Московский государственный университет
имени М.В. Ломоносова
Механико-математический факультет

На правах рукописи

Деветьяров Дмитрий Александрович

ИСПОЛЬЗОВАНИЕ НЕЧЕТКОЙ ЛОГИКИ ПРИ ОПИСАНИИ МОЛЕКУЛ В ЗАДАЧЕ «СТРУКТУРА-СВОЙСТВО»

05.13.17 – теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва - 2010

Работа выполнена на кафедре вычислительной математики механико-математического факультета Московского государственного университета имени М.В.Ломоносова.

Научный руководитель: доктор физико-математических наук,
Кумсков Михаил Иванович

Официальные оппоненты: доктор физико-математических наук,
профессор *Персианцев Игорь Георгиевич*

кандидат физико-математических наук,
старший научный сотрудник,
Афонин Сергей Александрович

Ведущая организация: Вычислительный центр
имени А. А. Дородницына РАН

Защита состоится 28 апреля 2010 г. в 16 час. 45 мин. на заседании диссертационного совета Д.501.002.16 в Московском государственном университете имени М.В. Ломоносова по адресу: Российская Федерация, 119991, Москва, ГСП-1, Ленинские горы, д.1, Московский государственный университет имени М.В. Ломоносова, механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в библиотеке механико-математического факультета МГУ (Главное здание, 14 этаж).

Автореферат разослан 26 марта 2010 г.

Ученый секретарь
диссертационного совета Д.501.002.16 при МГУ
доктор физико-математических наук А.А. Корнев

Общая характеристика работы

Актуальность

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

Особенно широкое развитие методы QSAR получили в последние 10-15 лет в связи с тем, что появились возможности для компьютерного хранения больших объемов данных о структуре всевозможных молекул и их активности, а также в связи с тем, что сильно повысилась производительность вычислительных систем, являющаяся критичной для ряда методов решения задачи QSAR.

В настоящее время разработано несколько разных подходов к решению QSAR-задачи. Как правило, QSAR-задача разбивается на две подзадачи:

1) преобразование информации о молекулярной структуре в вектора численных признаков (дескрипторов);

2) анализ полученных данных (построение предсказывающей модели для биологической активности – функции в векторном пространстве признаков). Предсказывающая модель строится с использованием стандартных методов машинного обучения (линейные и нелинейные регрессии, нейронные сети и т.д.).

За последние несколько десятилетий разработано большое число методов решения QSAR-задачи, при этом методы различаются, главным образом, методом описания молекул в векторном пространстве признаками (дескрипторами). Классический подход был предложен Розенблитом и Голендером[2], которые использовали понятие «фармакофор» — набор структурных признаков в молекуле, которые отвечают за биологическую активность молекулы. Данный метод выделяет группы или цепочки атомов в структуре молекулы и находит функциональную зависимость между наличием тех или иных групп или цепочек и биологической активностью.

---

[1] Karelson M. Molecular Descriptors in QSAR/QSPR. Wiley-interscience, 2000

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

Разработанный в работе метод развивает данное направление, однако направлен на избавление от ряда недостатков, которыми обладают классические структурные дескрипторы:

1. *Проблема автоматического поиска оптимального описания молекул*. При описании дескрипторами параметры описания, как правило, выбираются оператором исходя из априорной информации об обучающем множестве или из других соображений. В частности, описание молекулы структурными дескрипторами существенно зависит от выбора параметров описания – интервалов расстояний. При этом, затруднена возможная оптимизация выбора такого разбиения, так как значения дескрипторов не связаны непрерывно с выбором параметров – точек разбиений. Данная проблема называется *проблемой дискретизации расстояний*. Необходимость вмешательства оператора в описание молекул снижает прогностичную силу и скорость работы моделей «структура-свойство». Таким образом, является актуальной задача автоматического поиска оптимального описания молекул.

2. *Невозможность учитывать подвижность пространственной структуры молекулы*. При моделировании биологической активности задача «структура-свойство» осложняется тем, что молекулы могут незначительно менять конформацию (пространственную укладку). В результате, при изменении конформации даже несущественное изменение взаимного расположения атомов может привести к значительному изменению значений дескрипторов и прогнозирующая функция может работать ошибочно. Следовательно, актуальной является разработка методов представления информации о структуре молекул, нечувствительных к небольшим сдвигам атомов относительно положения равновесия.

Таким образом, актуальной является разработка нового метода представления информации о структуре молекулы, который не обладает вышеописанными недостатками. Данный метод предлагается разработать с помощью использования аппарата нечеткой логики при определении так называемых «нечетких» дескрипторов.

Кроме того, сформулированы следующие требования к разработанному методу:

1. Метод должен позволять содержательную интерпретацию дескрипторов, используемых в моделях, отражающих функциональную зависимость между структурой и свойством. Некоторые современные методы (например, топологические индексы) не обладают данным свойством.

2. Помимо нахождения структурных признаков, отвечающих за биологическую активность, метод должен также осуществлять проверку гипотезы о локальной значимости того или иного физико-химического свойства (например, электростатического заряда, липофильности, способности принимать/отдавать электрон).

Цель работы

Разработка метода представления информации о пространственных структурах молекул, основанного на нечетких структурных дескрипторах, в задаче обнаружения функциональной зависимости «структура-свойство». Для достижения этой цели сформулированы и решаются следующие **задачи**:

1. Разработать метод представления информации о пространственных конфигурациях молекул с помощью нечетких структурных 3D-дескрипторов.
2. Разработать алгоритм формирования алфавита нечетких структурных 3D-дескрипторов.
3. Разработать алгоритм оптимизации нечеткого описания молекул с целью поиска локально лучшей модели в некотором классе предсказывающих функций.
4. Оценить вычислительную сложность разработанных алгоритмов.
5. Реализовать разработанные алгоритмы, провести вычислительные эксперименты.

Научная новизна

1. Предложен новый метод представления информации о структуре молекулярных графов семействами четких и нечетких структурных 3D-дескрипторов.
2. В рамках предложенного метода разработан алгоритм описания молекул в задаче «структура-свойство» и проведена оценка вычислительной сложности алгоритма.
3. Подтверждена практическая значимость подхода в серии вычислительных экспериментов по прогнозированию биологической активности органических соединений.

Обоснованность и достоверность научных положений и полученных результатов обеспечивается обоснованной с точки зрения химии и биологии постановкой задачи и результатами тестирования использованных методов.

Практическая значимость

Разработанные алгоритмы решения QSAR-задачи могут быть использованы для решения прикладных задач предсказания физико-химической или биологической активности веществ по их структуре. Это позволяет отказаться от дорогостоящих и длительных исследований внеэкспериментальным скринингом на больших наборах химических соединений. Архитектура программного комплекса, созданного в рамках выполнения диссертационной работы, может служить основой для

автоматической системы предсказания активности соединений. Предложенный эволюционный алгоритм построения дескрипторов может быть использован для повышения вычислительной эффективности подобной системы.

Апробация работы

Материалы диссертации докладывались и обсуждались на 8-ой международной конференции «Распознавание образов и анализ изображений: новые информационные технологии» ("Pattern Recognition and Image Analysis: New Information Technologies", PRIA-8-2007), Международной научной конференции «Компьютерные науки и информационные технологии» (2009 г.), 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009 (2009 г.), Молодежной конференции «Молекулярный дизайн и синтез веществ с заданной физиологической активностью» (химический факультет МГУ им. М,В. Ломоносова, 2006 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.

Публикации по теме диссертации

По материалам диссертации опубликовано 12 научных работ [1-12]. Из них – три работы [10, 11, 12] представлены в журналах из перечня ведущих научных журналов и изданий, рекомендованных ВАК РФ.

Структура и объем диссертации

Работа состоит из введения, 3 глав, заключения, списка литературы и приложения. Общий объем диссертации – 123 страницы. Список литературы содержит 79 наименований.

Краткое содержание работы

Во **введении** дано описание основных результатов, приведены научная новизна и практическая значимость диссертации.

Первая глава является вводной и представляет собой обзор существующих методов решения задачи поиска функциональной зависимости «структура-свойство».

В **разделе 1.1** приведена общая постановка задачи «структура-свойство» для молекул, указаны основные подходы к описанию обучающего множества

молекул – молекулярными графами[3], молекулярными поверхностями[4], наборами особых точек.

Определение (задача «структура-свойство»). Пусть задано обучающее множество молекул $LS = (M_i, y_i),\ i = 1,...,N$, в котором каждая молекула M_i задана одним из описанных выше способов и отнесена к некоторому классу активности $A_k, k = 1,\ldots,K$, который можно описать меткой $y_i, i = 1,\ldots,N$, или имеет некоторое значение свойства $y_i \in \square, i = 1,\ldots,N$. Пусть также задан F – некоторый класс функций $f : \square^M \to \{A_1,\ldots,A_K\}$ при рассмотрении классов или $f : \square^M \to \square$ при рассмотрении свойства. Необходимо:

1. (Этап описания) Построить вектор признаков-дескрипторов для каждой молекулы $M_i \to (x_{i1},...,x_{iM})$, где $x_{i1},...,x_{iM}$ – значения соответствующих дескрипторов.

2. (Этап анализа) Выбрать функцию $f \in F$, получающую в качестве аргумента вектор дескрипторов молекулы и наилучшим образом относящую молекулу к одному из K классов активности или предсказывают значение свойства $y_i \in \square, i = 1,\ldots,N$ в смысле некоторого функционала качества $\varphi(f)$.

В **разделе 1.2** проводится обзор существующих методов представления информации о молекуле на этапе описания задачи «структура-свойство», а также перечисляются их недостатки. Рассмотрены методы на основе описания молекул топологическими дескрипторами (теоретико-графовыми индексами)[5], а также структурными дескрипторами[6], характеризующими наличие, количество, и взаимное расположение в молекуле определенных структурных фрагментов (атомов, связей и т.д.). Описаны методы решения задачи 3D-QSAR[7] путем вложения молекул в трехмерное пространство с регулярной сеткой.

В **разделе 1.3** приведены основные классы F классифицирующих функций f, используемых в задаче «структура-свойство» на этапе анализа, и методы машинного обучения, которые осуществляют поиск оптимальной классифицирующей функции. Обосновывается предпочтение классу линейных функций и эволюционным алгоритмам, строящим линейные модели (например, МГУА[8]).

---

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

[4] Lee, B., Richards F.M. The interpretation of protein structures: Estimation of static accessibility. Journal of Molecular Biology, vol. 55, 1971, pp.379-400

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

[6] 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

[7] Lowis D. R. HQSAR. A New, Highly Predictive QSAR Technique. Tripos Technical Notes; Oct. 1997; Vol. 1, No. 5

[8] Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. Принятие решений на основе самоорганизации. М.: Сов. Радио, 1976

Во **второй главе** диссертации приведена постановка задачи данной работы, описаны разработанные методы и алгоритмы решения поставленной задачи

Постановка задачи изложена в **разделе 2.1**. Исходя из недостатков методов представления информации о структуре молекул в задаче «структура-свойство», приведенных в разделе 1.2, перечислены актуальные проблемы разработки подобных методов: невозможность учитывать гибкость трехмерной структуры молекулы, а также необходимость участия оператора в выборе описания, что приводит к сложности оптимизации описания.

В результате, целью работы является разработка метода описания молекул для решения задачи «структура-свойство», который позволяет избавиться от вышеописанных недостатков. За основу метода предложено взять модель «ключ-замок» о наличии активного центра, представляющий собой комбинацию структурных фрагментов – особых точек. При использовании данной модели представляется естественным взять за основу разрабатываемых дескрипторов структурные дескрипторы, описанные в разделе 1.2.2. Наконец, вышеописанные проблемы существующих методов предлагается решать с помощью применения аппарата нечеткой логики.

Также в разделе выдвигаются дополнительные требования к разрабатываемому методу решения задачи, среди которых требование о содержательной химико-биологической интерпретации дескрипторов и о проверке гипотезы о локальной значимости физико-химического свойства.

Далее, приведены разработанные методы и алгоритмы описания молекулярных поверхностей.

Пусть задано обучающее множество вида (M_i, y_i), $i = 1,...,N$, где каждая молекула представлена набором выбранных по предварительно заданному алгоритму особых точек $M_i = \{P_i^j\}_{j=1}^{n_i}$, для каждой точки P_i^j заданы ее координаты $(x_i^j . y_i^j, z_i^j)$ и вектор физико-химических свойств $(p_i^{j,1},..., p_i^{j,L}) \in R^L$. Положим $p_{\min}^l = \min\limits_{i,j} p_i^{j,l}$, $p_{\max}^l = \max\limits_{i,j} p_i^{j,l}$, $d_{\max}$ – максимум по всей выборке всех возможных евклидовых расстояний между особыми точками одной молекулы.

В **разделе 2.2** приведен разработанный метод описания молекулярной структуры четкими структурными 3D-дескрипторами, построенных без использования преимуществ нечеткой логики.

В случае четких дескрипторов для каждого l, $1 \le l \le L$, отрезок $[p_{\min}^l, p_{\max}^l]$ разбивается на n_l подотрезков – классов значений свойств. В зависимости от принадлежности значения каждого свойства классам значений, каждой особой точке P_i^j присваивается метка "$i_1 \ldots i_L$", такая что значение свойства $p_i^{j,l}$ принадлежит отрезку значений i_l. Символьное кодирование меток порождает алфавит дескрипторов первого уровня $AD^1 = \{A_1, A_2, ..., A_l\}$, $A_1 < A_2 < ... < A_l$ – множество всех полученных символьных меток. Положим, что дескриптору A_i

соответствуют те и только химические функциональные группы G, которые состоят ровно из одной особой точки и $G = \{A_i\}$.

На отрезке [0, d_{max}] вводятся P интервалов расстояний. По индукции по уровню дескрипторов формируются алфавиты дескрипторов высших уровней AD^2, AD^3, … (для пар, троек особых точек и т.д.). Пусть уже построены алфавиты AD^2, AD^3, …, AD^n и необходимо построить алфавит дескрипторов $(n+1)$-ого уровня AD^{n+1} и задать соответствие между сформированными дескрипторами и химическими функциональными группами. К каждому из дескрипторов в AD^n добавляется новая особая точка $A, A \in AD$, которая лексикографически не меньше, чем метка любой особой точки дескриптора из AD^n; алфавит дескрипторов следующего уровня определяется как $AD^{n+1} = \{(D, A, c) \mid D \in AD^n, \quad A \in AD, \quad A \geq B \quad \forall B \in D, \quad c = 1,..,P)$.

Теперь, для того чтобы определить соответствие между химической функциональной группой G и произвольным дескриптором $D = (\hat{D}, A, c) \in AD^{n+1}$, необходимо проверить, можно ли разбить G на 2 такие группы G_1 и G_2 (состоящие из n и 1 особых точек соответственно), что фрагменту G_1 соответствует дескриптор $\hat{D}$ и $G_2 = \{A\}$. Если такое разбиение возможно, вычисляется расстояние $\rho(A, G_1)$ между G_1 и $G_2 = \{A\}$ (в качестве расстояния рассматривается наименьшее, наибольшее или среднее из всех расстояний между A и каждой из особых точек G). Химическая функциональная группа G соответствует дескриптору D тогда и только тогда, когда расстояние $\rho(A, G_1)$ принадлежит интервалу разбиения c.

Наконец, для каждой молекулы и каждого структурного дескриптора перечисляются все химические функциональные группы молекулярного графа, состоящие из n особых точек, соответствующих данному дескриптору, и значение дескриптора для данной молекулы определяется равным количеству подобных фрагментов.

Предложенный метод ориентирован на подтверждение гипотезы о биологической модели «ключ-замок», осуществляет проверку гипотезы о локальной значимости того или иного физико-химического свойства и обеспечивает содержательную интерпретацию полученной модели. Однако метод не решает проблемы дискретизации расстояний (автоматической оптимизации описания) и некорректной обработки гибких молекул.

Вышеизложенный метод модифицирован в **разделе 2.3** с помощью аппарата нечеткой логики в метод описания нечеткими структурными 3D-дескрипторами.

В **подразделе 2.3.1** даны общие понятия аппарата нечеткой логики[9]: даны определения нечеткого множества, функций принадлежности, операций над нечеткими множествами и систем логического вывода. Кратко описаны

[9] Zadeh L.A. Fuzzy sets. Information and Control, 1965, pp. 338-353

существующие методы решения задачи «структура-свойство» с использованием методов нечеткой логики. Рассмотрены подходы к описанию молекул структурными дескрипторами с введением нечетких множеств и функций принадлежности на множестве расстояний; подходы, основанные на использовании систем нечеткого логического вывода Мамдани[10] и Такаги-Сугено[11]. Указаны недостатки таких подходов, в частности, невозможность сформулировать до решения задачи «структура-свойство» правила логического вывода экспертно и ограниченная применимость нечеткого логического вывода в силу большого числа дескрипторов.

В **подразделе 2.3.2** вводятся понятия нечетких классов особых точек и расстояний.

Для каждого свойства l, $1 \le l \le L$, необходимо выбрать некоторое число n_l нечетких классов и n_l функций принадлежности $\mu_1^l,...,\mu_{n_l}^l,\mu_i^l:[p_{\min}^l,p_{\max}^l]\to[0,1]$, задающие нечеткие множества $A_i^l\, 1\le i\le n_l, 1\le l\le L$. В результате, для каждой особой точки P с вектором свойств $(p^1,...,p^L)\in R^L$ в L-мерном параллелепипеде $[p_{\min}^1,p_{\max}^1]\times...\times[p_{\min}^L,p_{\max}^L]$ можно вычислить $\sum_{l=1}^{L} n_l$ чисел, характеризующих принадлежность точки к различным нечетким классам особых точек по каждому из физико-химических свойств.

Аналогично вводятся нечеткие классы расстояний: на отрезке $[0,d_{\max}]$ определяются Q нечетких множеств $D_1,...,D_Q$, заданных функциями принадлежности $\nu_1,...,\nu_Q;\nu_i:[0,d_{\max}]\to[0,1]$. Введенные нечеткие множества определяют степени принадлежности $\nu_1(d),...,\nu_Q(d)$ произвольного расстояния $d\in[0,d_{\max}]$ к Q нечетким классам расстояний.

На основе введенных нечетких классов особых точек и расстояний, разработан метод построения алфавита нечетких структурных 3D-дескрипторов, изложенный в **разделе 2.3.3**.

При формировании всевозможных декартовых произведений вида $A_{i_1}^1\times A_{i_2}^2\times...\times A_{i_L}^L$, где $1\le i_1\le n_1,...,1\le i_j\le n_j,...,1\le i_L\le n_L$, функция принадлежности точки $P=(p^1,...,p^L)\in R^L$ к нечеткому множеству такого вида записывается в виде $\mu_{i_1,...,i_L}(P)=\mu_{i_1}^1(p^1)\cdot...\cdot\mu_{i_L}^L(p^L)$. Множество $R=\prod_{i=1}^{L} n_i$ построенных нечетких множеств обозначим через A^1.

---

[10] Mamdani E.H. Application of fuzzy algorithms for control of a simple dynamic plant. Proceedings of IEEE, vol.121, pp.1585-1588, 1974

[11] Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modelling and control, IEEE Transactions on Systems, man, and Cybernetics, vol.15, pp.116-132, 1985

Далее, для каждой особой точки, определяется степень ее принадлежности к каждому нечеткому множеству $\mu_{i_1,\ldots,i_L}, 1 \le i_l \le n_l$, и производится суммирование данных степеней принадлежности для каждого нечеткого множества. Таким образом, алфавит дескрипторов первого уровня AD^1 сформирован перечислением символьных строк вида $"i_1 i_2 \ldots i_L"$, $1 \le i_l \le n_l$, соответствующих нечетким множествам $\mu_{i_1,\ldots,i_L}$. Для каждого дескриптора такого вида его значение для молекулы M с особыми точками $P_1,\ldots,P_n$ определяется как

$$N("i_1 i_2 \ldots i_L") = \sum_{i=1}^{n} \mu_{i_1,\ldots,i_L}(P_i) \tag{1}$$

Алфавит дескрипторов AD^2 для пар особых точек строится следующим образом. Рассмотрим декартово произведение вида $A^1 \times A^1 \times D$. Его элементами являются нечеткие множества – декартовы произведения нечетких множеств вида

$$\left(A^1_{i_1} \times A^2_{i_2} \times \ldots \times A^L_{i_L}\right) \times \left(A^1_{j_1} \times A^2_{j_2} \times \ldots \times A^L_{j_L}\right) \times D_k \tag{2}$$

для всевозможных наборов $(i_1, i_2, \ldots, i_L, j_1, j_2, \ldots, j_L, k)$. Чтобы избежать повторения, рассмотрены только элементы, в которых $(i_1, i_2, \ldots, i_L) \le (j_1, j_2, \ldots, j_L)$ (лексикографический порядок). Таким образом, получен набор из $\frac{Q\left|AD^1\right|(\left|AD^1\right|+1)}{2}$ нечетких множеств вида (2).

Для произвольной пары особых точек (P_1, P_2), находящихся на расстоянии $\rho(P_1, P_2)$ друг от друга, ее степень принадлежности к множеству вида (2) записывается в виде

$$\mu_{i,j,k}(P_1, P_2) = \mu_{i_1,\ldots,i_L}(P_1)\,\mu_{j_1,\ldots,j_L}(P_2)\,\nu_k(\rho(P_1, P_2)) \tag{3}$$

Алфавит дескрипторов второго уровня AD^2 формируется перечислением всевозможных символьных строк вида $"i_1 i_2 \ldots i_L j_1 j_2 \ldots j_L k"$, где сохраняется лексикографическое упорядочивание $(i_1, i_2, \ldots, i_L) \le (j_1, j_2, \ldots, j_L)$ и $1 \le k \le Q$. Для каждого дескриптора такого вида его значение для молекулы M с особыми точками $P_1,\ldots,P_n$ определяется как сумма степеней принадлежности всех структурных 2-фрагментов молекулы соответствующему нечеткому множеству, т.е.:

$$N("i_1 i_2 \ldots i_L j_1 j_2 \ldots j_L k") = \sum_{l,m=1}^{n} \mu_{i,j,k}(P_l, P_m) \tag{4}$$

Аналогично, можно построить алфавит AD^3, рассмотрев декартовы произведения вида $A^1 \times A^1 \times D \times A^1 \times D$ и в них элементы, для которых сохраняется лексикографический порядок. В результате, формируется набор из

$\frac{Q^2(|AD^1|+1)AD^1(|AD^1|-1)}{6}$ нечетких множеств с функциями принадлежности вида

$$\mu_{i,j,h_1,k,h_2}(P_1,P_2,P_3)=\mu_i(P_1)\mu_j(P_2)\nu_{h_1}(\rho(P_1,P_2))\mu_k(P_3)\nu_{h_2}(\rho((P_1,P_2),P_3)) \quad (5)$$

Каждый дескриптор представляет собой символьную строку "$i_1i_2...i_Lj_1j_2...j_Lh_1k_1k_2...k_Lh_2$", где сохраняется лексикографическое упорядочивание $(i_1,i_2,...,i_L)\le(j_1,j_2,...,j_L)\le(k_1,k_2,...,k_L)$ и $1\le h_1,h_2\le Q$. Значение такого дескриптора для молекулы M с особыми точками $P_1,...,P_n$ равно сумме степеней принадлежности всех структурных 3-фрагментов молекулы соответствующему нечеткому множеству, т.е.:

$$N("i_1i_2...i_Lj_1j_2...j_Lh_1k_1k_2...k_Lh_2")=\sum_{l,m,q=1}^{n}\mu_{i,j,h_1,k,h_2}(P_l,P_m,P_q) \quad (6)$$

где $\rho((P_l,P_m),P_q)$ - расстояние от пары особых точек (P_l,P_m) до P_q в смысле наименьшего, наибольшего или среднего из расстояний $\rho(P_l,P_q)$ и $\rho(P_m,P_q)$.

Сходным образом можно усложнять описание далее, построив алфавиты $AD^4,...,AD^P$. Объединив построенные алфавиты, получим алфавит дескрипторов $AD=AD^1\cup AD^2\cup...\cup AD^P$.

Доказана теорема о том, что алфавит четких дескрипторов является частным случаем алфавита нечетких дескрипторов.

Теорема. Для любого алфавита четких дескрипторов, порожденного разбиением отрезков $[p^l_{\min},p^l_{\max}]$ на n_l подотрезков и отрезка $[0,\ d_{\max}]$ на Q подотрезков, существуют семейства нечетких функций принадлежности $\mu^l_1,...,\mu^l_{n_l},l=1,\dots,L,\mu^l_i:[p^l_{\min},p^l_{\max}]\to[0,1]$ и $\nu_1,...,\nu_Q;\nu_i:[0,d_{\max}]\to[0,1]$, такие что алфавит и значения нечетких дескрипторов, порожденных функциями принадлежности $\mu^l_1,...,\mu^l_{n_l},l=1,\dots,L$; $\nu_1,...,\nu_Q$ совпадают с алфавитом и значениями четких дескрипторов.

Описание нечеткими дескрипторами позволяют разработать алгоритм оптимального описания – выбора функций принадлежности μ^l_i и ν_k, от которого значительно зависит качество конечного прогноза. Предложенный алгоритм приведен в **разделе 2.4.**

В **подразделе 2.4.1** приводится общая схема алгоритма оптимизации описания:

1. Формируется начальный алфавит дескрипторов AD_0, в котором функции принадлежности строятся на основе гипотез о пространственных структурах, отвечающих за исследуемую активность.

2. Строится матрица «молекула-признак» обучающего множества в текущем алфавите.

3. По построенной матрице «молекула-признак» выбирается наилучшая линейная предсказывающая модель f для свойства y в смысле функционала качества $\varphi(f)$.

4. На основе построенной модели формируется двухслойная схема из функциональных элементов особого вида, соединенных между собой, так что на входе схема получает молекулу, представленную набором особых точек, на выходе первого слоя – выдает значения дескрипторов, а на выходе второго слоя – результат применения классифицирующей функции.

5. Проводится обучение схемы с тем, чтобы максимизировать функционал качества классификации.

Общий вид рассматриваемой схемы приведен на рис. 1.

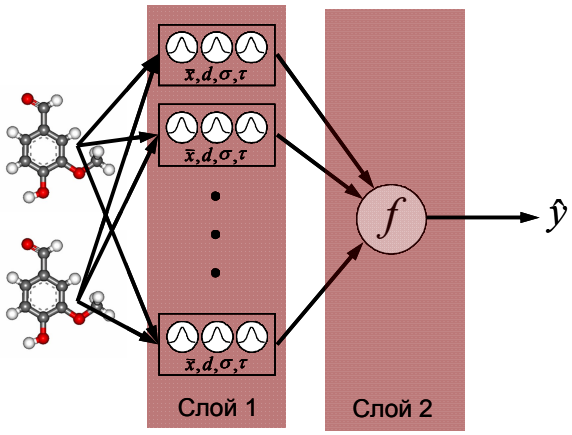

Рис.1. *Схема функциональных элементов для оптимизации алфавита дескрипторов*

В **подразделе 2.4.2** приведен алгоритм формирования начального алфавита дескрипторов. В **подразделе 2.4.3** доказана

Теорема. Для вышеописанной двухслойной схемы функциональных элементов возможно обучение (т.е. подбор параметров функциональных элементов) последовательным применением метода наискорейшего градиентного спуска для определения параметров первого слоя схемы и оценки методом наименьших квадратов для функциональных элементов второго слоя схемы.

На первом проходе обучаются параметры первого слоя (p,d,σ,τ) методом наискорейшего градиентного спуска: при зафиксированных параметрах второго слоя вычисляются частные производные по каждому из параметров α из

первого слоя и каждый из параметров сдвигается на $\Delta\alpha = \eta \frac{\partial \varphi}{\partial \alpha}$, где η – коэффициент сдвига, выбираемый отдельно.

На втором проходе ищется оценка наименьших квадратов, получаемая вычислением $w^* = (X^T X)^{-1} X^T y$, где X – матрица «молекула-признак» размера $N \times M$, построенная применением к молекуле функциональных элементов слоя 1, $\overline{w} = (w_1, ..., w_M)$, y – прогнозируемый вектор классов активности молекул обучающего множества.

Оценка наименьших квадратов может быть получена либо прямым вычислением, либо итерационно, в случае если велико число обусловленности матрицы $(X^T X)^{-1}$.

Раздел 2.5 посвящен реализованному алгоритм решения задачи «структура-свойство» с использованием нечетких дескрипторов. Детально описаны особенности алгоритма:

- Реализовано эволюционное построение дескрипторов: дескрипторы *n*-ого порядка формируются на основе наиболее информативных дескрипторов (*n*-1)-ого порядка.
- Осуществляется поиск разных функций принадлежности для расстояний между разными типами структурных фрагментов. В частности, происходит отдельное разбиение интервала расстояний для каждой пары меток ОТ / пары дескриптора (*n*-1)-ого порядка и метки ОТ.
- Рассматриваются только структурные фрагменты (пары и тройки ОТ), присутствующие не менее чем в определенной доле соединений.
- Осуществляется оптимизация функций принадлежности по угловому коэффициенту.

Особенности данного алгоритма позволяют добиться следующего:

- устранить эффект «комбинаторного взрыва» – ситуации, когда при обработке дескрипторов высших уровней формируется большое число дескрипторов, что приводит к значительным вычислительным затратам при вычислении значений дескрипторов, а также делает вычислительно неэффективным применение многих методов классификации и регрессии;
- обеспечить распространенность задействованных дескрипторов;
- оптимизировать тип используемых функций принадлежности.

В **разделе 2.6** приведена оценка сложности предложенного алгоритма. Обозначим через T число меток (типов) ОТ; D_k – число нечетких множеств, заданных на интервале значений расстояния между структурным фрагментом *k*-ого уровня (ОТ, парой или тройкой ОТ) и ОТ; Q – количество наиболее информативных дескрипторов, на основе которых формируются дескрипторы следующего уровня (в случае применения МГУА в качестве классифицирующей функции, равно глубине МГУА, умноженной на рассматриваемое число лучших моделей). $S^k_{\text{опис}}$, $F^k_{\text{опис}}$, $E^k_{\text{опис}}$ – количество

операций этапа описания при использовании структурных дескрипторов, нечетких дескрипторов и нечетких дескрипторов при эволюционном построении, соответственно; $S^k_{\text{мгуа}}$, $F^k_{\text{мгуа}}$, $E^k_{\text{мгуа}}$ – аналогичные показатели для этапа построения классифицирующей модели с МГУА как алгоритмом построения модели. Справедлива

Теорема. Имеют место следующие оценки:

а) $F^k_{\text{опис}} = S^k_{\text{опис}} O\left(\prod_{i=1}^{k} D_i\right)$;

б) $S^k_{\text{мгуа}} = F^k_{\text{мгуа}} = E^k_{\text{мгуа}} O\left(\frac{T^{k-1}\prod_{i=2}^{k-1} D_i}{Q}\right)$.

Из утверждения теоремы следует, что применение нечетких дескрипторов увеличивает сложность построения алфавита дескрипторов по сравнению с использованием четких дескрипторов, однако не увеличивает сложность построения прогнозирующей функции. При этом эволюционное построение нечетких дескрипторов позволяет существенно снизить общее количество операций, за счет снижения количества дескрипторов, и как следствие, уменьшения вычислительной сложности этапа анализа.

Третья глава посвящена изучению эффективности предложенных автором методов на практике. В главе описана программная реализация предложенных алгоритмов, приведено описание вычислительных экспериментов, проведен анализ их результатов.

В **разделе 3.1** описана программная реализация алгоритма. **Раздел 3.1.1** детально приводит этапы расчета пространственной структуры и электростатического заряда молекулярных графов, построения триангулированных молекулярных поверхностей, нахождения и маркировки особых точек. Реализация этапов формирования матрицы «молекула-признак» и поиска классифицирующей функции в среде MATLAB приведена в **разделе 3.1.2**.

В **разделе 3.2** описаны использованные методы построения классифицирующей функции:

- МГУА на кластерах;
- МГУА, использующий в качестве опорных функций конъюнкции и дизъюнкции ряда дескрипторов [10];
- ANFIS[12] на главных компонентах[13];

[12] J.-S. Roger Jang, C.-T. Sun and E. Mizutani, ``Neuro-Fuzzy and Soft Computing: a computational approach to learning and machine intelligence," 1996, to be published by Prentice-Hall

- МГУА с использованием метода ближайших соседей (МГУА-kNN) [9].

В **разделе 3.3** приведены результаты применения вышеописанных методов к алфавитам четких и нечетких дескрипторов, построенных для следующих выборках химических соединений:

- выборка гликозидов, протестированная на противоопухолевую активность;
- выборка соединений бициклической мочевины, протестированных на общую токсичность и транквилизирующую активность.

Проведено сравнение результатов при применении четких дескрипторов, а также при различных модификациях алгоритма с использованием нечетких дескрипторов. Результаты численных экспериментов подтвердили эффективность и перспективность методов и алгоритмов, разработанных на основе аппарата нечеткой логики: при обработке определенными методами машинного обучения (например, МГУА-kNN для выборки бициклических бисмочевин и ANFIS на главных компонентах для выборки гликозидов) наблюдалось заметное улучшение качества прогноза при переходе от четкого описания молекулярной структуры к нечеткому. При использовании остальных методов машинного обучения четкие и нечеткие дескрипторы приводят к сопоставимым результатам.

В частности, на выборке бициклической мочевины среднее качество прогноза методом МГУА-kNN улучшается при продвижении от более четких функций принадлежности к более нечетким: 78.0% для четких функций принадлежности, 82.6% для нечетких трапециевидных, 86.3% для нечетких треугольных. При этом максимальное значение качества прогноза 96.9% также достигается при использовании треугольных функций принадлежности

В **заключении** сформулированы результаты, полученные в рамках настоящей диссертационной работы и приведено обсуждение перспективы развития данного метода – адаптации к описанию молекул с множеством устойчивых пространственных конфигураций.

В **приложении** приведено описание обработанных выборок химических соединений.

Основные результаты диссертации, выносимые на защиту

1. Разработаны методы представления информации о пространственных конфигурациях молекул и молекулярных поверхностях с помощью четких и нечетких структурных 3D-дескрипторов.

2. В рамках предложенных методов предложены алгоритмы формирования четких и нечетких структурных 3D-дескрипторов – новых моделей молекулярных дескрипторов, последняя из которых учитывает гибкость

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

пространственной структуры молекулы, а также алгоритм оптимизации нечеткого описания молекул с целью поиска локально лучшей модели в некотором классе предсказывающих функций.

3. Проведена оценка вычислительной сложности алгоритмов представления информации о пространственных структурах молекул в виде четких и нечетких структурных 3D-дескрипторов и последующего анализа полученных данных.

4. Проведено исследование предложенных алгоритмов: в ходе тестовых испытаний по обнаружению функциональной зависимости «структура-свойство» на четких и нечетких структурных 3D-дескрипторах подтверждена перспективность последних.

Благодарность

Автор выражает глубокую признательность своему научному руководителю Кумскову Михаилу Ивановичу на постановку задач, постоянное внимание к работе и многочисленные плодотворные обсуждения. Автор также выражает благодарность заведующему кафедрой вычислительной математики профессору Кобелькову Георгий Михайловичу и всем сотрудникам кафедры за творческую атмосферу и поддержку, а также к.б.н. Апрышко Галине Николаевне (Российский онкологический научный центр имени Н.Н. Блохина), д.х.н. Кравченко Ангелине Николаевне и к.х.н. Свитанько Игорю Валентиновичу (Институт органической химии имени Н.Д. Зелинского РАН) за предоставление выборок химических соединений.

Список опубликованных работ по теме диссертации

Основные результаты диссертации содержатся в следующих статьях:

1) I.V. Svitanko, D.A. Devetyarov, D.E. Tcheboukov, M. S. Dolmat, A.M. Zakharov, S.S. Grigoryeva, V.T. Chichua, L.A. Ponomareva, M.I. Kumskov. QSAR Modeling on the Basis of 3D Descriptors Representing the Electrostatic Molecular Surface (Ambergris Fragrances) // Mendeleev Communications. – 2007. – Vol.17, No. 2. – P. 90-91. (Автору диссертации принадлежит реализация алгоритма и проведение численных экспериментов)

2) D.A. Devetyarov, A.M. Zaharov, M.I. Kumskov, L.A. Ponomareva. Fuzzy logic application for construction of 3D descriptors of molecules in QSAR problem. // Proceeding of the 8th International Conference "Pattern Recognition and Image

Analysis: New Information Technologies" (PRIA-8-2007) – 2007. – Vol.2. – P.249-252. (Автору диссертации принадлежат разработанный алгоритм и результаты вычислительных экспериментов)

3) S.S. Grigoreva, M.I. Kumskov, A.M. Zaharov, D.A. Devetyarov, L.A. Ponomareva, I.V. Svitanko. Search of 3D structure representation of flexible molecules adequate to the given biological activity // Proceeding of the 8th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-8-2007) – 2007. – Vol.2. – P.262-265. (Автору диссертации принадлежат реализация алгоритма и проведение численных экспериментов)

4) Григорьева С.С., Чичуа В.Т., Деветьяров Д.А., Кумсков М.И. Выбор оптимального описания структуры молекулы в задаче структура-свойство для заданной биологической активности // Вестник Московского Университета. Серия 2. Химия. – 2007. – Т. 48, N 5. – С. 305-307. (Автору диссертации принадлежат реализация алгоритма и проведение численных экспериментов)

5) Деветьяров Д.А., Григорьева С.С., Пермяков Е.А., Кумсков М.И., Пономарева Л.А., Свитанько И.В. Решение задачи «структура-свойство» для молекул с множеством пространственных конформаций // Система прогнозирования свойств химических соединений: Алгоритмы и модели. Сборник научных работ. – М.: МАКС Пресс, 2008. – С. 3-9. (Автору диссертации принадлежат алгоритмы 1 и 2 и результаты экспериментов по выборке зеленого запаха)

6) Григорьева С.С., Деветьяров Д.А., Свитанько И.В., Пермяков Е.А., Апрышко Г.Н., Кумсков М.И. Поиск представлений 3D структур гибких молекул в задаче прогнозирования биологической активности // Система прогнозирования свойств химических соединений: Алгоритмы и модели. Сборник научных работ. – М.: МАКС Пресс, 2008. – С. 10-36. (Автору диссертации принадлежат разработанные метод и алгоритмы описания гибких молекул)

7) Захаров А.М., Деветьяров Д.А., Кумсков М.И. Решение задачи «структура-активность» с использованием нечетких функций близости (kernel-функций) // Система прогнозирования свойств химических соединений: Алгоритмы и модели. Сборник научных работ. – М.: МАКС Пресс, 2008. – С. 37-50. (Автору диссертации принадлежат разработанный алгоритм (раздел 4) и доказательство теоремы 1)

8) 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. (Автору диссертации принадлежит метод и алгоритм формирования дескрипторов, экспериментальные результаты на этапе описания, а также экспериментальные результаты на этапе анализа методом МГУА)

9) Носеевич Ф.М., Деветьяров Д.А., Кумсков М.И., Апрышко Г.Н., Пермяков Е.А. Двоичный метод группового учета аргументов в задаче «структура-свойство» // Доклады 14-ой Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. – М: МАКС Пресс. – 2009. – С. 575-578. (Автору диссертации принадлежит метод и алгоритм формирования дескрипторов, экспериментальные результаты на этапе описания)

10) Деветьяров Д.А. Нечеткие дескрипторы молекул в задаче «Структура-свойство» // Информационные технологии. – 2010. – №3. – С. 73-74.

11) Деветьяров Д.А., Кумсков М.И. Использование нейронных сетей в задаче «структура-свойство» с использованием нечеткого описания пространственных структур молекул // Нейрокомпьютеры: разработка, применение. – № 3. – С. 14-19. (Автору диссертации принадлежат метод создания искусственных нейронных сетей на основе кусочно-линейных моделей «Структура-свойство» и экспериментальные результаты)

12) Деветьяров Д.А. Эволюционное построение алфавита дескрипторов, сформированных на основе аппарата нечеткой логики, в задаче «структура-свойство» // Системы управления и информационные технологии. – 2010. – 2010. – № 1.1 (39). – С. 131-134.