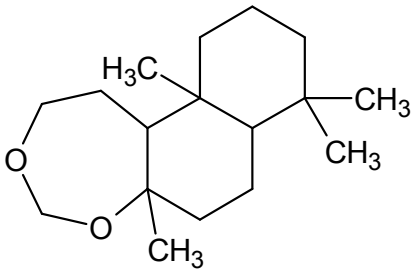


$C_{19}H_{32}O_2$



$C_{18}H_{30}O_3$





$C_{17}H_{30}O_2$

Рис 11. Структурные формулы некоторых из соединений выборки амбровых одорантов.

По выборке было сформировано 8 матриц. Параметры этапа описания наряду с показателями качества построенных классифицирующих функций приведены в таблице 1. Функционал качества вычислялся методом скользящего контроля. Для обоих методов проводился эволюционный отбор дескрипторов. Для универсальности при анализе каждой матрицы отбирались 30 столбцов-дескрипторов. В первом случае на них запускалась линейная регрессия. Во втором применялся разработанный в работе нечёткий классификатор. Кластеры формировались с помощью иерархической

кластеризации clusterdata с параметрами linkage = “average” и maxclust = 3. На её основе автоматически выбирались параметры нечёткой кластеризации и строились правила отказа от прогноза. Локальной прогнозирующей моделью для нечёткого классификатора являлась линейная регрессия.

	Регрессия	Нечёткий классификатор		параметры матрицы
№	прогноз	прогноз	отказ	
01	0,8605	0,8992	0	Fuzzy, Fuzzy_triangle, 2, linear, 2, linear
02	0,8915	0,9225	0	Fuzzy, Fuzzy_triangle, 2, linear, 3, linear
03	0,8992	0,9552	62	Fuzzy, Fuzzy_triangle, 3, linear, 2, linear
04	0,8450	0,8696	14	Fuzzy, Fuzzy_triangle, 3, linear, 3, linear
05	0,8140	0,8372	0	Fuzzy, Linear, 2, linear, 2, linear
06	0,8372	0,8947	72	Fuzzy, Linear, 2, linear, 3, linear
07	0,8837	0,9068	11	Fuzzy, Linear, 3, linear, 2, linear
08	0,8992	0,9421	8	Fuzzy, Linear, 3, linear, 3, linear

Таблица 1. Сравнительный анализ качества пргноза линейной регрессии и нечёткого классификатора на выборке амбровых одорантов.

Как видно из таблицы, предложенный метод значительно улучшает качество прогноза линейной регрессии. Более того, улучшение наблюдалось на всех предложенных матрицах «молекула - дескриптор».

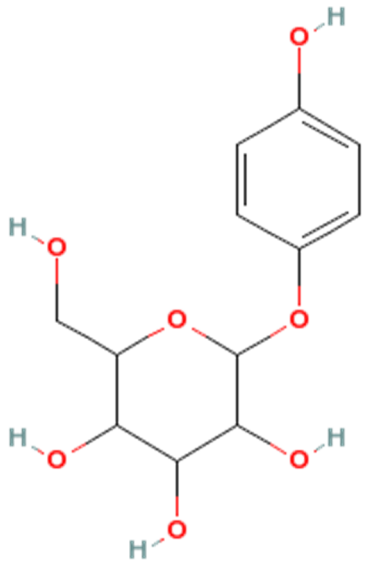
4.2. Выборка гликозидов

Выборка гликозидов предоставлена группой Банка данных отдела экспериментальной химиотерапии НИИ ЭДиТО ГУ РОНЦ им. Н.Н. Блохина РАМН (руководитель к.б.н. Апрышко Галина Николаевна). Выборка из 76 веществ класса гликозидов извлечена из Базы данных РОНЦ, в которой содержатся структурные формулы, номенклатурные характеристики, физико-химические свойства и результаты изучения цитотоксической активности in vitro и противоопухолевой активности in vivo около 12000 оригинальных отечественных синтетических веществ и природных экстрактов, которые изучались в РОНЦ или других учреждениях России и стран СНГ [17].

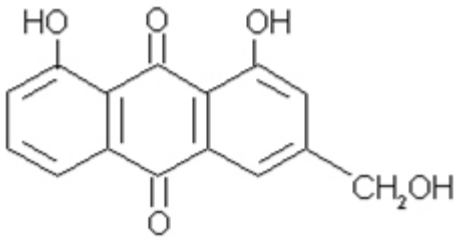
Имеются также количественные данные по результатам изучения общей токсичности веществ на лабораторных животных.

Особенностью информации по биологической активности, представленной в БД РОНЦ, является то, что результаты экспериментального изучения получены в стандартизованных экспериментальных условиях одного учреждения по одним и тем же методикам и имеют количественный характер.

Гликозиды представляют собой обширную группу органических веществ, встречающихся в растительном (реже в животном) мире и/или получаемых синтетическим путём. Ниже представлены некоторые структурные формулы



Арбутин



Алоэ – эмодин

Рис 12. Структурные формулы некоторых из соединений выборки гликозидов.

Изначально вся выборка была разбита на 5 классов, но это были не классы активности, а классы по схожести неуглеводородного фрагмента. Каждый из классов описывался 24 матрицами (то есть всего было 120 матриц), в зависимости от:

–способа разбиения интервала электростатического заряда (2 варианта);

– типу функции принадлежности – четкие, нечеткие треугольные, нечеткие трапециевидные (3 варианта);

– количества разбиений интервала расстояний между особыми точками (ОТ) и между ОТ парой ОТ (еще 4 варианта).

Ввиду того, что некоторые классы были совсем малочисленны, было предложено их объединить и рассматривать в совокупности. В результате чего была сформированы всего 24 матрицы, описывающие 76 молекул (57 неактивных и 19 активных). Мы приведём результаты работы метода на первых 10 матрицах. Нумерация молекул начинается с 9, так как номера с 1 по 8 остались в репозитории за выборкой амбровых одорантов.

Здесь, как и ранее использовалась иерархическая кластеризация clusterdata с параметрами linkage = “average” и maxclust = 3. На её основе автоматически выбирались параметры нечёткой кластеризации и строились правила отказа от прогноза. Методом МГУА было отобрано 15 «лучших» дескрипторов для построения прогнозирующей модели. Локальной прогнозирующей моделью являлась линейная регрессия. Таблица 2 показывает преимущества предложенного подхода для предсказания данной химической активности.

Номер матрицы	прогноз	отказ
09	0.9231	11
10	0.9275	7
11	0.9868	0
12	0.9737	0
13	0.9677	14
14	0.9710	7
15	0.9545	32
16	0.9868	0
17	1	1
18	1	5

Таблица 2. Результат работы нечёткого классификатора на выборке гликозидов.

4.3. Выборка токсичных соединений

Данная выборка была представлена на международном конкурсе по прогнозированию активности химических соединений Environmental Toxicity Prediction Challenge [18].

Химическая токсичность может вызывать множество биологически опасных эффектов, таких как повреждение генов, или даже привести смерти человека или животных. Около 120000 химических соединений будут протестированы в ближайшие 10 лет. Для этого потребуется до 45 миллионов лабораторных животных. Альтернативой может послужить прогноз токсичности на основе моделей построенных на уже протестированных соединениях. Исследуемая нами выборка поделена на две части. Первая из 644 молекул предназначена для построения моделей, а вторая (449 соединений) для их проверки. Таким образом, функционал качества вычислялся в данном случае на проверочном множестве. На конкурсе было предложено 5 различных признаковых пространств.

Сначала приведём сравнительные результаты работы регрессии без эволюционного отбора дескрипторов с МГУА-регрессией и нечётким классификатором.

Вид дескрипторов	Регрессия	МГУА + регрессия	Нечёткий классификатор
E-state (60)	R2 = 0.1451;	R2 = 0,7017;	R2 = 0.7334; reject = 30;
DRAGON (1664)	R2 = 0.0671;	R2 = 0.8276;	R2 = 0.8289; reject = 1;
SimulationsPlus (221)	R2 = 0.0979;	R2 = 0.8004;	R2 = 0.7685; reject = 7;
QuantumChemistry (23)	R2 = 0.7793;	R2 = 0,7919;	R2 = 0.7849; reject = 1;
MOE (254)	R2 = 0.0253;	R2 = 0.8470;	R2 = 0.7564; reject = 15;

Таблица 3. Сравнение различных подходов на базе линейной регрессии на различных пространствах признаков выборки ETPC.

В поле вид дескрипторов в скобках указана мощность данного пространства дескрипторов, reject – означает здесь и далее число молекул, которым отказано в прогнозе. Из таблицы видно, что регрессия без отбора дескрипторов не может быть использована для построения моделей, когда число дескрипторов велико. Также отметим, что не всегда нечёткий классификатор однозначно лучше регрессии с эволюционным отбором дескрипторов. Для сравнения приведём результаты работы и других алгоритмов на пространстве дескрипторов QuantumChemistry.

Метод, которым получен прогноз	Качество прогноза
Линейная регрессия	R2 = 0.7793;
Решающее дерево	R2 = 0.6346;
МГУА + регрессия	R2 = 0,7919;
Нечёткий классификатор	R2 = 0.7849; reject = 1;
Реграссия + k-means (k=3)	R2 = 0.7625; reject = 4
Регрессия + иерарх. кластеризация (3 кластера)	R2 = 0.7664; reject = 9;
Дерево + k-means (k=3)	R2 = 0.6942; reject = 235;
Дерево + иерарх. кластеризация (3 кластера)	R2 = 0,6261; reject = 9;
Регрессия + fuzzy-кластеризация	R2 = 0,7940; reject = 1;
Дерево + fuzzy-кластеризация	R2 = 0.6688; reject = 1;

Таблица 4. Другие методы построения прогнозирующей функции.

Отметим, что здесь лучший результат получен с использование нечёткой классификации, но на всей матрице (без эволюционного отбора дескрипторов). Если число дескрипторов ограничить (в данном случае 15-ю) качество прогноза снижается. Это означает в частности, что все дескрипторы содержательны для предсказания активности данных соединений.

4.4. Выборка TOMOL

Исходные данные: выборка химических соединений, предоставленная к.х.н. И.В. Свитанько. Рассматриваемая активность – ингибиторы фермента деления клеток. В выборке представлено 120 соединений с известной активностью – 86 активных и 34 неактивных. Также представлены 196 молекул с неизвестной активностью.

Задача: построить модели для прогнозирования активности и применить построенные модели для молекул с неизвестной активностью.

В качестве дескрипторов для описания соединений выбран метод выделения линейных фрагментов с введением маркировки вершин молекулярных графов.

Регулируемые *параметры описания*:

1) длина линейных фрагментов (k= 2, 3, 4);

2) маркеры, участвующие в описании (d, b, r).

В соответствии с выбором параметров построено 24 матрицы молекула-дескриптор (МД-матриц) – 8 вариантов включения маркеров и 3 варианта длины линейных фрагментов. Строкам матрицы соответствуют молекулы выборки, столбцам – дескрипторы. Значение матрицы на пересечении *i*-ой строки и *j*-го столбца – количество повторений *j*-го дескриптора в *i*-ой молекуле.

Количество различных дескрипторов, полученных в результате построения моделей, приведены в таблице.

Маркер	***	**r	*b*	d**	*br	d*r	db*	dbr
Длина фрагмента	k=2							
Количество дескрипторов	15	42	34	43	83	80	78	121
Длина фрагмента	k=3							
Количество дескрипторов	30	114	84	113	213	216	199	296
Длина фрагмента	k=4							
Количество дескрипторов	47	237	158	226	405	432	376	562

Таблица 5. Количество дескрипторов первого уровня

МД-матрицы, полученные с помощью различных наборов дескрипторов, обрабатывались с помощью подхода, сочетающего использование для прогноза активности деревьев решений, строящихся локально внутри нечетких кластеров, с эволюционным отбором дескрипторов.

Отметим, что в настоящем исследовании число кластеров было выбрано эмпирически, зафиксировано и равнялось 3.

Для оценки качества моделей, полученных в ходе описанной выше процедуры, использовался функционал качества со скользящим контролем.

Ниже в таблице представлены модели, полученные с помощью нечетких деревьев решений при эволюционном отборе дескрипторов.

Маркер	***	**r	*b*	d**	*br	d*r	db*	dbr
Длина фрагмента	k=2							
Качество прогноза	0.85833	0.89167	0.91667	0.875	0.9750	0.85833	0.94167	0.975
Длина фрагмента	k=3							
Качество прогноза	0.9	0.90833	0.875	0.91667	0.95833	0.9333	0.95	0.94167
Длина фрагмента	k=4							
Качество прогноза	0.825	0.875	0.875	0.825	0.94167	0.9	0.9333	0.9667

Таблица 6. Качество прогноза моделей, построенных с помощью нечетких деревьев решений по дескрипторам первого уровня

В результате работы было получено несколько моделей прогнозирования активности с хорошим качеством прогноза, с помощью построенных моделей рассчитана активность новых соединений.

Выводы

Приведённые результаты показывают практическую значимость предложенного в работе подхода. Новые методы позволили получить прогнозирующие модели высокого качества. В большинстве случаев наблюдаем значительное улучшение качества прогноза по сравнению с классическими методами. Проведено сравнение различных реализаций подхода и отобраны значимые для прогноза дескрипторы.

Заключение

В работе предложен метод построения классифицирующей функции в задаче обнаружения функциональной зависимости между структурой молекулярного графа и биологическими свойствами веществ «структура-свойство». Данный метод основан на нечеткой кластерной структуре обучающей выборки и позволяет избавиться от недостатков существующих методов: прогнозирование соединений, заведомо не относящихся к изучаемому классу веществ и невозможность оптимизировать кластерную структуру обучающей выборки под исследуемое свойство. Разработанный метод построения классифицирующей функции прогнозирует значение активности нового соединения исходя из активности наиболее схожих с ним соединений из обучающего множества, а также удовлетворяет требованию устойчивости к наличию в обучающей выборке выбросов, и не учитывает их при построении прогнозирующих моделей. Помимо этого в работе предложены быстрые с вычислительной точки зрения правила отказа от прогноза, которые осуществляют отказ в случае, когда предсказание свойства нового соединения нецелесообразно.

Метод состоит в нахождении классифицирующей функции на базе локальных прогнозирующих моделей, построенных внутри кластеров, используя параметризованную нечёткую кластерную структуру обучающего множества для взвешенного прогноза нового соединения. Показано, что

нечеткое описание кластерной структуры обучающей выборки может быть оптимизировано под целевое свойство.

При работе над данным материалом был разработан и программно реализован эволюционный алгоритм, реализующий построение нечёткой классифицирующей функции на наиболее информативных дескрипторах. Было проведено тестирование алгоритма, проверенна его эффективность на реальных выборках химических соединений, замечено улучшение качества прогноза классических методов при переходе к кластерам, а затем и нечётким кластерам. В ходе вычислительных экспериментов отмечена также пригодность полученных моделей для прогнозирования активности новых соединений с различными целевыми свойствами/активностями.

Одной из перспектив развития предложенного метода поиска функциональной зависимости в задаче «структура - свойство» является более детальное тестирование подхода, в частности:

1) проведение расчетов на большем числе выборок из разных областей химии и медицины;
2) реализация и использование при построении нечёткого классификатора других алгоритмов кластерного анализа;
3) построение локальных прогнозирующих моделей с помощью других алгоритмов распознавания образов;

Также интерес представляют другие параметризации нечёткой кластерной структуры обучающей выборке и оптимизация нечёткой классифицирующей функции по новым параметрам.

Продолжением работы также являются испытания быстрых правил отказа и нечёткой классифицирующей функции при скрининге больших баз соединений с неизвестной активностью.

Литература

1. Прохоров Е.И., Перевозников А.В., Воропаев И.Д., Кумсков М.И., Пономарёва Л.А. Поиск представления молекул и методы прогнозирования активности в задаче «структура–свойство» // Доклады 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. – М: МАКС Пресс. – 2009. – С. 589-591.

2. Деветьяров Д.А., Кумсков М. И., Апрышко Г. Н., Носеевич Ф.М. и др. Сравнительный анализ применения нечетких дескрипторов при решении задачи «структура–активность» для выборки гликозидов // 14-я всеросс. конф. ММРО-14. - М.: МАКСПресс, 2009. – С. 575-578.

3. 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.

4. Devetyarov D. A., Zaharov A. M., Kumskov M. I., Ponomareva L. A. Fuzzy logic application for construction of 3D descriptors of molecules in QSAR problem // 8th Intern. Conf. «Pattern Recognition and Image Analysis: New Information Technologies». - 2007._Vol. 2._Pp. 249–252.

5. Прохоров Е.И., Перевозников А.В., Пономарева Л.А. Кумсков М.И. Нейронная сеть как инструмент реализации кусочно-линейного классификатора при массовом скрининге молекул в задаче «структура-свойство» // Нейрокомпьютеры: разработка, применение. – № 3. – С. 39-45.

6. Штовба С.Д. Введение в теорию нечетких множеств и нечеткую логику. Винница: Издательство винницкого государственного технического университета, 2001. - 198 с.

7. В.А. Кохов. Метод количественного определения сходства графов на основе структурных спектров // Известия РАН, «Техническая Кибернетика ». – 1994. - № 5. - С. 143-159.

8. Деветьяров Д.А., Григорьева С.С., Пермяков Е.А., Кумсков М.И., Понаморёва Л.А., Свитанко И.В. Решение задачи «структура – свойство» для молекул с множеством пространственных конформаций. // Система прогнозирования свойств химических соединений: Алгоритмы и модели: Сборник научных работ/ Под ред. М.И. Кумскова. Москва: МАКС Пресс, 2008.

9. Григорьева С.С., Кумсков М.И., Захаров А.М. Применение метода главных компонент при построении кластерной структуры обучающей выборки молекул. // Математические методы распознавания образов. 13-я Всероссийская конференция: Сборник докладов. Москва: МАКС Пресс, 2007.

10. Деветьяров Д.А., Кумсков М.И., Апрышко Г.Н., Носеевич Ф.М., Прохоров Е.И., Перевозников А.В., Пермяков Е.А. Сравнительный анализ применения нечетких дескрипторов при решении задачи «структура-свойство» // Доклады 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. – М: МАКС Пресс. – 2009. – С. 511-514.

11. Кумсков М.И., Митюшев Д.Ф. Применение метода группового учета аргументов для построения коллективных оценок свойств органических соединений на основе индуктивного перебора их “структурных спектров”. / Проблемы управления и информатики, 1996, №4, с.127-149.

12. Харман Г. Современный факторный анализ: Пер. с англ., - М.: Статистика, 1972, 486с.

13. P. Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, 2002.

14. Kim K. H., Greco G., Novellino E. A Critical Review of Recent CoMFA Applications. 3D QSAR in Drug Design, Kubinyi, H.; Folkers, G.; Martin, Y. C. (Eds), Kluwer Academic Publishers, Great Britain vol. 3, p. 257., 1998.

15. Geladi P., Kowalski B.R. Partial least squares: a tutorial. Anal. Chim. Acta. vol. 185, pp. 1-15, 1986.

16. Yang, T. "Computational Verb Decision Trees" // International Journal of Computational Cognition (Yang's Scientific Press) – 2006 - 4 (4): 34-46

17. Апрышко Г.Н. Информационная система РОНЦ им. Н.Н. Блохина РАМН по противоопухолевым агентам. Общий обзор // НТИ. Сер. 2. – 2007. – № 1. – С. 18-22.

18. http://www.cadaster.eu/node/65

19. Randic M. On Characterization of Molecular Branching. Journal of the American Chemical Society, 1975, vo.97, pp.6609-6615.

20. Kier L.B., Hall L.H. Molecular Connectivity in Chemistry and Drug Research; Academic Press: New York, 1976.

21. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Топологические индексы в органической химии. Успехи химии. 1988. Т.57. №3. С.337-366.

22. Moriguchi I., Kanada Y. Use of van der Waals volume in structure-activity studies. Chem. Pharm. Bull. (Tokyo), vol.25, pp.925-936, 1977.

23. Labute P. A widely applicable set of molecular descriptors. J. Mol. Graph. Mod., vol.18, pp. 464-477, 2000.

24. Wiener H. Structural determination of paraffin boiling points. J Am Chem Soc 1947, vol.69, pp.17–20

25. Lowis D. R. HQSAR. A New, Highly Predictive QSAR Technique. Tripos Technical Notes; Oct. 1997; Vol. 1, No.

26. Rouvray D.H. (Ed.) Computational Chemical Graph Theory. / Nova Publ., New York, 1989.

27. Balaban A. T., Applications of Graph Theory in Chemistry, J. Chem. Inf. Comput. Sci., 25, p.334, 1985.

28. Розенблит А. Б., Голендер В. Е. Логико-комбинаторные методы в конструировании лекарств.— Рига: Зинатне, 1984.— 352 с.

29. Кумсков М.И., Смоленский Е.А., Пономарева Л.А., Митюшев Д.Ф., Зефиров Н.С. Системы структурных дескрипторов для решения задач «структура-активность». Доклады Академии Наук, 1994, 336, п.1., с.64-66.

30. Carhart R et al. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. J. Chem. Inf. Comput. Sci.; 1985; 25(2) pp 64–73

31. 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, n.3, pp.477-484.

32. 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. p.795-808.

33. Cramer R.D., III, D.E. Patterson, and J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc., 1988. 110(18): pp. 5959-5967.

34. Y. Martin, Distance Comparisons: A New Strategy for Examining Three-Dimensional Structure-Activity Relationships," in Classical and Three-Dimensional QSAR in Agrochemistry. ACS Symposium Series, C. Hansch & T. Fijita, eds., pp 318-329, (1995).

35. Y.C. Martin, K.-H. Kim, C.T. Liu, "Comparative Molecular Field Analysis: CoMFA" in Advances in Quant. Str.-Property Relationships, 1, 1-52. JAI Press (1996).

36. U. Norinder, "Single and Domain Mode Variable Selection in 3D QSAR Applications", J. Chemometrics, 10, 95-105 (1996).

37. Харман Г. Современный факторный анализ: Пер. с англ., - М.: Статистика, 1972, 486с.

38. Шараф М.А., Иламен Д.А., Ковальский Б.Р. Хемометрика: Пер. с англ. Ленинград: Химия, 1989, 269 с.

39. Klebe G., Abraham U. and Mietzner T. J. Med. Chem., vol. 37, pp. 4130–4146., 1994.

40. Klebe, G. Comparative Molecular Similarity Indices: CoMSIA, in "3D QSAR in Drug Design" H. Kubinyi et al. (eds), Kluwer Academic Publishers, Great Britain, 1998, Volume 3, pages 87-104.

41. Kim K. H., Greco G., Novellino E. A Critical Review of Recent CoMFA Applications. 3D QSAR in Drug Design, Kubinyi, H.; Folkers, G.; Martin, Y. C. (Eds), Kluwer Academic Publishers, Great Britain vol. 3, p. 257., 1998.

42. Pastor M., Cruciani G., McLay I., Pickett S., Clementi S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem, Vol. 43, No. 17., pp. 3233-3243, 2000.

43. Ortiz A. R., Pisabarro M. T., Gago F., Wade R. C. Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem vol. 38, pp.2681-2691, 1995.

44. 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.