ФГБОУ ВПО «Московский государственный университет
имени М.В. Ломоносова»

На правах рукописи

ПРОХОРОВ Евгений Игоревич

Адаптивная двухфазная схема решения задачи «структура – свойство»

Специальность 05.13.17 – теоретические основы информатики

ДИССЕРТАЦИЯ

на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
доктор физико-математических наук,
профессор М.И. Кумсков

Москва – 2014

Содержание

Введение

Стремительное развитие средств вычислительной техники, происходящее в последние десятилетия, позволило широко применять методы и алгоритмы информатики для анализа данных в больших хранилищах. В частности появились технологии и вычислительные системы для хранения и анализа данных о структуре различных химических соединений. Для обозначения применения методов информатики для решения химических задач используется специальный термин хемоинформатика [1]. В общем смысле хемоинформатика – название научных исследований, охватывающих процессы дизайна, создания, организации, управления, поиска, анализа, распространения, визуализации и использования информации о химических соединениях [2]. В частном случае под хемоинформатикой подразумевают также использование информационных ресурсов для преобразования данных в знания для принятия наилучших решений при поиске соединений-лидеров в разработке лекарств [3]. Методы хемоинформатики в настоящее время начинают активно внедряться во все области химии, и, прежде всего, в органическую химию. Одной из ключевых задач хемоинформатики является задача поиска количественных соотношений «структура – свойство» [4].

С точки зрения математики задача состоит в поиске численной зависимости между структурой молекулы химического соединения и её физико-химическими свойствами или биологической активностью. В англоязычной литературе для обозначения этих двух разновидностей рассматриваемой задачи существуют термины QSPR (Quantity Structure Property Relationship) и QSAR (Quantity Structure Activity Relationship), соответственно [5, 6].

Математические модели «структура – свойство» и «структура – активность» позволяют выявлять потенциально активные молекулы в больших ба-

зах химических соединений, а также осуществлять синтез веществ с заранее заданными свойствами. Поэтому модели «структура – свойство» / «структура – активность» применяются в процессе разработки новых лекарственных препаратов для поиска химических соединений, обладающих нужным видом биологической активности. Вычислительная процедура, которая включает автоматизированный просмотр базы данных химических соединений и отбор тех из них, для которых прогнозируется наличие желаемых свойств, носит название виртуальный скрининг [2, 7]. Использование виртуального скрининга позволяет существенно сократить объем длительных и дорогостоящих экспериментальных исследований в области химии, медицины и гии [8].

В настоящей диссертационной работе рассматривается задача «структура – свойство», которая состоит в поиске численной зависимости между структурой химических соединений, представленных своими молекулярными графами (*М-графами*), и их химическими свойствами, представленными заданным конечным набором классов. Под молекулярным графом подразумевается помеченный граф, вершины которого интерпретируются как атомы, а ребра как валентные связи между парами атомов. Метки вершин и ребер (числа или символы) кодируют атомы и связи различной химической природы.

В работе рассматриваются М-графы с числом вершин, не превосходящих заданной величины T. Такое ограничение с одной стороны обусловлено необходимостью изъять из рассмотрения М-графы, соответствующие высокомолекулярным соединениям (молекулы которых содержат сотни и тысячи атомов), а с другой позволяет более точно оценить вычислительную сложность предлагаемых алгоритмов. Множество М-графов с числом вершин, не превосходящих T, обозначим TG.

Зависимость ищется на ограниченном подмножестве TG, называемом обучающей выборкой $LS \subset TG$. Полученный в результате поиска набор решающих правил называют моделью «структура – свойство». Модель «структура – свойство» осуществляет прогнозирование свойств молекулярных графов из TG (отнесение М-графа к одному из заданных классов). Процесс отнесения М-графа к одному из заданных классов называется также классификацией М-графов.

В рамках данной диссертационной работы рассматривается подход к описанию структур М-графов на базе фрагментных дескрипторов (различных уровней) особых точек М-графов [9]. Особыми точками выступают цепочки вершин М-графа (атомов). Значения дескрипторов задаются как число повторений фрагментов, соответствующих особым точкам, их парам, тройкам и четверкам. При переходе к каждому следующему уровню описания, вычислительная сложность дескрипторов увеличивается пропорционально количеству различных меток вершин М-графов в степени p, где p – длина цепочки атомов, задающей особую точку (является параметром описания). Далее отображение, ставящее в соответствие *M*-графу $G \in TG$ его вектор дескрипторов $x = (x_1, \dots x_M) \in \mathbb{R}^M$ называется *описывающим* и обозначается $D : TG \to \mathbb{R}^M$. Процесс вычисления значений дескрипторов для множества М-графов называется *дескрипторным описанием*.

Формально *моделью «структура – свойство» или распознающей моделью RM* в настоящей работе называется совокупность решающих правил, полученная на обучающей выборке *LS* и обладающую следующими свойствами.

- Для молекулярного графа $G \in TG$ и его описания в виде вектора признаков $x = (x_1, \dots x_M) \in \mathbb{R}^M$ с помощью фиксированного описывающего отображения $D : TG \to \mathbb{R}^M$, распознающая модель *RM* ли-

бо осуществляет прогноз его свойства (отнесение М-графа к одному из H классов $\{Cl_1, Cl_2, \ldots Cl_H\}$), либо производит отказ от прогноза.

- Для распознающей модели может быть вычислен показатель качества $\phi(RM)$, характеризующий её качество прогноза на обучающей выборке. В настоящей работе для определения качества модели используется процент верно классифицированных М-графов в процессе выполнения процедуры скользящего контроля.

Процедура скользящего контроля (leave-one-out cross-validation) [10] заключается в следующем: из обучающей выборки последовательно удаляется каждый М-граф, по оставшимся М-графам строится распознающая модель, и с помощью этой модели прогнозируется свойство удаленного М-графа.

Модель, построенную с помощью фиксированного алгоритма обучения по обучающей выборке, будем обозначать $RM(LS)$. В общем случае показатель качества модели может быть вычислен на контрольной выборке – множестве М-графов, отличном от обучающей выборки, при условии, что в процессе классификации каждого М-графа из контрольной выборки, обучающая выборка не содержит классифицируемый М-граф (аналог скользящего контроля). Значение показателя качества, вычисленное по выборке CS, обозначим $\varphi(RM, CS) = \varphi(RM(LS), CS)$. При этом $\varphi(RM) := \varphi(RM, LS)$.

В случае, когда заранее задано дескрипторное описание молекулярных графов, задача «структура – свойство» сводится к задаче классификации [11]. В свою очередь задача «структура – активность» сводится к задаче регрессии [12]. Для обеих задач могут быть применены математические методы теории распознавания образов и методы машинного обучения [13]. Одним из оригинальных подходов к решению задачи можно считать предложенный В.К.Финном ДСМ-метод автоматического порождения гипотез [14]. Наряду с методами, использующими дескрипторное описание, существуют также мно-

гочисленные беспризнаковые подходы [15, 16, 17] и подходы, в которых вместо дескрипторов напрямую используются молекулярные графы и их «проекции», задающиеся с помощью специально определенной операции пересечения [18, 19].

Классические модели «структура – свойство» обладают существенными недостатками при практическом применении. Во-первых, модель «структура – свойство» представляет собой классификатор, обученный на некоторой ограниченной выборке М-графов. По этой причине она предсказуемо неэффективна на М-графах, принципиально отличных от тех, что использовались при обучении классификатора. При использовании таких моделей на практике, например, для скрининга больших баз химических соединений, модель осуществляет прогноз всех молекулярных графов без ограничений, и большинство таких прогнозов оказывается несостоятельными.

Вторым существенным недостатком является то обстоятельство, что для обеспечения высокого качества прогнозирования необходимо использовать вычислительно сложные дескрипторы. Например, такое химическое свойство, как хиральность – правосторонняя или левосторонняя ориентация М-графа, может быть представлено в рамках рассматриваемого подхода только при использовании фрагментов четвертого уровня. При этом описание неоднородных выборок М-графов может содержать сотни и тысячи дескрипторов. Вычислительная сложность прогнозирования свойств новых неизученных М-графов в этом случае очень высока, что делает полученные модели «структура – свойство» практически неприменимыми для задач виртуального скрининга.

Таким образом, актуальной является разработка нового подхода к решению задачи «структура – свойство», реализующего построение и использование ограничений допустимости для конкретной модели, а также автоматическую адаптацию дескрипторного описания под задачу прогнозирования кон-

кретного химического свойства с целью уменьшения вычислительной сложности прогнозирования свойств неизученных М-графов.

Ограничением допустимости или правилом отказа для распознающей модели *RM* в задаче классификации «структура – свойство» назовём некоторую функцию $g : TG \to \{0,1\}$ со следующей интерпретацией: $g(G) = 1$ будет означать отказ от прогноза свойства данного молекулярного графа, в противном случае прогноз может быть осуществлён. Множество допустимых М-графов обучающей выборке обозначим LSG. Правило отказа g назовем *эффективным*, если для него выполнено неравенство $\varphi(RM, LSG) > \varphi(RM, LS)$.

С учетом изложенного выше **целью диссертационной работы** являлась разработка метода построения моделей «структура – свойство» с использование эффективных в смысле данного выше определения ограничений допустимости, а также разработка метода выбора дескрипторного описания, снижающего вычислительную сложность процесса прогнозирования свойств неизученных М-графов.

Научная новизна. Предложен новый подход к решению задачи «структура – свойство» на базе описания структур молекулярных графов фрагментными дескрипторами особых точек, использующий ограничения допустимости для моделей «структура – свойство» и позволяющий сократить вычислительную сложность прогнозирования неизученных М-графов за счет использования неоднородного описания. В его рамках предложен оригинальный метод построения моделей «структура – свойство», включающий разработку решающих правил для определения допустимости М-графов для моделей, а также автоматический выбор дескрипторного описания. Проведена оценка качества прогнозирования для получаемых моделей «структура – свойство» и оценка вычислительной сложности разработанных алгоритмов. Предложенный подход позволяет избавиться от основных недостатков существующего решения на базе фрагментных дескрипторов особых точек, связанных с осо-

бенностями прикладных задач прогнозирования свойств химических соединений.

Основные результаты диссертации, выносимые на защиту

- Построена формальная модель, описывающая ограничения допустимости, которые необходимы для математического моделирования функциональной зависимости «структура – свойство». Дано определение эффективности использования для таких ограничений.
- Разработан метод решения задачи «структура – свойство», реализующий построение и использование ограничений допустимости для моделей «структура – свойство». Получены теоретические оценки эффективности использования предложенных ограничений и качества моделей.
- Разработан подход к описанию структуры молекулярных графов на базе фрагментных дескрипторов особых точек, позволяющий снизить вычислительную сложность построения моделей «структура – свойство».
- Представлены алгоритмы построения ограничений допустимости, а также алгоритмы адаптации дескрипторного описания под задачи поиска функциональной зависимости «структура – свойство» и построения ограничений допустимости. Приведена оценка вычислительной сложности для данных алгоритмов.
- На базе представленных в диссертации методов и алгоритмов построены и программно реализованы модели «структура – свойство» для прогнозирования противоопухолевой активности и способности ингибировать активность поли-(АДФ-рибоза)-полимеразы-1. Полученные оценки эффективности использования ограничений допустимости, качества моделей и вычислительной сложности разработанных алгоритмов подтверждены результатами тестирования.

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 работы состоит в том, что разработанные алгоритмы решения задачи «структура – свойство» 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 экспериментальным скринингом на больших наборах химических соединений. Практическая значимость работы подтверждена в серии прикладных научных исследований совместно с учеными из Института Органической Химии им. Н.Д. Зелинского РАН и Российского Онкологического Научного Центра им. Н.Н. Блохина РАМН.

Материалы работы докладывались и обсуждались на следующих всероссийских и международных конференциях.

1. Международная научная конференция «Компьютерные науки и информационные технологии» (1 – 4 июля 2009 г., Саратов).
2. 14-ая Всероссийская конференция «Математические методы распознавания образов» (21 – 26 сентября 2009 г., Суздаль).
3. XVII Международная конференция студентов, аспирантов и молодых учёных «Ломоносов» (12 – 15 апреля 2010 г., Москва).
4. 10-ая Международная конференция «Распознавание образов и анализ изображений: новые информационные технологии» (5 – 12 декабря 2010 г., Санкт-Петербург).
5. Специальный семинар «The International Workshop on Soft Computing Applications and Knowledge Discovery» в рамках 4-ой Международной конференции «Pattern Recognition and Machine Intelligence» (June 24, 2011, Moscow).

6. Международная конференция «Ломоносовские чтения 2012» (16 – 25 апреля, 2012 г., Москва).

7. 9-ая международная конференция «Интеллектуализация обработки информации» (16 – 22 сентября 2012 г., Будва, Черногория).

8. XX российский национальный конгресс «Человек и Лекарство» (15 – 19 апреля 2013 г., Москва).

9. 23-я Международная конференция по компьютерной графике и зрению ГрафиКон'2013 (16 – 20 сентября 2013 г., Владивосток).

Полученные результаты прошли апробацию также на специальном семинаре механико-математического факультета МГУ им. Ломоносова «Методы решения задачи «структура – свойство»» под руководством проф. д.ф.-м.н. М.И. Кумскова (2010 – 2013, неоднократно), на научно-исследовательском семинаре «Дискретная математика и математическая кибернетика» кафедры математической кибернетики факультета вычислительной математики и кибернетики МГУ им. Ломоносова под руководством проф. д.ф.-м.н. В.Б. Алексеева, проф. д.ф.-м.н. А.А. Сапоженко и проф. д.ф.-м.н. С.А. Ложкина (2014 г.), на семинаре «Теория автоматов» кафедры математической теории интеллектуальных систем механико-математического факультета МГУ им. Ломоносова под руководством академика В.Б. Кудрявцева (2014 г.), на учебно-исследовательском семинаре кафедры математических методов прогнозирования факультета вычислительной математики и кибернетики МГУ им. Ломоносова «Интеллектуальный анализ данных: новые задачи и методы» под руководством к.ф.-м.н. С.И. Гурова и к.ф.-м.н. Майсурадзе (2014 г.), на «Объединенном семинаре по проблемам химической информатики» физического факультета МГУ им. Ломоносова под руководством д.ф.-м.н. И.И. Баскина (2013 г.), на научном семинаре «Проблемы современных информационно-вычислительных систем» под руководством проф. д.ф.-м.н. В.А. Васенина (2013 г.), на научном семинаре «Математические модели информационных технологий» отделения прикладной математики и инфор-

матики НИУ ВШЭ под руководством проф. д.ф.-м.н. С.О. Кузнецова (2013 – 2014, неоднократно), на семинаре проблемной комиссии «Биоинформатика в создании новых лекарств» российской секции «The Cheminformatics and QSAR Society» под руководством проф. д.б.н. В.В. Поройкова (базовая организация – ИБМХ им. Ореховича РАМН, 2013 г.).

Основные результаты, выносимые на защиту, содержатся в следующих работах:

1. **Прохоров Е.И.** Нейронные сети для построения ограничений допустимости в задаче «структура – свойство» // Нейрокомпьютеры: разработка, применение. – 2012. – № 10. – С. 46 – 56.

2. **Prokhorov E.I.**, Ponomareva L.A., Permyakov E.A., Kumskov M.I. Fuzzy classification and fast rejection rules in the structure-property problem // Pattern Recognition and Image Analysis, 2013, Volume 23, Number 1, Pp. 130–138. (Е.И. Прохорову принадлежит разработка нечеткого классификатора).

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. – 2010. – № 3. – С. 39-45. (Е.И. Прохорову принадлежит разработка нечеткого классификатора).

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. (Е.И. Прохорову принадлежит разработка нечеткого классификатора).

5. **Прохоров Е. И.** «Нечеткое» прогнозирование свойств химических соединений: Использование нечеткой функции классификации на кластерах обучающего множества в задаче «структура – свойство», Saarbrucken, Germany: LAP Lambert Academic Publishing, 2012, – 80 с.

6. 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. (Е.И. Прохорову принадлежит разработка метода нечеткого прогнозирования активности).

7. 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. (Е.И. Прохорову принадлежит реализация алгоритма нечеткого логического вывода для построения моделей «структура – свойство»).

8. Prokhorov E.I., Ponomareva L.A., Permyakov E.A., Kumskov M.I. The fuzzy classification of molecular graphs and fast rejection rules in «structure – property» problem // Proc. 10th Int. Conf. Pattern Recognition And Image Analysis: New Information Technologies – V. 2. – St. Petersburg, 2010. – P. 217–220. (Е.И. Прохорову принадлежит разработка нечеткого классификатора).

9. Eugeny Prokhorov, Ludmila Ponomareva, Eugeny Permyakov and Mikhail Kumskov Fuzzy Predicting Models in «Structure – Property» Problem // Proceedings of the International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD 2011) Pages 89–94 // http://ceur-ws.org/Vol–758/ (the CEUR–Workshop web site). (Е.И. Прохорову принадлежит разработка нечеткого классификатора).

10. Прохоров Е.И., Кумсков М.И., Беккер А.В. Построение и использование адаптивных распознающих моделей для решения задачи «структура – свойство» // Интеллектуализация обработки информации: 9-я международная конференция. Черногория, г. Будва, 2012

г.: Сборник докладов. – М.: Торус Пресс, 2012. (718 с.) С. 581 – 584. (Е.И. Прохорову принадлежит разработка адаптивного подхода к описанию М-графов).

11. Прохоров Е.И., Кумсков М.И., Беккер А.В., Перевозников А.В., Пугачева Р.Б., Апрышко Г.Н. Согласованное прогнозирование противоопухолевой активности по семейству моделей «структура-свойство» // Прогнозирование свойств химических соединений. Унифицированный Репозиторий моделей «структура – свойство»: – Сборник научных работ. – М.: МАКС Пресс, 2012. – С. 25–56. (Е.И. Прохорову принадлежит разработка моделей «структура – свойство» на базе метода опорных векторов).

12. Прохоров Е.И., Беккер А.В., Перевозников А.В., Свитанько И.В., Захаренко А.Л., Суханова М.В., Кумсков М.И. Приложения метода эволюционного отбора дескрипторов в математическом моделировании зависимости биологической активности соединения от его структуры // Прогнозирование свойств химических соединений. Унифицированный Репозиторий моделей «структура – свойство»: – Сборник научных работ. – М.: МАКС Пресс, 2012. – С. 3–24. (Е.И. Прохорову принадлежит разработка моделей «структура – свойство» на базе нечеткого классификатора).

13. Е.И. Прохоров, Г.Н. Апрышко, Р.Б. Пугачева, А.В. Беккер, А.В. Перевозников, М.И. Кумсков Математические методы прогнозирования противоопухолевой активности // XX российский национальный конгресс Человек и Лекарство: Сборник материалов конгресса. – ЗАО РИЦ Человек и лекарство. – Москва, 2013. – С. 415–415. (Е.И. Прохорову принадлежит разработка моделей «структура – свойство» на базе метода опорных векторов).

14. Е. Прохоров, М. Кумсков Двухфазная схема решения задачи классификации \\ Conference Proceedings GraphiCon’2013 \ Труды Конференции ГрафиКон’2013. – Владивосток, 2013. – С. 241–243. (Е.И. Прохорову принадлежит доказательство теоремы и следствий).

Работа поддержана Российским Фондом Фундаментальных Исследований (РФФИ) по грантам №07-07-00282 и №10-07-00694.

Работа состоит из введения, 3-х глав основного текста, заключения и списка литературы. Общий объем диссертации – 137 страниц. Список литературы содержит 67 названий.

В **первой главе** приведена общая постановка задачи «структура – свойство». В **разделе 1.1** кратко рассматриваются этапы решения задачи. В **разделе 1.2** приводятся ключевые особенности задачи «структура – свойство». **Раздел 1.3** содержит основные определения и постановки задач, используемые для формулирования теоретической части работы.

Во **второй главе** приведены различные подходы к решению задачи построения распознающих моделей, поставленной в **разделе 1.3**. Приводятся теоретические результаты. В частности, в **разделе 2.4** дается двухфазная схема решения задачи «структура – свойство». Оценка качества результирующей модели при использовании двухфазной схемы доказана в **2.4.2**. В **разделе 2.7** описан метод понижения вычислительной сложности при обработке неоднородных выборок.

Третья глава содержит результаты практического тестирования подхода. В **разделе 3.1** описана программная реализация предложенных методов. В **разделах 3.2 – 3.4** даны подробные описания проведенных совместных научных исследований с Институтом Органической Химии им. Н.Д. Зелинского РАН и Российским Онкологическим Научным Центром им. Н.Н. Блохина РАМН.

В **Заключении** описаны результаты, полученные в рамках настоящей диссертационной работы, а также приведено описание основных направлений дальнейшей работы.

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 к.б.н. Апрышко Галине Николаевне (Российский онкологический научный центр имени Н.Н. Блохина), к.х.н. Свитанько Игорю Валентиновичу (Институт органической химии имени Н.Д. Зелинского РАН) за предоставление выборок химических соединений.

Глава 1. Задача «структура – свойство»

В главе приведена постановка задачи «структура – свойство». Изложены общие принципы прогнозирования свойств молекулярных графов методами машинного обучения. Кратко рассматриваются этапы решения задачи. Приводятся ключевые особенности задачи «структура – свойство».

1.1 Этапы решения задачи «структура – свойство»

В рамках рассматриваемого в настоящей работе подхода задачу «структура – свойство» можно разбить на две подзадачи: задачу представления информации о структуре М-графа в виде векторов признаков (*этап описания*) и задачу поиска функциональной зависимости f между значениями признаков и значением свойства (*этап поиска функциональной зависимости*).

На первом этапе формируется так называемая *МД*-матрица или матрица «молекулярный граф – дескриптор», содержащая по строкам описание М-графов в виде векторов дескрипторов. Столбцы МД-матрицы, соответственно, содержат значения конкретного дескриптора, вычисленного для каждого из М-графов обучающей выборки. Второй этап посвящён анализу этой матрицы методами машинного обучения и классификации. Указанные этапы решения задачи иллюстрирует **рисунки 1а, 1б**.

Молекула химического соединения

Структура химического соединения представляется в виде молекулярного графа

Этап описания структуры соединения

Формируется алфавит из M дескрипторов, по обучающей выборке строится матрица «молекулярный граф – дескриптор»

Вектор признаков M-графа

$x = (x_1, \dots x_M)$

Вектора признаков используются в качестве представления M-графов для построения распознающих моделей

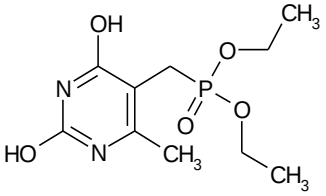
Этап поиска функциональной зависимости

$y = f(x_1, \dots, x_M)$

Результатом этапа является установленная функциональная зависимость между вектором признаков и исследуемым свойством

Рисунок 1а. Этапы решения задачи «структура – свойство»

Молекулярный граф | Вектор признаков | Функциональная зависимость



$\Longrightarrow \quad x = (x_1, \dots x_M) \quad \Longrightarrow \quad y = f(x_1, \dots, x_M)$

Рисунок 1б. Процесс решения задачи «структура – свойство»

Разделение решения задачи «структура – свойство» на перечисленные этапы осуществляется с позиций рассмотренного в настоящей работе подхода с применением фрагментных дескрипторов на особых точках М-графов и большого числа классических решений. Однако существуют подходы к решению задачи, которые в указанную схему не укладываются. Среди них можно отметить подходы на базе анализа формальных понятий (ДСМ-метод автоматического порождения гипотез [20]), а также беспризнаковые подходы, подразумевающие, как правило, выбор способа вычисления меры схожести двух М-графов, не использующего описание в виде дескрипторов [15].

Кроме того даже в рамках классических подходов к основным этапам может добавляться решение следующих подзадач:

- сокращение размерности дескрипторного описания (отбор значимых дескрипторов, поиск условного базиса в пространстве дескрипторов, различные разложения и преобразования МД-матрицы);
- анализ обучающей выборки (степень однородности, разбиения на подмножества, поиск выбросов, кластерный анализ);
- построение локальных прогнозирующих функций (функциональная зависимость устанавливается лишь для некоторых М-графов обучающей выборки).

1.2 Ключевые особенности решения задачи «структура – свойство»

Настоящий раздел посвящен основным особенностям решения задачи «структура – свойство» на базе фрагментных дескрипторов на особых точках М-графов. Данные особенности позволяют уточнить общую постановку задачи и накладывают ограничения на разрабатываемые методы.

Во-первых, отметим, что задача «структура – свойство» оперирует М-графами в качестве объектов классификации. В виду ограниченности мощности обучаемой выборки на практике, предсказание свойств произвольного

молекулярного графа не представляется возможным. Поэтому важно каким-то образом ограничить область применимости построенных моделей [21].

Ключевой особенностью задачи «структура – свойство» является также её ориентированность на предсказание свойств новых неизученных молекулярных графов. При этом надо иметь в виду, что на практике в процессе поиска М-графов, потенциально обладающих рассматриваемым свойством, приходится просматривать огромные базы химических соединений. Этот процесс носит название виртуальный скрининг [2]. Эффективная организация скрининга с помощью построенных моделей «структура – свойство» является важным приоритетом.

Следующая особенность состоит в том, что в рамках использования фрагментных дескрипторов особых точек М-графов для получения качественных моделей «структура – свойство» необходимо вычислять дескрипторы высоких уровней, что влечет за собой повышение количества машинных ресурсов (времени работы алгоритмов и объему используемой памяти). Вычислительная сложность прогнозирования значительно затрудняет практическое использование получаемых моделей.

Ещё одной чертой рассматриваемого подхода является огромная размерность исходного пространства дескрипторов. Слишком большое количество дескрипторов не только увеличивает вычислительную сложность процесса обучения, но и, зачастую, препятствует построению эффективных моделей. Включение в модель избыточных (шумовых) признаков может «испортить» качество модели. Кроме того с ростом числа дескрипторов (и как следствие увеличением сложности модели) на практике часто наблюдается эффект переобучения – средняя ошибка обучения падает в то время, как ошибка на контрольных данных начинает расти.

1.2.1 Ограничения допустимости

За последние годы накоплено много знаний о моделировании «структура – свойство» и его приложениях. Модели «структура – свойство» используются, в частности, для нормативного регулирования в области химической промышленности. В 2007 году вступило в силу новое европейское законодательство (REACH – European Community Regulation on chemicals and their safe use [8]), по которому для целей исследования свойств химических соединений допускается и поощряется использование моделей «структура – свойство». Результаты прогнозирования на основе реализаций моделей «структура – свойство» используются, когда экспериментальные данные не являются достаточно доступными, или в качестве дополнительной информации, при условии, что прогноз модели достоверен [22].

Рост популярности моделирования «структура – свойство» сопровождается ростом вопросов, касающихся надежности предсказаний соответствующих моделей [23]. Как уже отмечалось выше, вывод модели базируется, прежде всего, на обучающем наборе соединений, который по понятным причинам является структурно ограниченным [24]. По этой причине сама модель является структурно ограниченной, и может быть, в лучшем случае, применена для какой-либо конкретной категории веществ. Надежные прогнозы, как правило, обеспечены только тем соединениям, чьи структуры являются наиболее схожими со структурами веществ, использованных при построении модели [25].

Принцип областей применимости (Applicability Domain) [23] обязывает исследователей определить рамки использования своих моделей, задав, таким образом, определенные ограничения на допустимые для модели структуры с учетом их представления. Если соединения (молекулярные графы) находится за пределами области применимости модели, прогноз его свойств/активности не может рассматриваться, как достоверный.

Многие стратегии к определению областей применимости были предложены в литературе. Далее представлен краткий перечень этих методов.

1) Методы ограничения диапазона действия (Range-based Methods)

a. Габаритный прямоугольник – область применимости определяется как *M*-мерный прямоугольник, заданный максимальными и минимальными значениями дескрипторов, используемых для описания структуры М-графа. В качестве недостатков метода выделяют невозможность идентификации пустых областей внутри области применимости, а также то, что не принимается во внимание информация о корреляции между дескрипторами.

b. Габаритный прямоугольник на главных компонентах – габаритный прямоугольник, построенный на проекции исходного пространства дескрипторов в пространство главных компонент. Решает проблему корреляции между дескрипторами, однако, не позволяет по-прежнему идентифицировать пустые области.

2) Геометрический метод – подход, при котором строится наименьшая выпуклая область, содержащая обучающую выборку. При таком подходе увеличение сложности данных сильно влияет на реализацию метода и увеличивает его сложность [26]. Метод эффективен для небольшого количества измерений пространства дескрипторов.

3) Методы, основанные на расстоянии (Distance-based Methods) – области применимости определяются с помощью задания порога расстоянии между новым М-графом и обучающей выборкой в пространстве дескрипторов. Тем не менее, в литературе не получено никаких строгих правил по выбору величины такого порога. Для задания расстояния часто используются расстояние Махаланобиса, Евклидово расстояние, расстояние городских кварталов и другие.

4) Методы, основанные на плотности вероятностного распределения (Probability Density Distribution based Methods) – области применимости задаются путем оценки значений функции плотности вероятности для задан-

ных данных [27]. Потенциальная функция вычисляется для всех обучающих М-графов, а глобальное распределение получают путем учета всех индивидуальных потенциалов. С учетом прогностической способности модели выбирается пороговое значение для плотности вероятности, с помощью которого и ограничивают применимость модели.

5) Подход «Расстояние до Модели» (Distance to Model Approach) – DM-подход, основан на использовании информации о целевом свойстве [28]. Предложены методы, которые используют стандартное отклонение вектора предсказания для набора моделей, построенных на одних и тех же данных. При этом значительное расхождение в результатах свидетельствует о нбенадежности прогноза. Другие методы оценивают надежность через нечеткий выход модели (чем ближе нечеткий прогноз к четким граница, тем он надежнее).

В рамках представления методов использования областей применимости можно перечислить также работы по одноклассовой классификации [29], метод ближайших соседей [25], деревья решений и случайные леса [30]. В работе [31] предложен также метод, предполагающий построение метаклассификатора, прогнозирующего успешность прогнозов исходной модели. В такой формулировке метод значительно перекликается с предложенной в данной работе двухфазной схемой решения задачи «структура – свойство». Следует, однако, отметить, что в известных работах отсутствуют какие-либо строгие оценки прогностической способности данного подхода. Кроме того, сам подход используется лишь с целью задания некоторой метрики для определения областей допустимости. В отличие от перечисленных работ представленный в настоящей диссертации подход глубоко исследован, дано строгое теоретическое обоснование его применимости, а также рассмотрены перспективы его комплексного использования, как в рамках моделирования «структура – свойство», так и для решения общей задачи классификации.

1.2.2 Виртуальный скрининг

В данной работе под виртуальным скринингом будем иметь в виду поиск потенциально активных соединений в больших базах химических структур. В общем смысле виртуальный скрининг представляет собой вычислительную процедуру, включающую автоматизированный процесс просмотра больших баз химических соединений и отбор тех из них, для которых прогнозируется наличие желаемых свойств.

Виртуальному скринингу посвящен ряд монографий [2, 32], а также обзорных статей [33, 34]. Общая классификация методов виртуального скрининга приведена в **таблице 1.**

Таблица 1. Методы виртуального скрининга

Подходы	Методы	Преимущества	Недостатки
Ligand-based design (методы основанные на структуре известных активных соединений)	Методы 2D структурного подобия активным молекулам ▪ Подструктурный поиск ▪ Методы подобия «отпечатков пальцев» ▪ Биоизостерный анализ	Концептуальная простота Высокая производительность	Тяготение к уже известным химическим типам соединений Невозможность учета молекулярных параметров
	Методы 3D структурного подобия активным молекулам ▪ 3D фармакофорное моделирование ▪ Методы подобия 3D формы молекул	Способность разрабатывать новые структурные хемотипы соединений на основе структур известных молекул	Повышенные требования к вычислительным ресурсам
	Специальные методы анализа данных и QSAR ▪ Статистические методы ▪ Деревья классификации ▪ Искусственные нейронные сети ▪ Метод опорных векторов ▪ Композиции классификаторов ▪ Смеси экспертов	Высокая производительность Возможность удобного визуального представления результатов анализа Возможность моделирования комплексных видов активности, которые не могут быть исследованы на теоретическом уровне (например, токсичность, фармакологические эффекты)	Должно быть известно достаточно большое количество активных соединений Для некоторых методов сложность интерпретации результатов анализа
Target structure-based design (методы основанные на структуре биомишеней)	Методы молекулярного докинга	Метод разработки новых мишень-специфичных соединений, не требующий знания структур активных молекул	Очень высокие требования к вычислительным ресурсам Недостаточная изученность некоторых теоретических аспектов лиганд-рецепторных взаимодействий Белки-мишени рассматриваются как стационарные, а не динамические системы (как это имеет место в реальности)

Виртуальный скрининг практически невозможен без использования правил отказа от прогноза (ограничений допустимости) для моделей «структура – свойство». Кроме того необходимость просеивания огромного количества

соединений из баз рождает требования к вычислительной сложности упомянутых правил отказа. Значение функций допустимости должны вычисляться с минимальными временными затратами. Быстрые правила отказа – важный приоритет при разработке практических систем виртуального скрининга. Идея быстрых правил отказа впервые предложена автором в [35], затем подход получил широкое развитие [21, 36, 37].

Виртуальный скрининг в целом крайне важен при разработке новых лекарственных средств. Он применяется для поиска соединений, обладающих нужными видами свойств и биологической активности.

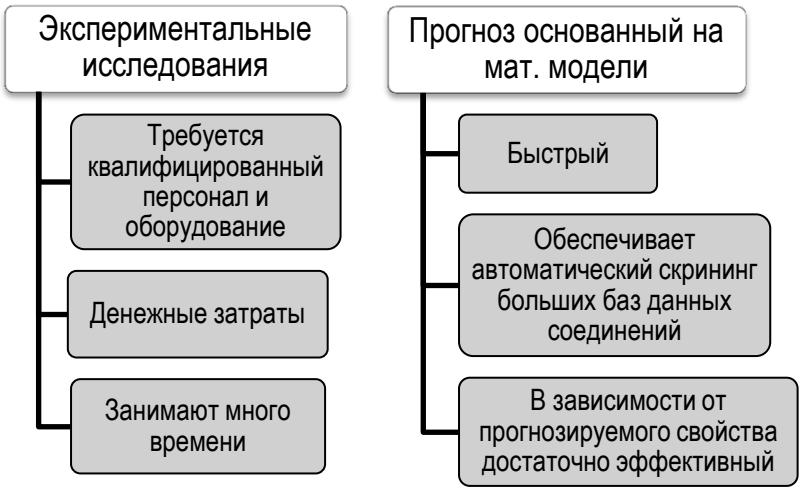


Рисунок 2. Преимущества использования методов моделирования «структура – свойство»

Использование компьютерного моделирования «структура – свойство» в процессе виртуального скрининга позволяет (см. **рисунок 2**):

- рассмотреть огромное количество соединений, затратив относительно небольшое время по сравнению с реальными экспериментами;
- дает возможность исследовать еще не синтезированные соединения, снизив при этом стоимость экспериментов, так как не требует средств на приобретение или синтез химических реактивов.

Ключевыми требованиями к моделям «структура – свойство» при виртуальном скрининге являются адекватность получаемого прогноза при условии ограниченной мощности обучаемой выборки (см. предыдущий раздел), а также – скорость прогнозирования.

1.2.3 Многоуровневое дескрипторное описания

В данном разделе приводится использованное в работе дескрипторное описание М-графов. Данный формат представления информации о структуре М-графов считался фиксированным и не представлял предмета исследований.

Ниже представлен подход к описанию структур М-графов на базе фрагментных дескрипторов особых точек. Подход включает в себя несколько уровней дескрипторного описания с последовательно возрастающей вычислительной сложностью. Данная особенность существенно влияет на выбор описания при построении конкретных моделей «структура – свойство». В качестве особых точек (ОТ) М-графа в общем случае могут выступать особые точки молекулярной поверхности, определяемой физико-химическими свойствами М-графа. В рассматриваемом (простейшем) случае в качестве ОТ выступают атомы (вершины) и цепочки атомов М-графа.

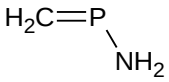
Особой точкой первого порядка молекулярного графа G назовем его вершину V_i. При этом кодом особой точки будет являться метка данной вершины.

В качестве меток атомов (вершин) М-графа в данной работе использовались символ соответствующего химического элемента, а также три маркера: d, b и r. Где d– степень вершины молекулярного графа, b–информация о виде химических связей атома (одинарная, двойная, тройная), r–положения атома в структуре (находится он в кольце или цепи). За счет включения или исклю-

чения данных маркеров из меток вершин М-графов можно сформировать 8 различных типов кодировки ОТ.

При этом количество различных меток вершин М-графа не превосходит: N_E (количество известных химических элементов) × 7 (6 значения маркера d + метка выключения маркера) × 6 (5 значения маркера b + метка выключения маркера) × 4 (3 значения маркера r + метка выключения маркера) = $168 \cdot N_E$. А для фиксированного типа кодировки ОТ это значение не превосходит $90 \cdot N_E$. При этом для оценки сложности описания может быть полезна и более грубая оценка, зависящая от числа М-графов в обучающей выборке – $90 \cdot N \cdot T$. Обозначим число различных меток для выбранного типа кодировки через M_c.

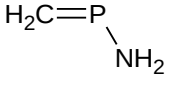
Пример 1.



Рассмотрим вершину P, код соответствующей ей ОТ может выглядеть как P2dc (все маркеры включены) и как P_d_ (включен только второй маркер).

Особой точкой порядка p назовем набор вершин $\{V_{i_1},\ldots V_{i_p}\}$ графа G, образующий путь в графе длины p. Кодом особой точки порядка p является конкатенация меток вершин, входящих в набор: <метка 1-ой вершины>+…+<метка p-ой вершины>.

Пример 2.



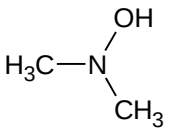
Данный М-граф содержит 2 ОТ второго порядка, им соответствует коды C1dcP2dc и P2dcN1sc.

На практике порядок особых точек при построении описывающего отображения фиксирован и принимает значения из множества $\{2,3,4\}$.

Количество различных кодов ОТ порядка p при этом не превосходит $M_c{}^p$.

Фрагментом первого уровня назовем особую точку М-графа G, кодом фрагмента первого уровня выступает код ОТ. *Дескриптором первого уровня* назовем число повторений фрагментов первого уровня (ОТ) с фиксированным кодом в М-графе G

Пример 3.



Значение дескриптора C1sc равно 2.

Расстояниям d между двумя особыми точками назовем минимальную длину пути в графе G, связывающую вершины первой ОТ с вершинами второй. Данное расстояние будем называть топологическим. Предполагаем, что задано конечное множество интервалов расстояния, имеющих уникальные коды $code(d)$.

Так как молекулярный граф $G \in TG$, то $d \le T$.И в качестве кода расстояния можно использовать его величину. В рассматриваемых дескрипторах использовалась пороговая кодировка расстояния: $code(d) = "1"$, если $d \le d_b$, $code(d) = "2"$, если $d > d_b$, где d_b – заданный порог расстояния, выбираемый на основе значений данного расстояния на обучающей выборке.

В случае, когда метки вершин исходного М-графа содержат трехмерные координаты атомов соответствующей молекулы, в качестве расстояния d можно использовать геометрическое расстояние, вычисленное по координатам атомов.

Структурным фрагментом второго уровня назовем пару ОТ молекулярного графа G, кодом фрагмента при этом выступает конкатенация кодов ОТ и кода расстояния между ними: *<код фрагмента 2 уровня> = <код ОТ1> + <код ОТ2> + <код расстояния>*.

Для управления эквивалентностью фрагментов используются указанные выше маркеры, а также коды расстояния. Два фрагмента называются эквивалентными, если их коды совпадают.

Фрагментом уровня 3 назовем тройку ОТ молекулярного графа G, при этом кодом фрагмента третьего уровня формируется из кода фрагмента второго уровня, соответствующего паре ОТ в данной тройке, кода третьей ОТ и кода расстояния между выделенной ОТ и фрагментом второго уровня. Расстояние между фрагментом и ОТ определяется, как и ранее, по минимальной длине пути, связывающем соответствующие вершины в графе. При этом каждой тройке ОТ соответствуют три фрагмента второго уровня, различающихся выделенной вершиной:

<код фрагмента 3 уровня> = <код ОТ1> + <код фрагмента ОТ2 + ОТ3> + <код расстояния>,

<код фрагмента 3 уровня> = <код ОТ2> + <код фрагмента ОТ1 + ОТ3> + <код расстояния>,

<код фрагмента 3 уровня> = <код ОТ3> + <код фрагмента ОТ1 + ОТ2> + <код расстояния>.

Аналогично определяются фрагменты более высоких уровней.

Для фрагментов 4-го и более высоких уровней к коду фрагмента, описанному выше, добавляется также флаг пространственной ориентации, задающий левосторонние и правосторонние четверки ОТ пространстве.

Дескриптором k-ого уровня назовем число повторений в молекулярном графе G фрагментов k-ого уровня с фиксированным кодом.

Заметим, что число дескрипторов k-ого уровня при этом не меньше, чем произведение количества различных ОТ на количество дескрипторов уровня

$k-1$: $M_{D_k} \geq M_c^{\,p} \cdot M_{D_{k-1}}$. Таким образом, при переходе к каждому следующему уровню дескрипторов, вычислительная сложность описания обучающей выборки увеличивается пропорционально количеству различных меток вершин М-графов в степени p.

При этом, для некоторых свойств М-графов, таких как хиральность (левосторонняя или правосторонняя ориентация М-графа в пространстве) необходимо использовать дескрипторы как минимум 4-го уровня.

1.2.4 Адаптация дескрипторного описания

На практике в любой задаче классификации, регрессии или прогнозирования возникают вопросы: какое признаковое описание использовать, а какое – нет?; необходимо ли как-то преобразовывать полученные вектора признаков?; и другие.

В задаче «структура – свойство» указанные вопросы приобретают особое значение, так как за годы формирования QSAR/QSPR было предложено очень большое число различных молекулярных дескрипторов. Основными типами молекулярных дескрипторов являются фрагментные дескрипторы, топологические индексы, физико-химические дескрипторы, квантово-химические дескрипторы, дескрипторы молекулярных полей, константы заместителей, фармакофорные дескрипторы, дескрипторы молекулярного подобия [38]. Всего на настоящее время в различных источниках можно встретить более 7000 типов молекулярных дескрипторов.

Понятно, что на практике перед исследователем стоит сложная задача подбора дескрипторного описания под конкретную изучаемую задачу (для одного свойства лучше подойдут одни дескрипторы, а для другого – совершенно иные). Кроме того, даже в рамках одного фиксированного подхода число дескрипторов, которыми описывается М-граф, исчисляется тысячами.

Поэтому задача отбора признаков (features selection) – одна из наиболее важных при построении моделей «структура – свойство».

Существенным является также то обстоятельство, что по мере увеличения мощности описания (количества используемых дескрипторов) ошибка прогнозировании на обучающей выборке, как правило, монотонно убывает, в то время как ошибка на контрольных данных сначала тоже убывает, однако с какого-то момента времени начинается её неограниченный рост. Это обстоятельство связано с так называемым эффектом переобучения. Избыточные степени свободы, возникающие при использовании большого числа дескрипторов «расходуются» не только на восстановление исходной зависимости «структура – свойство», но и на те зависимости, которые свойственны лишь данной конкретной обучающей выборке. Переобучение наблюдается при использовании большинства методов машинного обучения и встречается во всех предметных областях. Отбор дескрипторов позволяет находить оптимальную сложность модели (оптимальное количество дескрипторов), при которой переобучение минимально.

У процедуры отбора дескрипторов кроме уже отмеченных имеется ещё ряд важных преимуществ. Во-первых, она позволяет сократить затраты на вычисления ненужных признаков и повышает скорость алгоритмов прогнозирования. Во-вторых, такая процедура приводит к более простым и понятным моделям, которые легче интерпретировать с помощью аппарата хемоинформатики и для проверки конкретных химических гипотез.

Основная же сложность отбора дескриптора в его вычислительной сложности. Как уже было отмечено ранее, в задаче «структура – свойство» рассматриваются описывающие отображения, оперирующие сотнями и тысячами дескрипторов. В то время как перебор в множестве из M признаков потребует рассмотрения $2^M - 1$ вариантов подмножеств.

В качестве основных методов отбора дескрипторов можно перечислить такие как полный перебор (невозможно использовать на практике из-за «комбинаторного взрыва» по сложности), метод последовательного добавления признаков (также используются модификации вроде последовательного удаления признаков или поочерёдного добавления/удаления), анализ главных компонент, кластеризация признаков, генетический алгоритм и другие.

1.3 Постановка задачи построения адаптивных распознающих моделей

Данный раздел содержит основные определения и постановки задач, используемые для формулирования теоретической части работы. Постановки задач учитывают особенности задачи «структура – свойство», описанные в **разделе 1.2** настоящей работы. Предлагаемые решения поставленных задач, а также теоретические оценки их эффективности, содержатся в **Главе 2**.

1.3.1 Определения

Пусть химические соединения представлены своими *М-графами*. *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.). В качестве меток ребер могут использоваться любые характеристики соответствующих связей: кратность, длины, порядки связей, полученные из квантово-химических расчетов, и т.д. Множество всевозможных молекулярных графов обозначим $MG = \{G\}$. Предполагаем, что любому химическому соединению однозначно соответствуют некоторый граф из MG. Учитывая особенности предметной области (связанные с определенными трудностями в описании структур высокомолекулярных органических веществ, таких как белки) [39] далее формально рас-

сматриваются М-графы, с числом вершин, не превосходящих заданной величины T. Такое ограничение с одной стороны обусловлено необходимостью изъять из рассмотрения М-графы, соответствующие высокомолекулярным соединениям (молекулы которых содержат сотни и тысячи атомов), а с другой позволяет более точно оценить вычислительную сложность предлагаемых алгоритмов. Множество М-графов с числом вершин, не превосходящих T, обозначим TG.

Обучающая выборка $LS = \{(G_i, C_i)\}_{i=1}^{N}$, $G_i \in TG, C_i \in \{Cl_1, \ldots Cl_H\}$ – совокупность из *N* молекулярных графов, где *i*-ый М-граф отнесен к C_i – одному из *H* классов $\{Cl_1, Cl_2, \ldots Cl_H\}$ согласно исследуемому свойству.

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

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

Пусть алфавит дескрипторов состоит из *M* элементов. *Вектором признаков* молекулярного графа *G* будем называть вектор $x = (x_1, \ldots x_M) \in \mathbb{R}^M$, где x_i – значение *i*-ого дескриптора, вычисленное для *G*. *Описывающим отображением* $D: TG \to \mathbb{R}^M$ назовём отображение, ставящее в соответствие *M*-графу $G \in TG$ его вектор признаков $x = (x_1, \ldots x_M) \in \mathbb{R}^M$. Пространство $\mathbb{R}^M$ в данном случае будем называть *пространством дескрипторов*. Процесс вычисления значений дескрипторов для множества М-графов назовем *дескрипторным описанием*.

МД-матрицей или матрицей «молекулярный граф – дескриптор» (матрицей признаков) для рассматриваемой обучающей выборки $LS = \{(G_i, C_i)\}_{i=1}^{N}$

будем называть матрицу *X* размера *N* x *M*, в *i*-ой строке которой стоит вектор признаков $x_i = (x_{i1},..,x_{iM})$ *i*-ого молекулярного графа.

1.3.2 Распознающие модели как решение задачи «структура – свойство»

Традиционно решение задачи «структура – свойство» разбивается на два достаточно независимых этапа:

Этап описания обучающей выборки. Кратко, этап описания обучающей выборки состоит в выборе представления информации о структуре М-графа, то есть, набора признаков-дескрипторов. Более подробно, в ходе этапа описания обучающей выборки необходимо решить следующие задачи:

- выбрать и зафиксировать алфавит дескрипторов для данной обучающей выборки;
- построить описывающее отображение $D: TG \to \mathbb{R}^M$;
- для каждого молекулярного графа G_i из обучающей выборки $LS = \{(G_i, C_i)\}_{i=1}^N$ вычислить его вектор признаков $x_i = (x_{i1},..,x_{iM})$.

Результатом работы этапа описания, как правило, принято считать построенную по обучающей выборке матрицу «молекулярный граф – дескриптор» *MD*.

Этап поиска функциональной зависимости (этап анализа МД-матрицы), также его называют этапом построения распознающей модели. В ходе данного этапа может решаться большое число дополнительных подзадач. К их числу относятся: кластерный анализ обучающей выборки; поиск выбросов в обучающей выборке; отбор дескрипторов для прогнозирования; различные преобразования и разложения матрицы «молекулярный граф – дескриптор» с целью оптимизации описания обучающей выборки, и другие. Однако ключевым для данного этапа является решение задачи поиска функциональной зависимости f между значениями признаков и значением ис-

следуемого свойства и построение распознающей модели, осуществляющей классификацию рассматриваемых М-графов.

Рассмотрим реализацию данного этапа более формально.

Назовём *распознающей моделью RM* совокупность решающих правил, полученную на обучающей выборке и обладающую перечисленными далее свойствами.

- Для молекулярного графа $G \in TG$ и его описания в виде вектора признаков $x = (x_1, \ldots x_M) \in \mathbb{R}^M$ с помощью фиксированного описывающего отображения $D : TG \to \mathbb{R}^M$, распознающая модель *RM* либо осуществляет прогноз его свойства (отнесение М-графа к одному из *H* классов $\{Cl_1, Cl_2, \ldots Cl_H\}$), либо производит отказ от прогноза.
- Для распознающей модели может быть вычислен показатель качества $\phi(RM)$, характеризующий её качество на обучающей выборке.

Замечание. В настоящей работе для оценки качества модели используется процент верно классифицированных М-графов в процессе выполнения процедуры скользящего контроля. Соответствующий показатель качества определен ниже.

Определенные таким образом распознающие модели будем называть также *моделями «структура – свойство»*.

Функцию $f : TG \to \{Cl_1, Cl_2, \ldots Cl_H\}$ назовем *классифицирующей функцией*. Классифицирующая функция f *порождает* распознающую модель RM, если $RM(G) = f(G), \quad \forall G \in TG$. В случае, если не заданы ограничения допустимости, распознающая модель определяется своей классифицирующей функцией.

1.3.3 Адаптивные описывающие отображения

Предполагаем, что в ходе построения распознающей модели по данной обучающей выборке $LS = \{(G_i, C_i)\}_{i=1}^{N}$ с зафиксированным описывающим отображением $D : TG \to \mathbb{R}^M$ могли быть произведены некоторые преобразования матрицы «молекулярный граф – дескриптор» *MD*. В результате этих преобразований распознающая модель строится с использованием только некоторых дескрипторов, полученных описывающим отображением *D*, или на преобразованных дескрипторах МД-матрицы *MD*. Таким образом хотим учесть возможность различного рода отбора значимых дескрипторов, а также использование различных разложений и преобразований матрицы *MD* на этапе поиска функциональной зависимости в задаче «структура – свойство».

Сформулируем общее определение.

Пусть для обучающей выборки $LS = \{(G_i, C_i)\}_{i=1}^{N}$ получено описание в виде $G_i \to x_i = (x_{i1}, .., x_{iM}), \quad i = \{1, \ldots N\}$ с помощью описывающего отображения $D : TG \to \mathbb{R}^M$. Пусть в ходе построения распознающей модели описание обучающей выборки было преобразовано и задано преобразование из старого пространство дескрипторов $\mathbb{R}^M$ размерности M в новое пространство дескрипторов $\mathbb{R}^{M_A}$ размерности M_A: $A : \mathbb{R}^M \to \mathbb{R}^{M_A}$. Тогда определим *адаптивное описывающее отображение* обучающей выборки $AD : TG \to \mathbb{R}^{M_A}$, положив для произвольного молекулярного графа *G* по определению $AD(G) = A(D(G))$. Назовем также *адаптивной матрицей «молекулярный граф – дескриптор» МД*-матрицу MD_A, соответствующую адаптивному описанию *AD*, а пространство $\mathbb{R}^{M_A}$ – *адаптивным пространством дескрипторов*.

1.3.4 Ограничения допустимости и локальные классифицирующие функции

Ограничения допустимости для распознающих моделей служат для определения области допустимых значений для классифицирующей функции. Таким образом, задаются правила отказа от прогноза для недопустимых для прогнозирующей модели М-графов. Задание значений классифицирующей функции на области допустимых молекулярных графов будем называть локальным. Если классифицирующая функция задана на всем множестве М-графов *MG*, будем говорить о глобальном определении. Более формальные термины приведены ниже.

Ограничением допустимости или правилом отказа для распознающей модели *RM* в задаче классификации «структура – свойство» назовём некоторую функцию $g : TG \to \{0,1\}$ со следующей интерпретацией: $g(G) = 1$ будет означать отказ от прогноза свойства данного молекулярного графа, в противном случае прогноз может быть осуществлён. Заметим, что правило отказа g может быть определено на пространстве дескрипторов $\mathbb{R}^M$ или на адаптивном пространстве дескрипторов $\mathbb{R}^{M_A}$ с условием, что $g(G) := g(D(G))$ и $g(G) := g(AD(G))$, соответственно.

Пусть теперь в ходе построения распознающей модели *RM* принято правило отказа g. Назовём молекулярный *граф* $G \in TG$ *допустимым* для распознающей модели *RM*, если согласно принятому ограничению допустимости g прогноз его свойств не может быть осуществлен с помощью данной распознающей модели, то есть $g(G) = 0$.

Пусть зафиксирована распознающая модель *RM* с ограничением допустимости g. Тогда множество $SG = \{G \in TG \mid g(G) = 0\}$ назовем *множеством допустимых молекулярных графов* для модели *RM*.

Пусть $F = \{f\}$ – некоторый класс функций $f : SG \to \{Cl_1, Cl_2, \ldots Cl_H\}$, где $SG \subset TG$ – множество допустимых для распознающей модели *RM* графов, в котором ищется классифицирующая функция.

Назовём *локальной классифицирующей функцией* функцию $f : SG \to \{Cl_1, Cl_2, \ldots Cl_H\}$, получающую в качестве аргумента допустимый для модели *RM* молекулярный граф *G*, и относящую *M*-граф к одному из классов $C \in \{Cl_1, Cl_2, \ldots Cl_H\}$. Как и прежде допускаем, что в качестве аргумента *f* может быть указан вектор в пространстве дескрипторов $\mathbb{R}^M$ или в адаптивном пространстве дескрипторов $\mathbb{R}^{M_A}$ с условием, что по определению $f(G) \coloneqq f(D(G))$ и $f(G) \coloneqq f(AD(G))$, соответственно.

1.3.5 Качество распознающих моделей

Для различных распознающих моделей, использующих разные классы локальных классифицирующих функций, применяют различные показатели качества. В данной работе остановим свой выбор на наиболее общих и универсальных, в то же время являющихся одними из самых прозрачными для понимания.

Пусть зафиксировано множество допустимых для модели *RM* молекулярных графов *SG*, тогда назовём *допустимым обучающим подмножеством* множество $LSG = \{(G_j, C_j) \mid G_j \in LS \cap SG\}_{j=1}^{\tilde{N}}$, которое содержит $\tilde{N}$ молекулярных графов обучающей выборки $LS = \{(G_i, C_i)\}_{i=1}^{N}$, $|LSG| = \tilde{N} \le N$.

Простым показателем качества $\varphi_s(RM)$ для распознающей модели *RM* с локальной классифицирующей функцией $f : SG \to \{Cl_1, Cl_2, \ldots Cl_H\}$, определенной на *допустимом обучающем подмножестве* $LSG = \{(G_j, C_j) \mid G_j \in LS \cap SG\}_{j=1}^{\tilde{N}}$, будем называть процент верно классифицированных данной функцией допустимых молекулярных графов из данной обучающей выборки:

$$\varphi_s(RM) = 1 - \frac{\sum_{G_i \in LSG} \varepsilon_i}{\tilde{N}}, \text{ где } \quad \varepsilon_i = \begin{cases} 0, \ f(G_i) = C_i, G_i \in SG \\ 1, \text{ в противном случае.} \end{cases}$$

Теперь пусть, зафиксирован алгоритм *A* построения локальной классифицирующей функции по произвольному подмножеству обучающей выборки в рамках построения распознающей модели.

Зафиксируем один из элементов *LSG* в качестве контрольного множества $LSG_T = (G_j, C_j)$, тогда текущее обучающее множество образуют элементы из исходной выборки LS без контрольного, а именно $LS_L = LS \setminus LSG_T = LS \setminus (G_j, C_j)$. Пусть с помощью алгоритма *A* на основе LS_L построена классифицирующая функция $f_A^{LS_L} \in F$, которая предсказывает для контрольного М-графа LSG_T класс $f_A^{LS_L}(LSG_T)$. Построив аналогичным образом предсказания для каждого М-графа множества *LSG*, определим показатель качества со *скользящим контролем (leave-one-out cross validation)* [10] как процент верно классифицированных таким образом М-графов множества *LSG*:

$$\varphi_{CV}(RM) = \varphi_{CV}(RM, LSG) = 1 - \frac{\sum_{G_i \in LSG} \varepsilon_i}{\tilde{N}}, \text{ где } \quad \varepsilon_i = \begin{cases} 0, & \text{если } f_A^{LS \setminus \{G_j\}}(G_j) = C_j; \\ 1, & \text{в противном случае.} \end{cases}$$

По умолчанию далее в тексте под показателем качестве подразумевается показатель качества со скользящим контролем.

1.3.6 Постановки задач

С учетом данных выше определений в настоящем разделе дадим постановку задачи построения распознающей модели и постановку задачи построения эффективных правил отказа, а также сформулируем задачу выбора оптимального представления структур М-графов в задаче «структура – свойство».

Задача построения распознающей модели

Задачу построения распознающей модели для решения задачи «структура – свойство» можно сформулировать следующим образом:

Пусть задана обучающая выборка из N молекулярных графов $LS = \{(G_i, C_i)\}_{i=1}^{N}$, для которой рассматривается фиксированное описание, то есть зафиксировано описывающее отображение $D : TG \to \mathbb{R}^M$. *Задача состоит в том, чтобы построить адаптивное описывающее отображение* $AD : TG \to \mathbb{R}^{M_A}$, *ограничение допустимости* $g : TG \to \{0,1\}$, а также *локальную классифицирующую функцию* $f : SG \to \{Cl_1, Cl_2, \ldots Cl_H\}$ для множества допустимых молекулярных графов *SG*, определенного ограничением g, так, чтобы по возможности увеличить значения показателя качества $\varphi_{CV}(RM, LSG)$ на допустимом обучающем подмножестве *LSG*.

Задача построения эффективных правил отказа

Рассмотрим обучающую выборку $LS = \{(G_i, C_i)\}_{i=1}^{N}$. Пусть задана распознающая модель *RM* и зафиксирован алгоритм *A* построения локальной классифицирующей функции по произвольному подмножеству обучающей выборки в рамках построения распознающей модели *RM*. Пусть также с помощью правила отказа $g : TG \to \{0,1\}$ определено допустимое обучающее подмножество $LSG \subset LS$. Тогда правило отказа g назовем *эффективным*, если для него выполнено неравенство $\varphi(RM, LSG) > \varphi(RM, LS)$.

Задача выбора оптимального представления

Пусть зафиксирован конкретный тип описывающих отображений. Предполагается, что в рамках этого типа конкретные отображения различаются своей вычислительной сложностью. При этом различные отображения можно упорядочить по их сложности от самых простых до самых сложных. Пусть заданы описывающие отображения $D_1, D_2, \ldots, D_d$.

Обозначим через $CD_i = CD(D_i)$ – вычислительную сложность отображения D_i (время выполнения вычислений на М-графе с фиксированным числом вершин, выраженное в количестве элементарных операций). Пусть $CD_1 < CD_2 < \ldots < CD_d$.

Задача заключается в разработке алгоритма выбора описания, который обеспечивает снижение сложности вычислений при допустимой потере качества моделей.

Отметим здесь, что постановка задачи построения распознающей модели рассчитана на комплексное изучение обучающей выборки. Она предполагает, что традиционная задача классификации будет решаться множеством различных методов. В конечном итоге хотелось бы показать, как для решения задачи «структура – свойство» использовать семейства и множества распознающих моделей и получать с их помощью согласованный прогноз исследуемого свойства для новых молекулярных графов. Кроме того постановка задачи построения распознающей модели позволяют учитывать:

- возможные вспомогательные преобразования дескрипторного описания обучающей выборки, такие как отбор значимых дескрипторов, различные разложения и преобразования матрицы «молекулярный граф – дескриптор» на этапе поиска функциональной зависимости;
- возможное использование такой информации об обучающей выборке, как наличие или отсутствие в ней выбросов, степень однородности выборки, результаты кластерного анализа, анализа эффективности отдельных алгоритмов и методов для предсказания исследуемого свойства и тому подобных;
- возможность локального построения классифицирующей функции, определенной лишь на некотором подмножестве *M*-графов из обучающего множества;

- возможность отказа от прогнозирования.

1.4 Прогнозирование свойств М-графов методами машинного обучения

Данный раздел посвящен этапу поиска функциональной зависимости. Будем предполагать, что для рассматриваемой обучающей выборки молекулярных графов уже построено описание с помощью молекулярных дескрипторов и сформирована МД-матрица, содержащая по строкам вектора признаков М-графов из обучающей выборки. Для М-графов обучающей выборки считаем заданным вектор свойств (размерности равной числу строк МД-матрицы или количеству М-графов из обучающей выборки), содержащий значение исследуемого свойства для каждого М-графа обучающей выборки соответственно. Отметим, однако, что некоторые алгоритмы машинного обучения (например, SVM на базе ядер [40]) могут и не использовать дескрипторное описание обучающей выборки в виде МД-матрицы. В простой постановке (когда задано векторное описание объектов обучающей выборки и каждому объекту поставлен в соответствие определенный класс) задача «структура – свойство» сводится к общей задаче классификации.

Отмеченные выше соображения означают, что для решения задачи «структура – свойство» на этапе поиска функциональной зависимости с некоторыми поправками может быть использован любой метод машинного обучения. Однако наиболее часто можно видеть работы с применением следующих методов: байесовский классификатор, линейный дискриминантный анализ, искусственные нейронные сети, метод опорных векторов, деревья принятий решений, метод *k* ближайших соседей.

Для задачи «структура – активность» (регрессионной задачи): множественная линейная регрессия, искусственные нейронные сети, метод k ближайших соседей, деревья принятия решений.

Для решения одноклассовой задачи классификации: нейронные сети, одноклассовая машина опорных векторов (1-SVM).

Наиболее полное впечатление о многообразии методов машинного обучения можно составить, ознакомившись с курсом лекций К.В. Воронцова [41].

Далее приведем краткое описание двух методов, практическая реализация которых использовалась при проведении тестирования предложенного в работе подхода (см. **Главу 3**).

1.4.1 Линейная регрессия

Статистические методы начали применяться в 1930-х годах одними из первых на заре развития классификации по прецедентам, и были связаны с байесовской теорией принятия решений (работы Неймана, Пирсона [42]).

Простейшим примером решения задачи с вещественными значениями целевого вектора, является линейная регрессия [12]. Искомая функциональная зависимость ищется в виде линейной функции $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_M x_M$, наилучшим образом приближающей экспериментальную кривую. Для этого используется классический метод наименьших квадратов. Метод в ходе своей работы минимизирует сумму квадратов отклонений реально наблюдаемых y от их оценок $\hat{y}$ (под оценками понимаются оценки, полученные приближением прямой линией, выступающей в качестве представления искомой регрессионной зависимости): $\sum_{k=1}^{N} (y_k - \hat{y}_k)^2 \to min$.

1.4.2 Метод опорных векторов

Предложенный Владимиром Вапником в 1998 году метод опорных векторов (Support Vector Machine) [40] стал один из ключевых методов машинного обучения, показавший высокую эффективность в различных задачах классификации. Метод является практическим выходом теории Вапника-Червоненкиса [43], в рамках которой в частности Вапником теоретически до-

казана минимальность оценки функции общего риска для метода опорных векторов.

Вапник в работе [40] предложил бинарный классификатор для задачи классификации из N обучающих векторов $\{(x_1, y_1),(x_2, y_2),\ldots(x_N, y_N)\}$, где $y_i \in \{-1,+1\}$, $i = 1,\ldots N$. Если прецеденты данной обучающей выборки могут быть разделены в пространстве признаков линейной гиперплоскостью $w \cdot x + b = 0$, то такой случай называется случаем линейно разделимых классов. При этом очевидно существует бесконечное множество возможных разделяющих гиперплоскостей.

Из интуитивных соображений представляется, что лучшей будет та разделяющая плоскость, которая максимально отстоит от прецедентов обоих классов. В действительности первоначально данный принцип классификации и возник из следующих эвристических соображений: вполне естественно полагать, что максимизация зазора (margin) между классами должна способствовать более точной классификации. В дальнейшем же этот принцип получил мощное теоретическое обоснование [44, 45].

К несомненным преимуществам метода опорных векторов относится то обстоятельство, что с его помощью находится глобальный минимум в отличие от нейросетевых алгоритмов распознавания. Кроме того, для использования SVM на практике не обязательно знать исходное признаковое представление объектов обучающей выборки. Достаточно уметь вычислять «похожесть» прецедентов (значения соответствующих скалярных произведений).

Для случая линейно неразделимой выборки предлагается переход от исходного пространства признаковых описаний объектов X к новому пространству H с помощью некоторого преобразования $\psi : X \to \mathrm{H}$. Пространство H называют спрямляющим. В случае, когда пространство H имеет достаточно высокую размерность, можно рассчитывать на то, что в нём обучающая вы-

борка окажется линейно разделимой. Можно показать, что если выборка X не содержит противоречий, то всегда найдётся пространство размерности не большей N (размера выборки), в котором обучающая выборка будет линейно разделима.

Так как задача, которую решает метод опорных векторов и сам алгоритм классификации не зависят от признаковых описаний объектов, а зависят лишь от их скалярных произведений, то формально вместо скалярных произведений $\langle x, x' \rangle$ всюду в ходе применения метода можно использовать ядра $K(x, x')$. Функция $K : X \times X \to \mathbb{R}$ называется ядром (kernel function), если она представима в виде $K(x, x') = \langle \psi(x), \psi(x') \rangle$ при некотором отображении $\psi : X \to \mathrm{H}$, где H — гильбертово пространство.

В заключении отметим, что указанный подход может вообще не использовать признаковые описания объектов. Не так давно стали популярны также так называемые беспризнаковые методы классификации, которые используют только информацию о попарной схожести объектов. Из чего, однако, следует, что данная информация должна обладать свойствами некоторого скалярного произведения, симметричностью и неотрицательной определенностью).

1.5 Выводы

В данной главе приведена постановка задачи «структура – свойство», а также задач, решение которых направлено на построения распознающих моделей и ограничений допустимости для них. В **разделе 1.2** продемонстрированы существенные особенности, выделяющие задачу «структура – свойство» из числа абстрактных задач классификации. Таким образом, показана необходимость разработки подхода, позволяющего строить модели «структура – свойство», эффективные для практических задач прогнозирования свойств М-графов (таких как виртуальный скрининг).

Глава 2. Методы решения

В настоящей главе представлены различные подходы к решению задачи построения распознающих моделей, поставленной в **разделе 1.3.** Приводятся теоретические результаты. Описаны методы адаптации дескрипторного описания. Описан классификатор на базе нечеткой кластерной структуры обучающей выборки. Дается двухфазная схема решения задачи «структура – свойство». Доказана оценка качества результирующей модели при использовании двухфазной схемы. Предлагается решение задачи «структура – свойство» на базе семейств и множеств распознающих моделей. Описан метод снижения вычислительной сложности дескрипторного описания. Обсуждаются также перспективы предложенного подхода.

2.1 Общая методология прогнозирования

В данном разделе представлена общая методология прогнозирования свойств новых, неизученных М-графов на базе семейств и множеств распознающих моделей с ограничениями допустимости. В последующих разделах уточняются подходы к построению ограничений допустимости и построению классификаторов.

Постановка задачи «структура – свойство» с точки зрения химии предполагает формирования какого-либо обучающего набора (выборки) соединений, для которых значение свойства получено экспериментально. При формировании обучающей выборки необходимо учитывать цели исследования, особенности химических классов соединений, гипотезы о влиянии отдельных структурных элементов и физико-химических свойств на целевое свойство М-графов и другую актуальную с точки зрения химии информацию.

Теперь стоит задача выбора дескрипторного описания, которое будет использоваться при обучении распознающих моделей, а также методов самого машинного обучения.

Существенным является тот факт, что в ходе анализа исследователем выборки М-графов, так или иначе, используются различные описания, преимущества которых друг перед другом в данном конкретном случае заранее установить невозможно. Помимо этого приходится сравнивать между собой различные алгоритмы обучения моделей, наборы параметров и учитывать вспомогательные методы анализа. К их числу относятся: кластерный анализ обучающей выборки; поиск выбросов в обучающей выборке; отбор дескрипторов для прогнозирования; различные преобразования и разложения матрицы «молекулярный граф – дескриптор» с целью оптимизации описания обучающей выборки и другие.

Далее представим некоторое универсальное решение, позволяющее учесть описанные выше особенности.

Постановка задачи построения распознающей модели из **главы 1** предоставляет возможность использовать различные дескрипторные описания обучающей выборки, а также адаптировать эти описания под конкретное целевое свойство. В результате можно одновременно рассматривать несколько описывающих отображений. Для прогнозирования свойств новых М-графов предлагается использовать множества или семейства (параметрические) моделей, построенных на различных описаниях и с использованием различных методов обучения.

Пусть анализируется обучающая выборка из N молекулярных графов $LS = \{(G_i, C_i)\}_{i=1}^{N}$, для которой рассматривается несколько описаний, то есть, зафиксированы d описывающих отображений $D_1, D_2, \ldots, D_d$. Каждому такому описанию соответствует своя матрица «молекулярный граф – дескриптор» $MD_1, MD_2, \ldots MD_d$.

С помощью фиксированного метода адаптации и набора алгоритмов обучения, можно построить K распознающих моделей $RM_1, RM_2, \ldots RM_K$, где K

является произведением d на количество используемых алгоритмов обучения. Обозначим рассматриваемое множество моделей через *SRM*. Каждой модели $RM_i \in SRM$ соответствует свое адаптивное описывающее отображение AD_i, а также с помощью ограничений допустимости g_i задано множество допустимых М-графов SG_i, для которых и только для них, она осуществляет прогноз. Помимо этого известно значение показателя качества на обучающей выборке $\varphi(RM_i)$.

Теперь рассмотрим прогнозирование свойств нового М-графа $\hat{x}$. Назовем распознающую модель RM_i допустимой для М-графа $\hat{x}$, если $g_i(\hat{x}) = 0$. Таким образом, графу $\hat{x}$ можно поставить в соответствие набор допустимых моделей $SRM_x = \{RM_i \in SRM \mid g_i(\hat{x}) = 0\}$.

Далее, когда определено множество допустимых для молекулярного графа $\hat{x}$ моделей, следует осуществить согласованное прогнозирование целевого свойства. Схематично процесс прогноза показан на **рисунке 3**.

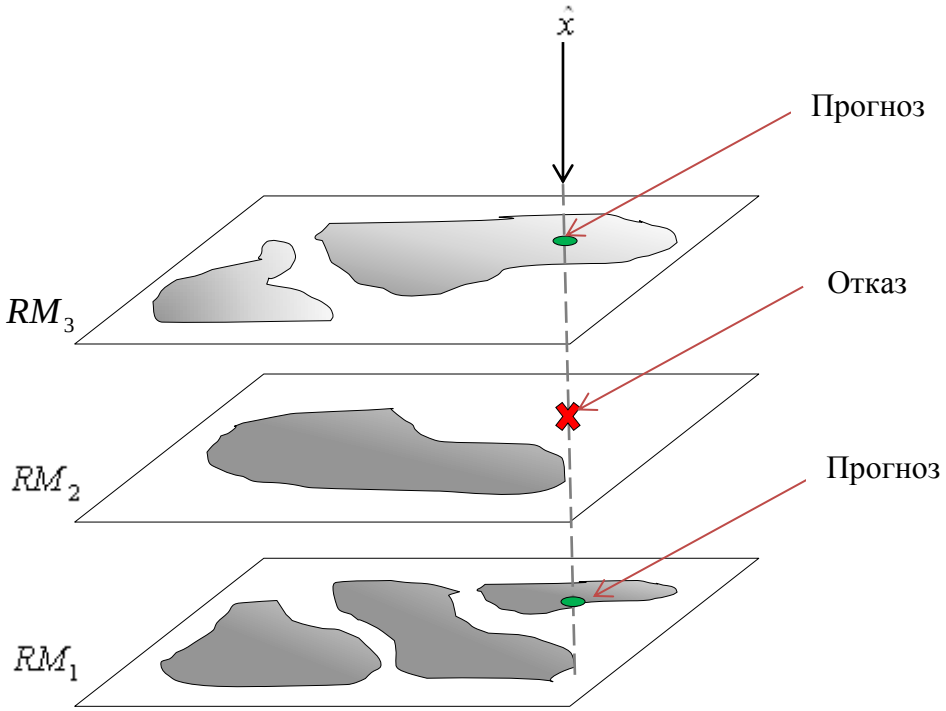


Рисунок 3. Прогнозирование свойств нового М-графа по множеству распознающих моделей

В качестве методов согласованного прогноза приведем наиболее простые, перечисленные далее.

Метод голосования, в соответствии с которым результирующая оценка значения свойства получается как покомпонентная сумма векторов прогноза свойства всех допустимых моделей. Если значение компоненты результирующего вектора положительно (то есть большинство моделей предсказывает наличие свойства у данного М-графа), то данный М-граф считается обладающим свойством, иначе – не обладающим.

Метод взвешенного голосования, когда в качестве результирующей оценки свойства выступает сумма прогнозов моделей с коэффициентами, равными значению показателей качества данных моделей.

Метод положительных оценок, в основе которого те же принципы, что и при взвешенном голосовании, однако суммируются исключительно поло-

жительные оценки свойства (другая интерпретация – вместо значения «-1» для графов, не обладающих свойством, используется значение «0»).

Голосование сильнейших состоит в том, что в голосовании принимают участие не все модели, а только несколько лучших с точки зрения оценки качества. Эта модификация может быть применена к любому из перечисленных выше методов.

Метод победителя, когда результирующий прогноз осуществляется моделью с наивысшим значением показателем качества.

Вероятностная оценка, метод, который заключается в том, что итоговым ответом служит отношение суммы или взвешенной суммы (в случае, когда такая сумма положительна) прогнозов к теоретическому максимуму такого значения.

Отметим, что на практике, как правило, результаты согласованного прогнозирования, которое получается с помощью перечисленных методов, отличаются не принципиально. По этой причине выбор метода согласованного прогнозирования не относится к числу принципиально важных задач. Однако сравнение этих методов может выступать в качестве вспомогательных целей практических исследований.

2.2 Эволюционный метод адаптации дескрипторного описания

Как уже отмечалось ранее, в качестве методов адаптации описывающих отображений могут быть использованы различные известные схемы. Наибольший интерес среди них представляют статистические оценки информативности дескрипторов [46], метод главных компонент, методы последовательного добавления/удаления дескрипторов. Кроме того, отдельного упоминания заслуживает выбор параметров описывающих отображений по результатам тестового прогнозирования [47, 48]. В случае используемых нами структурных дескрипторов [9] в качестве таких параметров выступают

интервалы разбиения межфрагментных расстояний, а также выбор способа маркировки фрагментов.

При проведении практических исследований в рамках данной работы, описанных в **главе 3**, в качестве основного метода адаптации дескрипторного описания (метода отбора признаков) использовался эволюционный отбор дескрипторов [49]. Для оптимизации такого отбора, в плане уменьшения времени проводимых расчетов, из исходных матриц «молекулярный граф – дескриптор» удалялись малоинформативные столбцы-дескрипторы (значительная часть элементов которых являлась нулевыми). Далее опишем процедуру адаптации описывающего отображения, в ходе которой по данной матрице «молекулярный граф – дескриптор» MD строится адаптивная МД-матрица MD_A.

Метод эволюционного отбора дескрипторов

В ходе эволюционного отбора дескрипторов для каждой МД-матрицы $MD_1, MD_2, \ldots MD_d$ производится следующая селективная процедура, схематично представленная на **рисунке 4**.

Рассматривается матрица MD. Её столбцы обозначим $d_1, \ldots d_M$, таким образом, для описания структуры М-графов из обучающей выборки использовались M дескрипторов. Назовем селекцией S произвольный набор столбцов матрицы MD, $S = \{d_{i_1}, \ldots d_{i_m}\}$. Далее считаем, что алгоритм построения распознающей модели по произвольной МД-матрице фиксирован. Будем обозначать через $RM(S)$ распознающую модель, построенную по МД-матрице, состоящей из столбцов, входящих в селекцию S. В качестве параметров алгоритма отбора дескрипторов выступают количество отбираемых на каждом шаге селекций sel, а также ограничение на общее количество итераций алгоритма $nevol$. Шаги алгоритма иногда называют эволюциями.

Первый шаг метода эволюционного отбора дескрипторов состоит в следующем. Формируются селекции, состоящие только из одного столбца-дескриптора исходной МД-матрицы MD:

$$S_1 = \{d_1\}, S_2 = \{d_2\}, \ldots S_M = \{d_M\}.$$

Затем по МД-матрицам, заданным каждой селекцией, строятся распознающие модели и вычисляются значения показателей качества для каждой из них:

$$r_i = \varphi(RM(S_i)), \quad i = 1, \ldots M.$$

Для следующего шага отбираются sel лучших в смысле показателя качества построенных моделей селекций $BS_1, BS_2, \ldots BS_{sel}$ (эти селекции условно будем называть лучшими). На последующих шагах процедура формирования набора селекций заключается в том, что для каждой лучшей селекции BS_i формируются M селекций путем присоединения к набору столбцов, входящих в BS_i каждого из столбцов исходной матрицы:

$$S_{ij} = \{BS_i, d_j\}, \quad i = 1, \ldots sel, \quad j = 1, \ldots M.$$

Процедура отбора лучших селекций проходит аналогично первому шагу. Алгоритм прекращает работу, когда добавление новых столбцов перестает улучшать качество рассматриваемых моделей, или при достижении допустимого количества итераций.

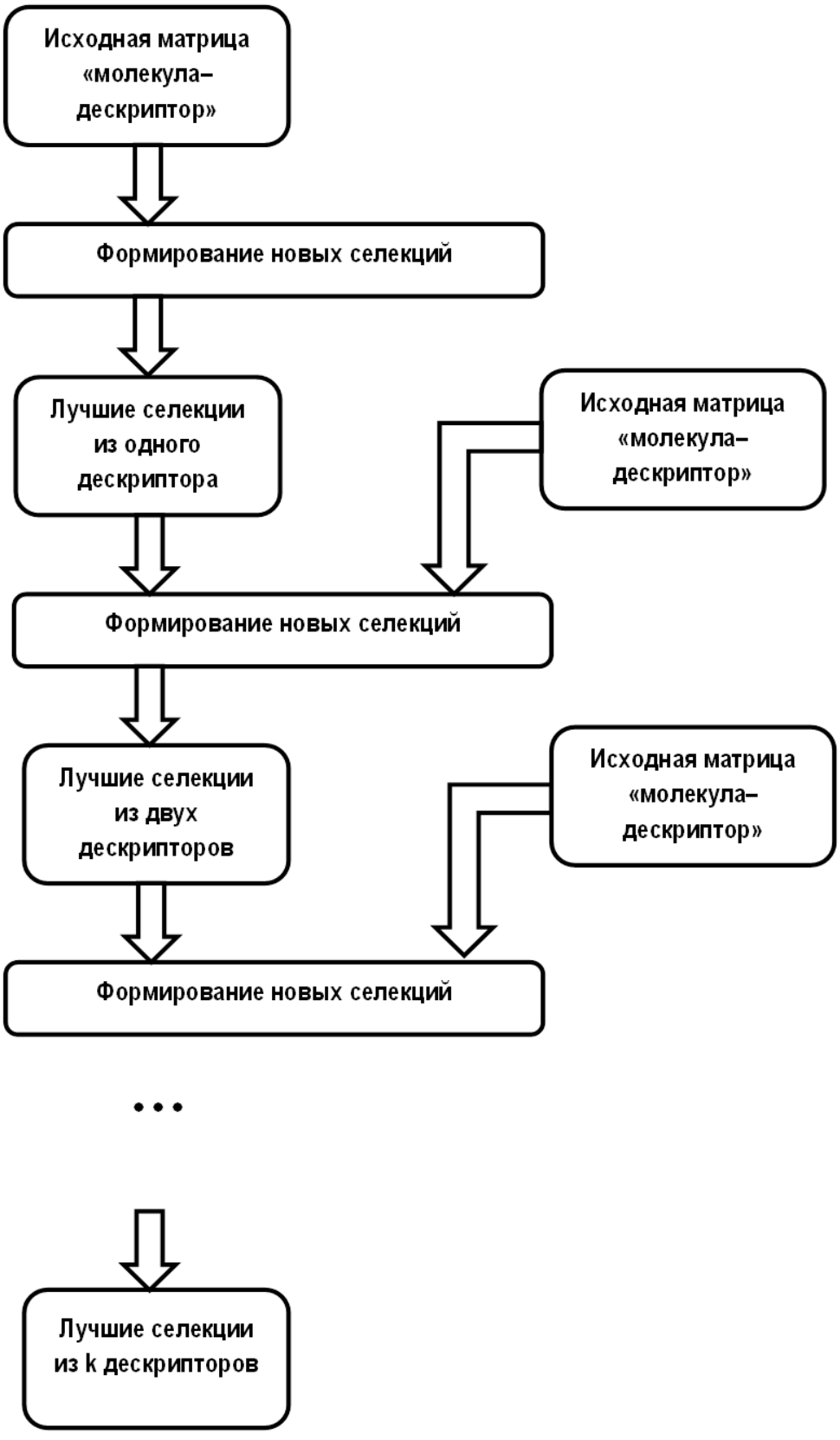


Рисунок 4. Схема алгоритма эволюционного отбора дескрипторов

Описанная процедура адаптации описывающих отображений реализует «обратную связь» этапа поиска функциональной зависимости с этапом описания структур М-графов. Это обстоятельство означает, что результаты тестирования построенных моделей влияют на выбор итогового описывающего отображения.

2.3 Модели «структура – свойства» на базе кластерной структуры обучающей выборки

Ниже описано построение моделей «структура – свойство» и ограничений допустимости для них на основе кластерной структуры обучающей выборки. Идея состоит в том, чтобы использовать кластерную структуру исходной обучающей выборки для построения правил отказа от прогноза для данного молекулярного графа и для построения локальных моделей «структура – свойство», определенных на подмножествах обучающей выборки. В **подразделе 2.3.2** представлен нечеткий классификатор на базе кластерной структуры обучающей выборки.

2.3.1 Ограничения допустимости на базе кластерной структуры обучающей выборки

Так как важным требованием к ограничениям допустимости является необходимость определять допустимость молекулярного графа с минимальными вычислительными затратами, предлагается вычислять правила отказа на специальном пространстве дескрипторов, гораздо меньшей размерности, чем исходное, например, только топологических [50].

Таким образом, строится два пространства дескрипторов, одно – для построения правил отказа, другое – для классификации и прогноза свойств. Здесь естественным образом возникает *сокращённая (специальная) матрица «молекулярный граф – дескриптор»,* строки которой являются векторами в специальном пространстве дескрипторов. Итак, пусть имеется сокращённая матрица «молекулярный граф – дескриптор» X и интерпретация её строк

$x_i = (x_{i1}, \ldots x_{iM})$ как векторов пространства $\mathbb{R}^M$. Пусть также в пространстве $\mathbb{R}^M$ зафиксирована метрика ρ. Для определения числа кластеров предлагается использовать алгоритм построения минимального покрывающего дерева [51].

Пусть для сокращённой МД-матрицы вычислены расстояния между любыми её двумя строками как точками в пространстве $\mathbb{R}^M$ в метрике ρ. Выберем точку и начнем наращивать «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*.Далее 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 [52]. Отличие заключено в том, что центром кластера считается ядро, состоящее из q точек обучающей выборки, а не одна точка пространства $\mathbb{R}^M$. 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*:

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

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

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

Алгоритм заключается в следующем:

1. 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

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

2. 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.

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

Сформулируем правила отказа. Пусть $\hat{x}$ – вектор признаков молекулярного графа *G*, свойства которого необходимо предсказать. В этом случае:

- $g_1(\hat{x}) = (\min_{i \in \{1,\dots N\}} \rho(\hat{x}, x_i) > R)$, то есть если $\min_{i \in \{1,\dots N\}} \rho(\hat{x}, x_i) > R$, где R – пороговое значение, то отказываемся от прогноза (новый М-граф посторонний для всей выборки).
- $g_2(\hat{x}) = (\rho(\hat{x}, Z_i) > r_i)$, то есть если $\rho(\hat{x}, Z_i) > r_i$ – отказ от прогноза в *i*-ом кластере (М-граф посторонний для данного кластера).

Окончательно имеем:

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

Таким образом, М-графы, заведомо отличные от используемых при построении модели, не будут с помощью неё предсказываться. Это обстоятельство способно улучшить качество прогноза. Однако, утверждать, что в данном случае, что для построенных ограничений допустимости выполнено условие эффективности $\varphi(LSG) > \varphi(LS)$ нельзя.

2.3.2 Нечеткий классификатор на базе кластерной структуры обучающей выборки

Для прогнозирования свойств М-графов на базе кластерной структуры обучающей выборки автором предложен следующий подход, представлен-

ный автором в [53]. Пусть $x_i = (x_{i1},\dots x_{iM})$ – описание *i*-го М-графа, y_i – значение его свойства. Тогда имеем матрицу $X = [x_{ij}]$, именуемую в соответствии с принятой ранее терминологией МД-матрицей или матрицей «молекулярный граф – дескриптор», а также вектор значений свойства $(y_1,\dots y_N)$.

Итак, дано описание обучающей выборки в виде МД-матрицы *X*, а также интерпретация строк данной матрицы $x_i = (x_{i1},\dots x_{iM})$ как векторов пространства $\mathbb{R}^M$. Пусть также в пространстве $\mathbb{R}^M$ зафиксирована метрика ρ. Для построения нечеткой кластерной структуры применяем некоторый алгоритм нечёткой кластеризации (*c-means fuzzy* или другой) [54].

Для описания нечётких кластеров будем использовать следующую *матрицу нечёткого разбиения:*

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

i-ая строчка которой содержит степени принадлежности М-графа $(x_{i1},\dots x_{iM})$ к кластерам $S_1,\dots S_k$. В отличие от соответствующей матрицы четкого разбиения, степень принадлежности М-графа к кластеру при нечетком разбиении принимает вещественные значения из интервала [0, 1], в то время как, при четком разбиении степень принадлежности выбирается из двухэлементного множества {0, 1}.

Матрица *S* должна соответствовать следующим условиям:

$$\sum_{j=1}^{k} \mu_{ij} = 1, i \in \{1,\dots N\};$$

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

Таким образом, при наличии М-графов, лежащих на границе кластеров на примерно равном удалении от их центров, им можно сопоставить степени принадлежности, равные *0,5*. Однако у приведенных выше условий есть свои

недостатки. Например, трудности возникают в том случае, когда в выборке оказались М-графы, удаленные от центров всех найденных кластеров. Такие графы имеют мало общего с любым из рассматриваемых кластеров. Однако, ограничение на сумму степеней принадлежности, равную единице, не позволяет назначить для них малые степени принадлежности. Для преодоления этого недостатка необходимо ослабить данное ограничение. Взяв, например, в качестве ограничения требование принадлежности хотя бы одному кластеру для любого М-графа, получим следующие условия для нечеткого разбиения:

$$\sum_{j=1}^{k} \mu_{ij} = 1, i \in \{1,\dots N\};$$

$$\exists j, \quad \mu_{ij} > 0 \quad \forall i .$$

Для оценки качества разбиения используется критерий разброса, показывающий сумму расстояний от М-графов до центра своего кластера. Для евклидового пространства этот критерий записывается следующим образом [52]:

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

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

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

В настоящее время предложено большое количество алгоритмов кластеризации, минимизирующих в процессе своей работы указанный критерий разброса. В общем случае задача поиска матрицы нечеткого разбиения, соответствующей минимальному значению критерия, является задачей нелинейной оптимизации, для решения которой можно использовать различные под-

ходы. Наиболее популярным на практике является известный алгоритм нечетких *c*-средних [54]. Алгоритм основан на методе неопределенных множителей Лангранжа и позволяет найти локальный минимум задачи оптимизации. Надо понимать, однако, что выполнение алгоритма из различных начальных приближений может приводить к различным результатам.

В результате нечеткой кластеризации получим разбивку исходного пространства на нечёткие кластеры. Внутри каждого из них строим локальную классифицирующую модель (далее предполагается, что это линейная регрессия, но может быть использован и любой другой алгоритм). Пусть для простоты имеем два возможных значения свойства, а именно – активное (обладает указанным свойством) и неактивное (не обладает указанным свойством), обозначим их соответственно числовыми значениями *1* и *-1*.

Для нового М-графа $\tilde{x} = (\tilde{x}_1, \ldots \tilde{x}_M)$ у нас есть *k* предсказаний свойства 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 данного М-графа 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.

2.3.3 Параметры нечёткой классификации

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

В работе [51] подробно описаны некоторые методы построения кластерной структуры обучающей выборки. В контексте настоящей работы, интерес

представляет «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 М-графа $(x_{i1},\dots x_{iM})$ 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, радиусом кластера будем называть значение $r_i = \max_{x_j \in S_i} \rho(x_j, Z_i)$.

Построим матрицу нечёткого разбиения $\tilde{S} = [\mu_{ij}]$, в которой *i*-ая строчка содержит степени принадлежности М-графа $(x_{i1},\dots x_{iM})$ 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.

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 содержательно использовать кластерную структуру выборки, не ограничиваясь при этом пределами лишь одного кластера. На **рисунке 5** приведен пример нечёткой кластеризации, полученной из заданного разбиения множества точек на чёткие кластеры.

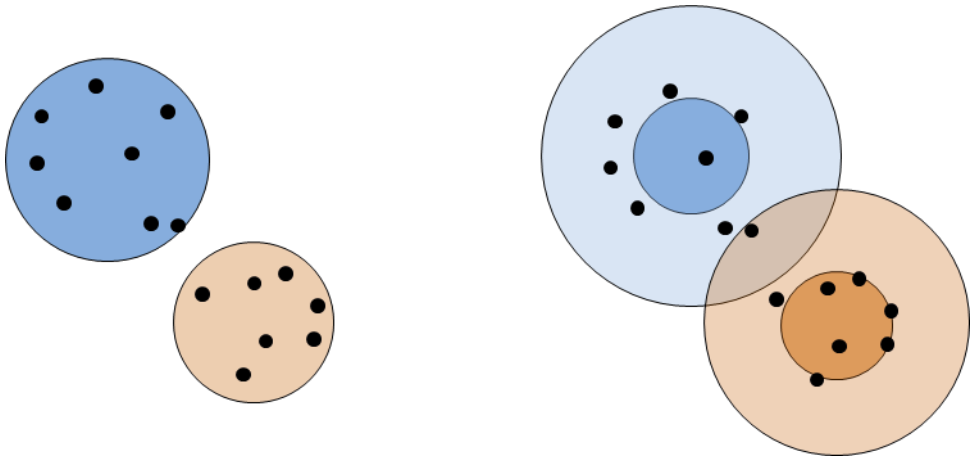


Рисунок 5. Чёткая и нечёткая кластерные структуры

«Нечёткость» классифицирующей функции на границах кластеров позволяет прогнозировать свойства одного и того же М-графа в разных локальных моделях и получать ответ взвешенным голосованием. На практике наблюдается значительное улучшение качества прогноза за счёт кусочной линейности строящихся моделей, а также с помощью оптимизации моделей по параметрам $\lambda_1, \lambda_2 \in \mathbb{R}$.

2.4 Двухфазная схема решения задачи «структура – свойство»

Ниже представлена двухфазная схема решения задачи «структура – свойство». В **подразделе 2.4.1** дается метод построения модели «структура – свойство» и ограничений допустимости согласно двухфазной схеме решения.

В **подразделе 2.4.2** доказана оценка качества прогнозирования с использованием двухфазной схемы.

2.4.1 Описание двухфазной схемы решения задачи «структура – свойство»

Пусть обучающая выборка LS состоит из N молекулярных графов x_i, $i = 1,\ldots,N$, каждому из которых поставлено в соответствие одно из значений: «1» или «–1». Значение «1» при этом соответствует М-графам, обладающим целевым свойством, значение «–1» соответствует М-графам, не обладающим целевым свойством. Вектор, последовательно содержащий значение целевого свойства всех М-графов обучающей выборки, обозначим $y = (y_1, y_2, \ldots, y_N)$, $y_i \in \{-1,1\}$.

Пусть также построена распознающая модель, решающая исходную задачу классификации, т.е. $RM_1(x_i) \in \{-1,1\}$ для любых $x_i \in LS$. Назовем RM_1 моделью первого уровня.

Напомним, что процедура скользящего контроля (leave-one-out cross-validation [10]), описанная ранее в **разделе 1.3**, заключается в следующем: из обучающей выборки последовательно удаляется каждый М-граф, по оставшимся М-графам строится распознающая модель, и с помощью этой модели прогнозируется свойство удаленного М-графа. Далее будет использован показатель качества моделей со скользящим контролем, равный отношению количества верных прогнозов к общему числу спрогнозированных М-графов.

Обозначим через R_1 – множество тех М-графов обучающей выборки x_i, для которых полученные в ходе процедуры скользящего контроля значения целевого свойства совпадают с действительными: $RM_1(x_i) = y_i$, т.е. множество верно классифицированных моделью первого уровня М-графов. Через W_1 обозначим множество ошибочно классифицированных моделью первого уровня М-графов: $W_1 = \{x_i \in LS \mid RM_1(x_i) \neq y_i\}$. Таким образом, показатель ка-

чества со скользящим контролем для модели первого уровня равен $\varphi_1 = |R_1|/N$.

Определим задачу классификации второго уровня. Всем М-графам обучающей выборки, для которых получен верный прогноз свойства моделью первого уровня (их $|R_1|$), поставим в соответствие значение «1», а М-графам, спрогнозированным неверно (их $|W_1|$), поставим в соответствие значение «–1». Сформируем, таким образом, вектор $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$, $\hat{y}_i \in \{-1,1\}$:

$$\hat{y}_i = \begin{cases} 1, & \textit{если} \quad RM_1(x_i) = y_i; \\ -1, & \textit{если} \quad RM_1(x_i) \neq y_i, \end{cases} \quad i = 1, \ldots, N.$$

Появившуюся в ходе реализации предлагаемого подхода новую задачу классификации назовем задачей классификации второго уровня.

Пусть построена распознающая модель RM_2, решающая задачу классификации второго уровня, т.е. $RM_2(x_i) \in \{-1,1\}$ для любых $x_i \in LS$. Назовем RM_2 моделью второго уровня.

Пусть в ходе процедуры скользящего контроля моделью второго уровня получено $|R_2|$ верных прогнозов, где $R_2 = \{x_i \in LS \mid RM_2(x_i) = \hat{y}_i\}$. Тогда показатель качества модели второго уровня $\varphi_2 = |R_2|/N$.

Наконец, определим результирующую распознающую модель RM_0. Результирующая модель решает исходную задачу классификации, однако, в отличие от модели первого уровня результирующая модель обладает опцией отказа от прогноза. То есть $RM_0(x_i) \in \{-1,0,1\}$ $\forall x_i \in LS$ и значение $RM_0(x_i) = 0$ интерпретируется как отказ от прогноза свойства М-графа x_i.

Для $x_i \in LS$:

$$RM_0(x_i) = \begin{cases} 1, & \textit{если} \quad RM_2(x_i) = 1 \quad \textit{и} \quad RM_1(x_i) = 1; \\ -1, & \textit{если} \quad RM_2(x_i) = 1 \quad \textit{и} \quad RM_1(x_i) = -1; \\ 0, & \textit{если} \quad RM_2(x_i) = -1. \end{cases}$$

Таким образом, результирующая модель осуществляет отказ от прогноза тогда, когда модель второго уровня предсказывает, что модель первого уровня ошибается, и осуществляет прогноз целевого свойства моделью первого уровня в противном случае (см. **рисунок 6**).

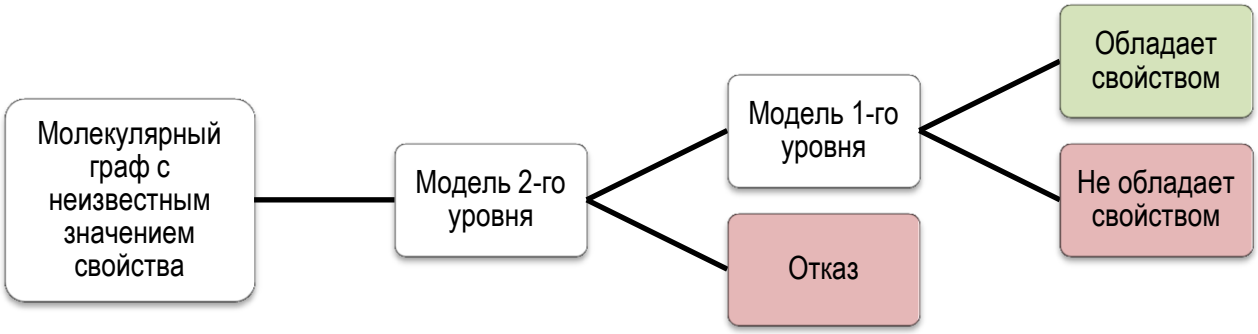


Рисунок 6. Двухфазная схема прогнозирования свойств молекулярных графов

Как и ранее, обозначим через $R_0 = \{x_i \in LS \mid RM_0(x_i) = y_i\}$ множество М-графов, верно классифицированных результирующей моделью. Пусть также через Reject обозначено количество отказов от прогноза. Тогда показатель качества результирующей модели $\varphi_0 = \mid R_0 \mid /(N - \text{Reject})$.

2.4.2 Оценка качества результирующей модели

Лемма 1. Если $\varphi_2 = 1$, то $\varphi_0 = 1$, а $\text{Reject} = \mid W_1 \mid$.

Доказательство. Условие $\varphi_2 = 1$ означает, в частности, что $\mid R_2 \mid = N$ и модель второго уровня классифицирует верно все М-графы обучающей выборки. Верная классификация моделью второго уровня означает, в свою очередь, что отказ от прогноза результирующей моделью осуществляется в том

и только в том случае, когда модель первого уровня классифицировала М-граф ошибочно. Таких М-графов по определению $|W_1|$. Отсюда Reject $=|W_1|$.

Таким образом, все М-графы, для которых прогноз осуществляется, попадают в множество R_1 и классифицируются верно. Получаем $|R_0| = |R_1| = N - |W_1| = N - \text{Reject}$ и $\varphi_0 = 1$. □

Лемма 2. Если $\text{Reject} = 0$, то $\varphi_0 = \varphi_1$.

Доказательство. Отсутствие отказов от прогноза означает, что $RM_2(x_i) = 1, \quad \forall x_i \in LS$ и прогноз для любого М-графа обучающей выборки всегда осуществляется моделью первого уровня $RM_0(x_i) = RM_1(x_i), \quad \forall x_i \in LS \quad \Rightarrow \quad \varphi_0 = \varphi_1$. □

Сформулируем более общий результат.

Теорема. Верна следующая оценка качества результирующей модели:

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \text{Reject}}{2(N - \text{Reject})}.$$

Доказательство. По определению $\varphi_1 N = |R_1|$, а $\varphi_2 N = |R_2|$. Кроме того $R_0 = R_1 \cap R_2$. Последнее следует из того что: если $x_i \in R_1 \setminus R_2$, то модель первого уровня осуществляет верный прогноз, однако модель второго уровня (ошибаясь) возвращает значение –1, в силу чего $RM_0(x_i) = 0$, таким образом, происходит отказ от прогноза. Если $x_i \in R_2 \setminus R_1$, то модель первого уровня ошибается, а модель второго уровня снова возвращает –1, что опять означает отказ от прогноза. Когда $x_i \notin R_1$ и $x_i \notin R_2$, модель первого уровня ошибается, в то время как модель второго уровня возвращает значение 1, таким образом, осуществляется неверное прогнозирование.

Учитывая изложенное выше, заметим, что отказам от прогноза соответствуют множества $R_2 \setminus R_1$ и $R_1 \setminus R_2$. Следовательно, Reject $=|R_1 \Delta R_2|$.

Таким образом, числитель дроби в формулировке теоремы приобретает вид $|R_1|+|R_2|-|R_1 \Delta R_2|$. Далее,

$$|R_1|+|R_2|-|R_1 \Delta R_2| = 2|R_1 \cap R_2| = 2|R_0|,$$

и, сокращая дробь на 2, имеем выражение для показателя качества φ_0. □

Следствие 1. Пусть $\varphi_{\min} = \min(\varphi_1, \varphi_2) > 1/2$. Тогда, если $\text{Reject} > 0$, то $\varphi_0 > \varphi_{\min}$.

Доказательство. Заметим, что $\text{Reject} < N$. Действительно, так как $\varphi_1 > 1/2$ и $\varphi_2 > 1/2$, то $R_1 \cap R_2 \neq \varnothing$ и существуют верные ответы результирующей модели.

Теперь по теореме имеем

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \text{R}eject}{2(N - \text{R}eject)}, \Rightarrow \varphi_0 \geq \frac{2\varphi_{\min} N - \text{R}eject}{2(N - \text{R}eject)} =$$

$$= \frac{2\varphi_{\min} N - 2\varphi_{\min}\text{R}eject + 2\varphi_{\min}\text{R}eject - \text{R}eject}{2(N - \text{R}eject)} =$$

$$= \frac{2\varphi_{\min}(N - \text{R}eject) + (2\varphi_{\min} - 1)\text{R}eject}{2(N - \text{R}eject)} =$$

$$= \varphi_{\min} + \frac{(\varphi_{\min} - 1/2)\text{R}eject}{N - \text{R}eject} > \varphi_{\min}.$$

□

Следствие 2. Если $\varphi_2 \geq \varphi_1 > 1/2$, то в случае $\text{R}eject > 0$ имеем $\varphi_0 > \varphi_1$.

Доказательство достаточно просто вытекает из следствия 1. □

Следствие 3. Если $\varphi_2 > \varphi_1 > 1/2$, то $\varphi_0 > \varphi_1$.

Доказательство вытекает из доказательства следствия 1. □

Таким образом доказано, что если модель первого уровня классифицировала М-графы из обучающей выборки хотя бы чуть лучше, чем случайным образом, и качество модели второго уровня не хуже качества модели первого уровня, то при условии, что количество отказов от прогноза больше нуля, результирующая модель демонстрирует более высокое качество классификации на исходной задаче, чем модель первого уровня. Также доказано улучшение качества результирующего прогноза в случае, когда качество модели второго уровня превосходит качество модели первого уровня.

Замечание 1. Несложно видеть, что вышеописанная схема решения, а также оценки качества для результирующей модели остаются в силе, если исходная задача классификации является задачей с несколькими классами. В таком случае компоненты вектора целевого свойства $y = (y_1, y_2, \ldots, y_N)$ принимают значения из заданного конечного числа меток классов $y_i \in \{Cl_1, Cl_2, \ldots, Cl_H\}$. При этом определение задачи классификации второго уровня остается прежним, то есть данная задача по-прежнему является задачей бинарной классификации.

Замечание 2. Также отметим, что все вышеприведенные рассуждения проходят в случае, когда рассматриваемый алгоритм обучения модели уже обладает опцией отказа от прогноза. В таком случае, вместо общего числа М-графов обучающей выборки N, во всех выкладках будет принимать участие величина $N_1 = N - \mathrm{Reject}_1$, равная числу М-графов, для которых осуществляет прогноз модель первого уровня.

Преимущества двухфазной схемы решения

Как уже было отмечено ранее, применение двухфазной схемы решения задача «структура – свойство» позволяет улучшить качество прогноза на обучающей выборке за счет осуществления отказа от прогноза. В **главе 3** настоящей работы приводятся результаты использования схемы на практике.

Кроме улучшения качества прогноза, двухфазная схема предполагает возможность независимой адаптации дескрипторного описания обучающей выборки под каждую из задач классификации (первого и второго уровня). Это обстоятельство означает, что распознающая модель второго уровня может использовать для прогнозирования совсем иные дескрипторы, нежели модель первого уровня. В свою очередь, это позволяет, накладывая определенные ограничения на отбор дескрипторов для модели второго уровня, добиться низкой вычислительной сложности для правил отказа, что является существенным при скрининге больших баз соединений.

Преимуществом предлагаемого подхода является также его универсальность. Описанная схема не зависит от конкретных алгоритмов классификации и предоставляет исследователю широкую свободу в выборе методов построения распознающих моделей. Кроме того, как будет показано ниже, двухфазная схема может быть использована последовательно для построения более сложных и эффективных классификаторов.

2.4.3 Интерпретация двухфазной схемы на примере метода опорных векторов

Рассмотрим бинарный классификатор для задачи распознавания из N обучающих объектов $\{(x_1, y_1),(x_2, y_2),\ldots(x_N, y_N)\}$, где $y_i \in \{-1,+1\}$, $i = 1,\ldots N$. Если объекты данной обучающей выборки могут быть разделены в пространстве признаков линейной гиперплоскостью $w \cdot x + b = 0$, то такой случай называется случаем линейно разделимых классов. При этом очевидно существует бесконечное множество возможных разделяющих гиперплоскостей.

Метод опорных векторов позволяет максимизировать зазор (margin) между классами[40]. Итоговый вид функции классификации, получаемой с помощью метода опорных векторов, описывается следующим выражением:

$$f(x) = sign(\sum_{i=1}^{N} a_i y_i \langle x_i, x \rangle + b),$$

где a_i – коэффициенты Лагранжа, полученные из решения задачи квадратичной оптимизации – максимизации зазора между классами:

$$w = \arg\min\{\|w\|^2 / 2\},\ y_i(w \cdot \mathrm{x}_i - b) \geq 1.$$

Для коррекции влияния ошибок на итоговую классификацию на практике также рассматривают другую задачу оптимизации:

$$w = \arg\min\{\|w\|^2 / 2 + C\sum_i \varepsilon_i\},\ y_i(w \cdot \mathrm{x}_i - b) \geq 1 - \varepsilon_i,$$

где переменные ε_i, характеризуют величину ошибки на прецеденте x_i, а параметр C регулирует отношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки.

Рассмотрим пример классификации методом опорных векторов, представленный на **рисунке 7**. Объектам x_i с $y_i = 1$ соответствуют синие точки на графике, а объектам с $y_i = -1$ соответствуют красные. Пространство дескрипторов задано значениями дескрипторов D_1, D_2.

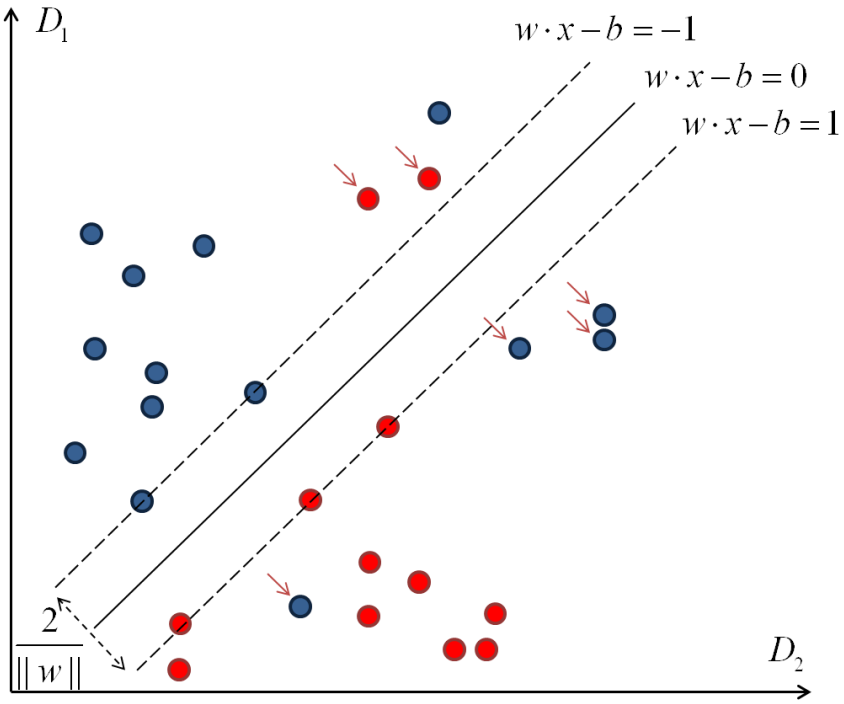


Рисунок 7. Пример классификации методом опорных векторов.

Можно видеть на **рисунке 7** объекты, которые были классифицированы неверно (отмечены стрелками). В соответствии с предложенной двухфазной схемой решения этим объектам (множество W_1 согласно введенным ранее обозначений) ставятся в соответствие значения нового вектора целевого свойства, равные «–1». Значение целевого вектора для объектов, классификация которых осуществлена верно (множество R_1), равны «1».

Рисунок 8 иллюстрирует переход от задачи классификации первого уровня к задаче классификации второго уровня. Объекты из множества R_1 обозначены зелеными точками на графике, объекты из множества W_1 – желтыми.

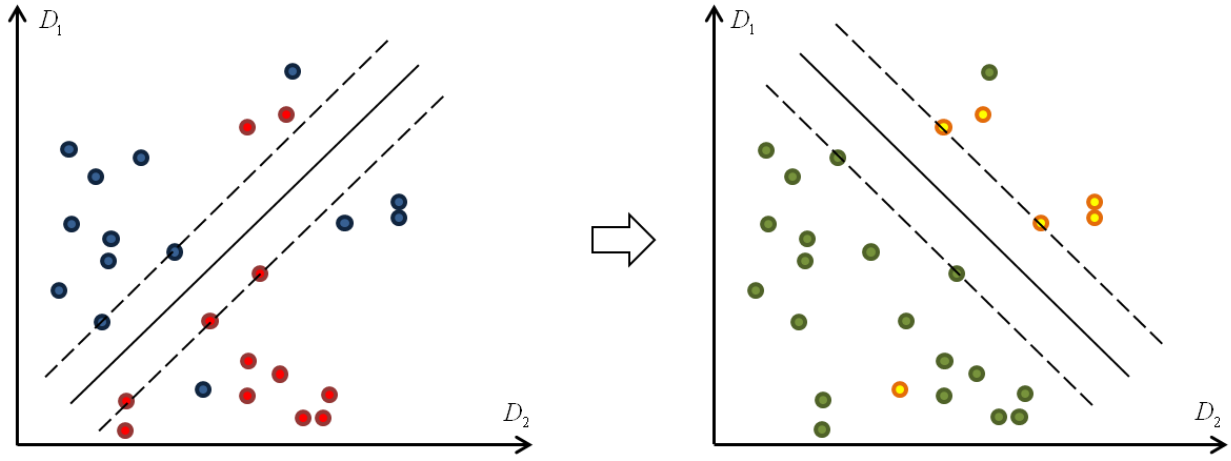


Рисунок 8. Решение задач классификации первого и второго уровня методом опорных векторов.

Качество классификации второго уровня в примере лучше качества исходной классификации (количество неверных прогнозов снизилось). Итоговая классификация результирующей моделью показана на **рисунке 9**. При этом для объектов x_i, для которых получено значение $RM_2(x_i) = -1$ производится отказ от прогноза (обозначено серым на графике).

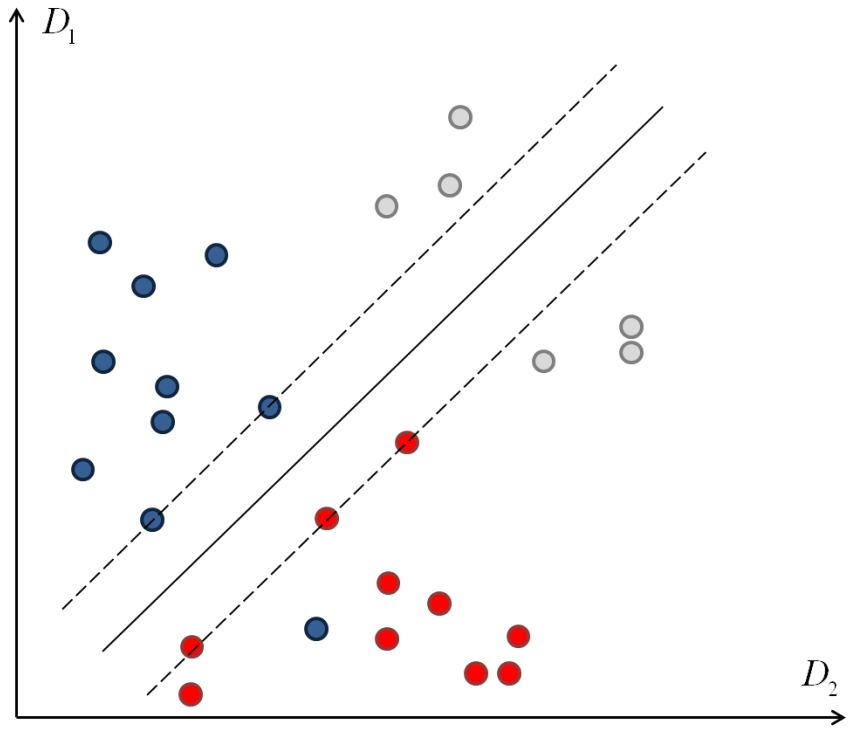


Рисунок 9. Итоговая классификация результирующей моделью.

Отметим, что в общем случае (когда в SVM используются ядра), прямые на представленных графиках принимают вид произвольных кривых. Также в примере для наглядности при решении задач классификации первого и второго уровня использовалось одно и то же пространство дескрипторов, что необязательно в общем случае.

Ключевым отличием двухфазной схемы от методов поиска выбросов (outliers)[55] является то обстоятельство, что разделяющая плоскость, построенная при решении исходной задачи классификации, в последствие не перестраивается.

2.4.4 Модификация двухфазной схемы без использования отказов от прогноза

Для случая бинарной классификации возможно также определить следующую модификацию двухфазной схемы, которая не предполагает использования отказов от прогноза.

Обозначим обратную распознающую модель к модели RM через $\overline{RM}$, положив по определению:

$$\overline{RM}(x_i) = \begin{cases} 1, & если \quad RM(x_i) = -1; \\ -1, & если \quad RM(x_i) = -1, \end{cases} \quad i = 1, \ldots, N.$$

Далее пусть, как и ранее, построена распознающая модель первого уровня, решающая исходную задачу классификации $RM_1(x_i) \in \{-1,1\}$ для любых $x_i \in LS$. Пусть также через R_1 обозначено множество верно классифицированных моделью первого уровня М-графов ($RM_1(x_i) = y_i$), а через W_1 обозначено множество ошибочно классифицированных моделью первого уровня М-графов: $W_1 = \{x_i \in LS \mid RM_1(x_i) \neq y_i\}$.

Рассмотрим распознающую модель $\overline{RM_1}$, обратную к модели первого уровня RM_1, обозначив аналогично через $\overline{R_1}$ и $\overline{W_1}$ множества верно и неверно классифицированных М-графов моделью $\overline{RM_1}$. Легко заметить, что $\overline{R_1} = W_1$ и $\overline{W_1} = R_1$ в силу определения обратной модели.

Пусть также, как ранее, построена модель второго уровня, решающая задачу классификации второго уровня. Тогда результирующую распознающую модель можно задать следующим образом.

Для $x_i \in LS$:

$$RM_0(x_i) = \begin{cases} RM_1(x_i), & если \quad RM_2(x_i) = 1; \\ \overline{RM_1}(x_i), & если \quad RM_2(x_i) = -1. \end{cases}$$

Это означает что, в зависимости от результатов классификации, произведенной с помощью модели второго уровня, результирующая модель производит прогноз целевого свойства либо моделью первого уровня, либо обратной к ней моделью.

Для качества классификации определенной таким образом результирующей модели верна следующая очевидная оценка.

Утверждение 1. В обозначениях, данных выше, качество классификации результирующей модели не зависит от качества исходной классификации и равно качеству классификации модели второго уровня ($\varphi_0 = \varphi_2$).

Как и прежде сформулируем также тривиальное следствие из полученного результата.

Следствие 1. Если $\varphi_2 > \varphi_1$, то $\varphi_0 > \varphi_1$.

В этом случае для улучшения качества исходной классификации достаточно, чтобы качество классификации моделью второго уровня превосходило качество классификации моделью первого уровня.

Отмеченный результат показывает, что есть возможность применять двухфазную схему и без использования опции отказа от прогноза. В таком случае также существуют основания полагать, что исходное качество классификации может быть улучшено с помощью применения классификатора второго уровня.

2.4.5 Приложения двухфазной схемы

Данный раздел посвящен возможным приложениям двухфазной схемы решения в задаче «структура – свойство». Основной результат, полученный в **разделе 2.4**, позволяет, прежде всего, констатировать два основных положительных свойства такого подхода. Во-первых, двухфазная схема решения является универсальным (не зависящим от конкретного типа распознающих моделей) инструментарием построения ограничений допустимости. А, во-вторых, ее применение позволяет в ряде случаев, как доказано, улучшить качество прогнозирования. Выше также показано, как двухфазная схема решения может быть использована для построения более эффективных по скорости их реализации на ЭВМ правил отказа от прогноза при скрининге боль-

ших баз химических соединений. Кроме этих, очевидных приложений, полученные результаты позволяют также использовать подход для более глубокого анализа обучающей выборки. Ниже рассмотрим возможные примеры использования разработанной схемы в решении задачи «структура – свойство».

Последовательное «вычерпывание» ошибки

Отметим, что двухфазная схема решения может быть применена тогда, когда качество исходной классификации является низким. Мощность множества М-графов, спрогнозированных неверно с помощью модели первого уровня, в таком случае велика и сравнима с мощностью исходной обучающей выборки. Тогда можно ожидать, что содержательное разделение М-графов на две группы с помощью модели второго уровня позволит нам получить множество допустимых М-графов, где классификация будет относительно качественной (в смысле величины показателя качества).

Здесь следует заметить, что для оставшихся М-графов прогноз не осуществляется (они не являются допустимыми для модели первого уровня).

Однако разбиение М-графов исходной выборки на два множества моделью второго уровня может быть использовано в качестве результата кластеризации. Тогда, в каждом из получившихся кластеров можно построить независимые локальные модели, ожидая что, качество прогнозирования каждой из них будет выше, чем качество прогнозирования на исходной обучающей выборке (что доказано утверждением теоремы из **2.4.2**, по крайней мере, для одного из кластеров).

Данный процесс можно использовать итерационно, выделяя на каждом шаге кластеры, где локальная модель является достаточно качественной, и области, в которых прогнозирование затруднено, продолжая в последствие работать со сложными областями.

Такой итерационный процесс назовем последовательным вычерпыванием ошибки.

Заметим, что указанный подход полностью согласуется с разработанной методологией согласованного прогнозирования с использованием множеств локальных ограниченных моделей и является одной из его реализаций.

Многоуровневая классификация

Помимо «горизонтального» распространения двухфазной схемы, когда процесс последовательно применяется для различных подмножеств обучающей выборки, можно также определить «вертикальную» композицию распознающих моделей. Идея заключается в том, что оценки выходов модели второго уровня могут быть в свою очередь использованы для определения задачи классификации более высокого уровня.

Пусть через R_2 обозначено множество М-графов обучающей выборки x_i, для которых полученные в ходе процедуры скользящего контроля значения прогнозов с помощью модели второго уровня совпадают со значениями вектора целевого свойства второго уровня: $RM_2(x_i) = \hat{y}_i$, т.е. множество верно классифицированных моделью второго уровня М-графов. Через W_2 обозначим множество ошибочно классифицированных моделью второго уровня М-графов: $W_2 = \{x_i \in LS \mid RM_2(x_i) \neq \hat{y}_i\}$.

Определим задачу классификации третьего уровня. Всем М-графам обучающей выборки, спрогнозированным верно с использованием модели второго уровня, поставим в соответствие значение «1», а М-графам, спрогнозированным неверно, поставим в соответствие значение «–1». Таким образом формируется вектор целевого свойства третьего уровня, характеризующий качество классификации модели второго уровня $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N)$, $\tilde{y}_i \in \{-1,1\}$:

$$\tilde{y}_i = \begin{cases} 1, & \textit{если} \quad RM_2(x_i) = \hat{y}_i; \\ -1, & \textit{если} \quad RM_2(x_i) \neq \hat{y}_i, \end{cases} \quad i = 1, \ldots, N.$$

Возникшую, таким образом, задачу классификации назовем задачей классификации третьего уровня и обозначим через RM_3 распознающую модель, решающую данную задачу.

Результирующая модель в случае трехуровневой классификации задается следующими соотношениями:

$$RM_0(x_i) = \begin{cases} 1, & \textit{если} \quad RM_1(x_i) = RM_2(x_i) = RM_3(x_i) = 1; \\ -1, & \textit{если} \quad RM_2(x_i) = RM_3(x_i) = 1 \quad \textit{и} \quad RM_1(x_i) = -1; \\ 0, & \textit{если} \quad RM_3(x_i) = -1 \quad \textit{или} \quad RM_2(x_i) = -1. \end{cases}$$

Описанную процедуру можно применять итерационно, наращивая количество уровней классификации до тех пор, пока качество классификации результирующей модели улучшается.

Кроме того предлагается другая вариация этого подхода. Когда для постановки задачи классификации следующего уровня используются не классификация с помощью модели предыдущего уровня, а классификация с помощью результирующей моделью. Для этого в представленном выше тексте, необходимо везде, где рассматривается RM_2, использовать RM_0. Сравнение этих вариаций представляет интерес для дальнейшего исследования, как и тестирование многоуровневой классификации на практике.

2.6 Оценки вычислительной сложности

В данном разделе приведем оценки вычислительной сложности алгоритмов построения распознающих моделей, а также алгоритмов прогнозирования свойств новых неизученных М-графов. Отметим, что, в первую очередь, ценность с позиции применимости данного подхода к задаче виртуального скрининга представляют «быстрые» алгоритмы классификации новых М-графов. Напротив, требования к процессу обучения при реализации моде-

лей и построения ограничений допустимости могут быть снижены, так как процесс обучения и построение ограничений допустимости для моделей происходит единожды и не повторяется в процессе скрининга.

Опишем подход в более формальных терминах.

Пусть для фиксированного способа описания структуры М-графа (фиксированного описывающего отображения) сложность вычисления каждого дескриптора одинакова и равна CD.

Описывающие отображения при этом могут отличаться по сложности, для этого рассмотрим несколько уровней дескрипторного описания. Самые «легкие» с вычислительной точки зрения дескрипторы вычисляются со сложностью CD_l, более сложные CD_h.

Пусть размерность алфавита дескрипторов для выбранного описывающего отображения составляет M, тогда при условии, что вычислительная сложность одного дескриптора равна CD, сложность построения матрицы «молекулярный граф – дескриптор» равна $CD \cdot M$.

Построение модели методом линейной регрессии сводится к умножению исходной МД-матрицы на себя транспонированную и обращению полученной матрицы. Таким образом, сложность построения распознающей модели методом линейной регрессии можно считать равной $O(M^3)$. Более точной оценкой является $O(M^2 N)$. Как правило, другие алгоритмы классификации и кластеризации также имеют полиномиальную сложность (например, опоминавшиеся ранее SVM и иерархическая кластеризация). В настоящей работе обозначим сложность построения распознающей модели $CRM(M,N)$, где M – количество дескрипторов, N – число М-графов обучающей выборки.

Задача адаптации дескрипторного описания в общем случае имеет комбинаторную сложность (так как в процессе адаптации необходимо перебрать

всевозможные подмножества множества дескрипторов), то есть, относится к классу NP-сложных задач. В настоящей работе для проведения вычислительных экспериментов применялся эволюционный отбор дескрипторов, имеющий существенно более низкую сложность. Сформулируем оценку для сложности построения распознающей модели фиксированным методом обучения с эволюционным отбором дескрипторов.

Утверждение 2. Сложность построения распознающей модели с эволюционным отбором дескрипторов не превосходит $O(CRM(N,M)\cdot N\cdot M)$, где $CRM(M,N)$ – сложность построения распознающей модели, зависящая от количества дескрипторов M и числа М-графов в обучающей выборке N.

Доказательство. Эволюционный отбор моделей осуществляется по значениям показателя качества, вычисленного для отбираемых моделей. Заметим, что при использовании показателя качества со скользящим контролем, сложность вычисления его значения может быть оценена, как $CRM(N-1,\hat{M})\cdot N$ ($\hat{M}$ – текущее число дескрипторов, по которым строится модель), так как при этом необходимо N раз построить модель по обучающей выборке, состоящей из $(N-1)$ М-графа.

Таким образом, на первом шаге эволюционного отбора потребуется примерно $CRM(N-1,1)\cdot N\cdot M$ операций. Пусть теперь номер шага k. Тогда на k-ом шаге необходимо произвести $sel\cdot CRM(N-1,k)\cdot N\cdot M$ операций. Тогда общее число операций не превосходит $O(sel\cdot CRM(N-1,nevol)\cdot N\cdot M)$ и так как sel является константой, а $nevol<M$, то имеем заявленное в утверждении.□

Далее приведем оценку сложности построения ограничений допустимости при использовании двухфазной схемы решения задачи «структура – свойство».

Утверждение 3. Сложность построения ограничений допустимости при использовании двухфазной схемы решения задачи «структура – свойство» не превосходит $O(CRM(N,M)\cdot N\cdot M)$.

Доказательство. Так как задача построения ограничений допустимости при использовании двухфазной схемы суть задача классификации второго уровня, то для её решения потребуется столько же операций, что и для решения задачи классификации первого уровня. Учитывая необходимость проведения независимой адаптации дескрипторного описания, имеем оценку из утверждения. □

Таким образом, учитывая представленные выше оценки, общее число операций необходимое для обработки обучающей выборки при использовании двухфазной схемы решения составляет:

$$CD\cdot M + O(CRM(N,M)\cdot N\cdot M) + O(CRM(N,M)\cdot N\cdot M) = \\ = CD\cdot M + O(CRM(N,M)\cdot N\cdot M).$$

Отметим, что, так как кластеризация, как правило, имеет сложность порядка $CRM(M,N)$, то использование нечеткого классификатора кажется более выгодным с вычислительной точки зрения по сравнению с двухфазной схемой, которая, напомним, предполагает независимую адаптацию дескрипторного описания под задачу второго уровня, что, в свою очередь, приводит ко второму эволюционному отбору. Однако, если параметры «нечеткости» кластеров оптимизируются, сложность таких правил отказа резко увеличивается. Необходимость использования прямых методов условной оптимизации по одному или нескольким параметрам «нечеткости» приводит к тому, что итоговая сложность построения правил отказа на базе кластерной структуры становится сопоставима со сложностью классификации второго уровня. При этом важно помнить, что классификация новых М-графов с использований

двухфазной схемы проходит в общем случае быстрее, чем с использованием кластерных правил отказа.

Рассматриваемые в контексте данной работы методы (регрессия, SVM) позволяют получить прогноз для нового М-графа за линейное количество операций. В таком случае, если M_1 – количество дескрипторов, эволюционно отобранных для решения задачи классификации первого уровня, а M_2 – количество дескрипторов, отобранных для решения задачи второго уровня, то имеет место следующая оценка.

Утверждение 4. Вычислительная сложность прогнозирования свойств М-графа с использованием двухфазной схемы равна $O(CD \cdot \max(M_1, M_2))$.

Доказательство. Сложность прогнозирования свойств нового М-графа складывается из сложности процедур вычисления дескрипторов для задачи классификации первого и второго уровня, а также из вычисления линейных функций данных дескрипторов. Таким образом, имеем:

$$CD \cdot M_1 + CD \cdot M_2 + O(M_1) + O(M_2) = O(CD \cdot \max(M_1, M_2)).$$ □

Замечание 3. Отказ от прогноза для недопустимых М-графов при этом осуществляется за $O(CD \cdot M_2)$ опреаций.

Рассмотрим теперь прогнозирование свойств нового М-графа с помощью нечеткого классификатора. Пусть для построения быстрых правил отказа использовалось специальное пространство дескрипторов, более «легких» с вычислительной точки зрения (сложность вычисления одного такого дескриптора CD_l, их число обозначим M_l), а для прогнозирования применялись эволюционно отобранные дескрипторы основного дескрипторного пространства (их сложность вычисления CD_h, отобраны для прогнозирования $\hat{M}$ из M дескрипторов). Пусть кластерная структура задана центрами и ра-

диусами *k* кластеров. Тогда для прогнозирования целевого свойства нечетким классификатором верна следующая оценка вычислительной сложности.

Утверждение 5. Для прогнозирования целевого свойства М-графа с помощью нечеткого классификатора потребуется $O(CD_l \cdot M_l) + O(CD_h \cdot \hat{M})$ операций.

На практике оба рассматриваемых подхода к прогнозированию свойств новых неизученных М-графов (нечеткий классификатор и двухфазная схема) показывают схожие результаты по вычислительной производительности. Причина в том, что, несмотря на использование пространства вычислительно легких дескрипторов в случае нечеткого классификатора, их число, как правило, на несколько порядков превышает количество дескрипторов, отбираемых в процессе адаптации.

2.7 Понижение вычислительной сложности дескрипторного описания обучающей выборки

В данном разделе обсуждается возможность понижения вычислительной сложности построения моделей «структура – свойство» с использование кластерной структуры обучающей выборки и двухфазной схемы решения задачи. В первую очередь представленный ниже метод оптимизации дескрипторного описания полезен при обработке больших неоднородных выборок молекулярных графов. Оптимизация понимается в настоящем разделе как понижение вычислительной сложности построения моделей «структура – свойство» без потери их качества.

Актуальной в задачах прогнозирования свойств молекулярных графов является обработка больших неоднородных выборок, содержащих в себе М-графы, имеющие существенно различную структуру, относящиеся к различным классам, обладающие различными физико-химическими свойствами. Традиционным подходом к обработке таких выборок можно считать предва-

рительный анализ, в ходе которого исследователи заранее группируют М-графы по тем или иным признакам в выборки меньшего размера. На данном этапе также может быть полезна кластеризация, выполненная по максимально упрощенному описанию.

Более формально, назовем обучающую выборку LS *неоднородной*, если стандартное отклонение качества моделей «структура – свойство», построенных с использованием дескрипторов первого уровня по случайным подмножествам обучающей выборки LS, выше заданного порога φ_p.

Естественным будет предположить, что в силу неоднородности исходной обучающей выборки для эффективного решения задачи «структура – свойство», для различных М-графов может понадобиться различное дескрипторное описание. Упомянутая выше адаптация дескрипторного описания не решает эту задачу, так как адаптирует описание выборки в целом.

Простой подход к обработке неоднородных выборок выглядит следующим образом:

- задается упрощенное описание обучающей выборки (дескрипторы невысокой сложности);
- по данному описанию выполняется кластеризация исходной обучающей выборки;
- в каждом из полученных кластеров независимо решается задача «структура – свойство» с использованием ограничений допустимости и адаптации дескрипторного описания.

Таким образом, на выходе имеем множество моделей, построенных по своим обучающим выборкам (непересекающимся в случае четкой кластеризации и имеющим пересечения в случае использования методов нечеткой кластеризации), использующих каждая свое адаптивное дескрипторное описанию.

Ключевым недостатком описанного подхода является необходимость построения общего множества дескрипторов для каждой модели с проведением последующего отбора. Таким образом, если появляется необходимость использовать дескрипторы высокого уровня, вычислительная сложность такого метода будет крайне высока. Ниже предлагается подход, позволяющий оптимизировать (сократить) общую вычислительную сложность обработки больших неоднородных выборок.

Пусть заданы описывающие отображения различного уровня сложности (определены дескрипторы 1-го, 2-го и последующих уровней). Причем для оценок вычислительной сложности дескрипторов разного уровня выполнено условие: $CD_1 < CD_2 < \ldots < CD_d$.

Зададимся некоторым значением показателя качества классификации, которого хотелось бы добиться на неоднородной обучающей выборке. Обозначим это значение φ_c. Как и в представленном выше подходе, первым этапом обработки выборки будет выделение общих кластеров с помощью наиболее простого дескрипторного описания. Здесь будем использовать наиболее простые с вычислительной точки зрения дескрипторы в силу того, что функции принадлежности М-графа кластерам впоследствии будут использоваться для задания ограничений допустимости для построенных моделей и поэтому должны вычисляться быстро.

Далее в каждом из полученных кластеров $K_1, K_2, \ldots K_k$ построим модель «структура – свойство» с помощью адаптации дескрипторного описания первого уровня. Применение адаптации здесь означает, что для классификации модель будет использовать только некоторые дескрипторы исходного пространства. Обозначим через $\mu_1, \mu_2, \ldots \mu_k$ функции принадлежности для кластеров:

$$\mu_i(x_j) = \begin{cases} 1, & \textit{если} \quad x_j \in K_i; \\ 0, & \textit{иначе}, \end{cases} \quad i = 1,\ldots,N,\ j = 1,\ldots k.$$

Вычислим значение показателя качества каждой из построенных моделей и обозначим значения показателей через $\varphi_i, \quad i = 1,\ldots k.$

Определим *обобщающую модель «структура – свойство»* следующим образом. Для $x_i \in LS:\ M(x_i) = \sum_{j=1}^{k} \mu_j(x_i) \cdot RM_j(x_i)$.

Показатель качества обобщающей модели обозначим через θ. Докажем промежуточный результат.

Утверждение 6. В обозначениях, данных выше, значение показателя качества обобщающей модели удовлетворяет неравенству: $\theta \geq \min_{i=1,\ldots,k}(\varphi_i)$.

Доказательство.

Количество М-графов в кластерах – $N_1, N_2, \ldots N_k$. По определению число верных прогнозов для каждой из моделей $R_i = \varphi_i N_i, \quad i = 1,\ldots k$. Рассмотрим

$$\theta = \frac{R_1 + R_2 + \ldots + R_k}{N} = \frac{\varphi_1 N_1 + \ldots + \varphi_k N_k}{N} =$$

$$= \frac{\varphi_1}{N} N_1 + \ldots + \frac{\varphi_k}{N} N_k \geq \frac{\varphi_{\min}}{N}(N_1 + \ldots + N_k) \geq \varphi_{\min}.\square$$

Теперь отберем те модели, для которых качество оказалось меньше ($\varphi_i < \varphi_c$). Для каждой из них можно построить классификатор второго уровня и результирующую модель с ограничениями допустимости. Если в результате этой процедуры для каких-то из моделей удалось повысить качество до требуемого, то они исключаются из рассматриваемого множества.

Составим новую обучающую выборку, в нее войдут М-графы из кластеров, для которых не удалось построить качественные модели на дескрипто-

рах первого уровня, а также М-графы, соответствующие отказам моделей использующих классификацию второго уровня.

Данную выборку будем рассматривать как выборку «трудных» М-графов, для которых не удалось осуществить качественную классификацию с помощью простого дескрипторного описания. Поэтому для указанной выборки вычисляются дескрипторы второго уровня. При этом для кластеризации новой выборки и для построений классификаторов второго уровня будем по-прежнему использовать самые «простые» дескрипторы (для того, чтобы результирующие ограничения допустимости вычислялись быстро).

Далее строится новая кластеризация и новые локальные модели (уже с использованием адаптации дескрипторного описания второго уровня). Те модели, качество которых удовлетворительно, добавляются к множеству моделей построенных на первом этапе, а для остальных производится процедура построения двухфазного классификатора.

Данный процесс продолжается до тех пор, пока не будет выполнен один или несколько критериев остановки. В качестве таких критериев можно предложить, например:

1) на новом этапе не удалось получить ни одной новой нетривиальной модели.
2) прогноз требуемого качества достигнут для большинства (заранее заданной части) М-графов обучающей выборки;
3) построены дескрипторы d-ого уровня сложности.

Таким образом, в результате работы указанного алгоритма, на выходе получим множество моделей со своими ограничениями допустимости (задающимися с помощью функций принадлежности к кластерам и соответствующих классификаторов второго уровня), удовлетворяющих требованиям качества, поставленным при обработке выборки. Для М-графов, недопустимых ни для одной из этих моделей, осуществляется отказ от прогноза.

Согласно **утверждению 6** качество обобщенной классификации будет также удовлетворять поставленным требованиям.

Кроме того, ограничения допустимости имеют низкую вычислительную сложность и дескрипторное описание оптимизировано по неоднородности выборки (сложные дескрипторы вычисляются только там, где это требуется).

2.8 Выводы

Таким образом, в настоящей главе предложен подход к прогнозированию свойств М-графов, включающая в себя метод адаптации дескрипторного описания, методы построения ограничений допустимости для моделей «структура – свойство», а также методы согласованного прогнозирования по множествам и семействам моделей «структура – свойство».

Описанный нечеткий классификатор и двухфазная схема решения задачи «структура – свойство» решают поставленную в **разделе 1.3** задачу построения адаптивных распознающих моделей, а также задачу построения ограничений допустимости для этих моделей. При этом двухфазная схема решения обладает рядом преимуществ. При некоторых условиях двухфазная схема гарантирует улучшение качества прогноза на обучающей выборке за счет осуществления отказа от прогноза, она предполагает возможность независимой адаптации дескрипторного описания под каждую из задач классификации, а также является универсальной – не зависит от того, насколько содержательно разбивается обучающая выборка на кластеры, а также от выбора конкретных алгоритмов обучения моделей. Представлены необходимые оценки вычислительной сложности алгоритмов построения распознающих моделей, построения ограничений допустимости и прогнозирования свойств новых соединений.

Предложен метод оптимизации сложности дескрипторного описания обучающей выборки, понижающий вычислительную сложность обработки неоднородных выборок М-графов.

Глава 3. Результаты использования предложенных подходов

Настоящая глава содержит результаты практического тестирования подхода. Описана реализация разработанных алгоритмов и приведены результаты их работы на реальных обучающих выборках соединений. Приведено сравнение результатов использования разработанного подхода с классическими аналогами, а также с полученными теоретическими оценками. Даны также подробные описания проведенных совместных научных исследований с Институтом Органической Химии им. Н.Д. Зелинского РАН и Российским Онкологическим Научным Центром им. Н.Н. Блохина РАМН. Показана практическая значимость разработанного подхода и перспективность его использования.

Все описанные в настоящей главе результаты получены на выборках химических соединений, предоставленных профильными институтами Российской академии наук и Российской академии медицинских наук. В качестве исходных данных предоставлялась информация о структуре соединений в форматах .sdf или .mol. Для обучающих выборок были доступны также значения активности.

Термин «активность» в данном разделе, несмотря на то, что обозначает вполне определенную биологическую или химическую активность, выраженную численно, в данной ранее терминологии должен восприниматься, как целевое свойство. В приводимых ниже экспериментах везде использовался порог, разделяющий неактивные и активные соединения по исходным числовым значениям. Таким образом, задача прогнозирования активности везде сводилась к задаче классификации с двумя классами. Также ниже понятие химическое соединение отождествляется с понятие его молекулярного графа.

3.1 Программная реализация предложенных методов

Как для вычислений дескрипторов М-графов на этапе описания, так и для построения распознающих моделей использовались скрипты и функции, реализованные в среде MATLAB. Расчеты были произведены с помощью программного обеспечения MathWorks MATLAB R2012b (Version 8.0) под управлением операционной системы Windows XP SP3. При написании скриптов использовались дополнительные пакеты Matlab Fuzzy Logic Toolbox и Statistics Toolbox, предлагающие широкий выбор инструментов для статистического исследования и работы с нечёткой логикой.

Всего в ходе работы над диссертацией было разработано порядка 50 модулей (.m) скриптов для реализации работы с химическими структурами, построения и анализа МД-матриц, классификации данных, построения распознающих моделей и ограничений допустимости, а также прогнозирования активности неизученных соединений.

3.1.1 Общее описания разработанного программного комплекса

Разработанный программный комплекс для построения моделей «структура – свойство» позволяет:

- загружать матрицу описания структур химических соединений из обучающей выборки в формате «молекула – дескриптор» (в качестве исходного файла используется формат хранения данных системы MATLAB – .mat). Согласно данному в **Главе 1** определению, элементом *(i, j)* МД-матрицы *X* выступает значение *j*-го дескриптора для *i*-го соединения обучающей выборки;
- выполнять отбор значимых дескрипторов и адаптацию дескрипторного описания под задачу прогнозирования конкретного свойства (эволюционный отбор дескрипторов);
- выполнять обучение и сохранение моделей «структура – свойство» с помощью стандартных алгоритмов множественной линейной регрес-

сии, алгоритмов построения деревьев решений и метода опорных векторов;

- оценивать качество полученных моделей с помощью процедуры скользящего контроля;
- построить и сохранить ограничения допустимости для стандартных моделей «структура – свойство»;
- использовать полученные модели «структура – свойство» и ограничения допустимости для них для прогнозирования свойств новых соединений.

Программный комплекс (набор модулей, разработанных автором), использует в качестве входных данных подготовленные МД-матрицы, сохраненные в формате хранения данных в системы MATLAB (.mat). На входе модулей-алгоритмов предполагаются файлы .mat, хранящие переменные системы MATLAB с именами *«matrix»*, содержащие обучающие МД-матрицы, построенные на этапе описания обучающей выборки.

Модули, разработанные автором, интегрированы в общую программную систему прогнозирования свойств химических соединений, включающую модули этапа описания обучающей выборки, которые реализуют расчет фрагментных дескрипторов особых точек М-графов. Общая схема прогнозирования свойств неизученных соединений представлена на **рисунке 10**.

Рисунок 10. Схема работы программного комплекса прогнозирования свойств химических соединений

Далее следует краткое описание процесса построения МД-матриц.

3.1.2 Предварительная обработка обучающей выборки

В качестве исходных данных при обработке выборок химических соединений использовалась информация о структуре соединений в форматах .sdf или .mol. Данные форматы широко распространены и являются в настоящее время стандартными форматами для обмена химической информацией. Описание химических соединений с помощью форматов SDF и MOL является способом представления информации о структуре химического соединения в виде текстового файла матрицы смежности молекулярного графа. Кроме того, в качестве файла, содержащего значения целевого вектора свойства ис-

пользовался текстовый файл, содержащий метки классов соединений из обучающей выборки, записанные столбцом.

Модуль расчета фрагментных дескрипторов на особых точках М-графов разработан коллегами автора [48] и представлен в виде 50 модулей-функций (файлов .m).

Необходимые исходные данные для работы модуля расчета дескрипторов:

1) выборка соединений обучающей выборки в формате sdf;
2) вектор активности для обучающей выборки – вектор-столбец значений в текстовом файле;
3) файл параметров обработки обучающей выборки;
4) файл дискретизации расстояния для облучающей выборки (для случая построения дескрипторов высокого уровня).

Описание основных функций модуля расчета фрагментных дескрипторов дано ниже.

OT1main(KeyFile, SDFFile) – построение МД-матрицы, дескрипторы первого уровня.

Вход: KeyFile – текстовый файл с параметрами обработки, SDFFile – необязательный параметр – имя файла sdf с обрабатываемой выборкой (если выборка уже представлена в виде нужного набора директорий, то обработка по параметрам, указанным в KeyFile)

Выход: в root_package сохранено описание соединений в виде заданных дескрипторов, МД-матрица.

Шаблон KeyFile:

```
root_package=demo_test_sample
mol_prefix=demo_test_sample
profile_file=SingleConnectivity
chain_length=2
```

distance_type=topologic
markers=db_
method=linear_fragments\level1,

где root_package – имя директории, в которой находятся полученные mol-файлы, а также список дескрипторов и МД-матрица, mol_prefix – начало имени формируемого mol-файла, profile-file – проверка условий. В данном случае – связность структуры, chain_length – длина формируемой цепочки связных атомов (значения: 2, 3, 4), distance_type – тип вычисляемого расстояния (значения: geometric/topologic), markers – включение маркеров (значения: ___, __r, _b_, d__, db_, d_r, _br, dbr), method – имя используемого метода и поддиректории, в которой будут храниться результаты.

OT2main(KeyFile, DistFile, SDFFile) – метод для построения особых точек второго уровня.

Вход: файл с параметрами запуска KeyFile, DistFile – значения и имена интервалов, расстояние между цепочками, SDFFile – необязательный параметр, используется при первичной обработке выборки, если sdf-файл не представлен в виде набора директорий.

Выход: в root_package сохранено описание соединений в виде заданных дескрипторов, МД-матрица.

3.1.3 Модуль построения и использования моделей «структура – свойство»

Модуль представляет собой набор скриптов .m, разработанных в системе MATLAB. В качестве алгоритмов обучения моделей «структура – свойство» использовались стандартные реализации методов множественной линейной регрессии, построения деревьев решений и метода опорных векторов, представленные в MATLAB соответственно функциями *regress*, *treefit* и *svmtrain*, входящими в пакет Statistics Toolbox. Кроме того, использовались

такие стандартные алгоритмы как *clusterdata* и *kmeans*, осуществляющие кластеризацию данных.

Отметим, что в общем случае в предложенном в работе подходе возможно применение других стандартных или нестандартных алгоритмов обучения классифицирующих моделей и кластеризации.

Общая схема работы модуля построения моделей «структура – свойство» представлена на **рисунке 11**.

Рисунок 11. Схема работы модуля построения моделей «структура – свойство»

Схема работы модуля прогнозирования представлена на **рисунке 12**.

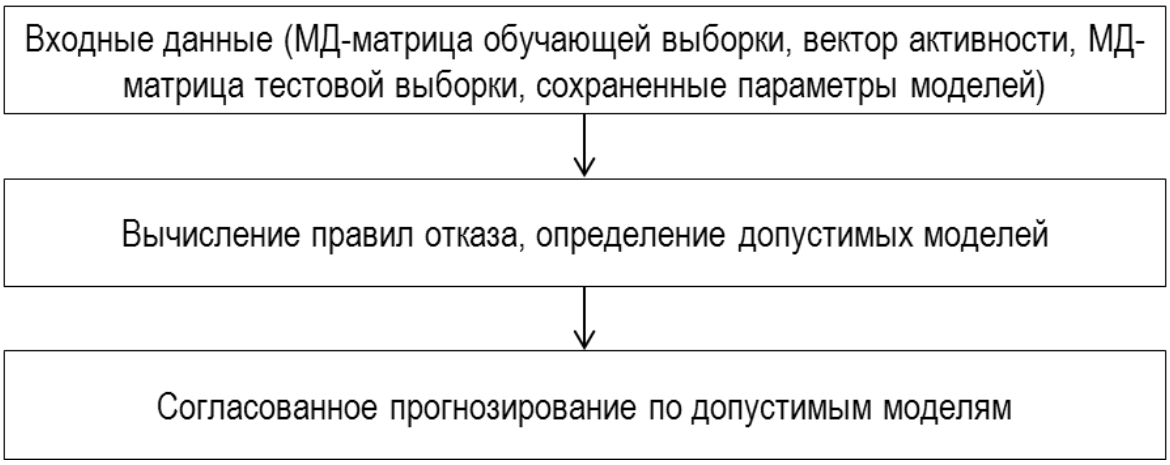


Рисунок 12. Схема работы модуля прогнозирования

Основной функционал программного комплекса реализуют следующие модули-функции.

1) model_regress.m

Выполняет скользящий контроль на данной МД-матрице с заданным вектором активности. В качестве прогнозирующей модели используется модель линейной регрессии. Функция получает в качестве параметров МД-матрицу и вектор активности, а возвращает значение показателя качества построенной модели.

2) model_tree.m

Выполняет скользящий контроль на данной МД-матрице с заданным вектором активности. В качестве прогнозирующей модели используется модель дерева решений. Функция получает в качестве параметров МД-матрицу и вектор активности, а возвращает значение показателя качества построенной модели.

3) model_svm.m

Выполняет скользящий контроль на данной МД-матрице с заданным вектором активности. В качестве прогнозирующей модели используется метод опорных векторов. Функция получает в качестве параметров МД-

матрицу и вектор активности, а возвращает значение показателя качества построенной модели.

4) features_selection.m

Функция осуществляет построение модели заданным методом с использованием эволюционного отбора дескрипторов. Функция получает в качестве параметров МД-матрицу, вектор активности, мощность селекций, число итераций эволюционного отбора и название метода построения модели, возвращает значение показателя качества построенной модели, сокращённую МД-матрицу, содержащую только отобранные дескрипторы, а также массив с индексами отобранных дескрипторов в исходной МД-матрице.

5) twophase_scheme.m

Функция осуществляет построение модели и ограничений допустимости для нее с помощью двухфазной схемы решения. Для задач классификации первого и второго уровня используется заданный метод машинного обучения и независимый эволюционный отбор дескрипторов. Функция получает в качестве параметров МД-матрицу, вектор активности, мощность селекций, число итераций эволюционного отбора и название метода построения модели, возвращает значение показателя качества построенной модели.

6) predict.m

Функция осуществляет прогнозирование активности соединений тестовой выборки, представленной своей МД-матрицей. Функция получает в качестве параметров МД-матрицу тестовой выборки, МД-матрицу обучающей выборки, вектор активности, набор параметров построения модели, включая индексы отобранных дескрипторов, и название метода построения модели, возвращает вектор активности, соответствующий соединениям, представленным в тестовой выборке.

В следующих далее разделах излагаются результаты исследований, проведенных с помощью описанного программного комплекса. Изложение результатов содержит в частности значения параметров основных модулей и описание других особенностей методологий, применяемых для обработки конкретных выборок химических соединений.

3.2 Прогнозирование противоопухолевой активности гликозидов

Выборка из 76 веществ класса гликозидов извлечена из Базы данных по противоопухолевым веществам НИИ ЭДиТО ГУ РОНЦ им. Н.Н. Блохина РАМН, 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 [59]. Имеются также количественные данные по результатам изучения общей токсичности веществ на лабораторных животных.

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

Противоопухолевая активность всех веществ была изучена в опытах на экспериментальных животных со следующими перевиваемыми опухолями:

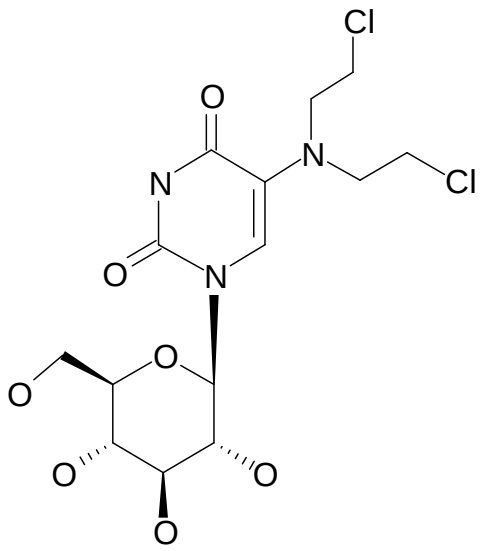
- мышиные лейкозы La, Lel, LZ, L321, L996, P388, L1210, лейкоз Мазуренко,
- солидные и асцитные опухоли мышей лимфома AKR, плазмацитома МОРС-406, опухоли NK/Ly, ЛИО-1, рак легкого РЛ-67, рак легкого LLC,

рак толстой кишки АКАТОЛ, рак шейки матки РШМ-5, плоскоклеточный рак желудка ПРЖ, опухоль желудка ОЖ-5, гепатома 22, рак молочной железы Ca755, карцинома НК, меланома S-91 (Cloudman), меланома Harding-Passey, меланома B16, Эрлиха опухоль, саркома 180, саркома 298,

- крысиные лейкозы Shay, L37, ИЛК,
- солидные и асцитные опухоли крыс альвеолярный рак печени RS-1, гепатома Zajdela, карцинома Герена, карциносаркома Walker-256, саркома 45, саркома М-1, саркома Таращанской, лимфосаркома Плисса.

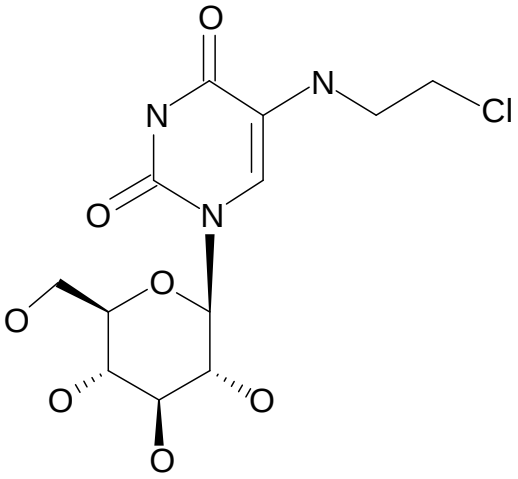
Исследование противоопухолевой активности проводили по стандартной методике, описанной в [56]. Вещество считали обладающим противоопухолевой активностью в случае, если оно ингибировало рост перевиваемой опухоли не менее, чем на 50%, по сравнению с не леченым контролем и/или увеличивало продолжительность жизни леченых животных не менее, чем на 25% по сравнению с не леченными. При меньших эффектах или их отсутствии вещество считалось не обладающим противоопухолевой активностью.

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



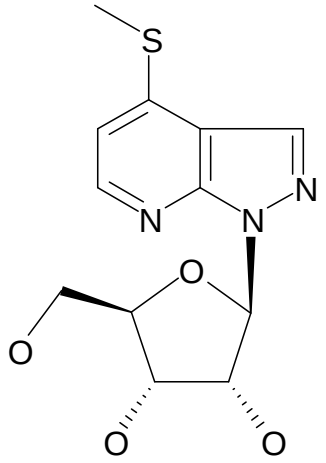
5-бис(2-хлорэтил)аминоурацил, 1-бета-D-глюкопиранозил

(обладает противоопухолевой активностью)



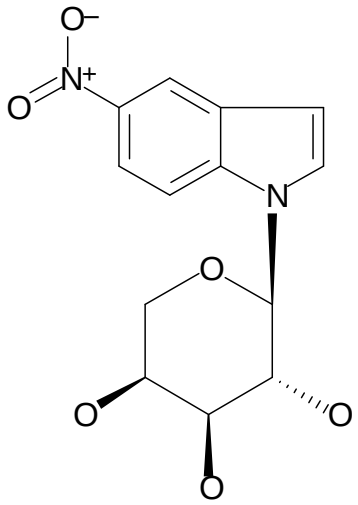
1-бета-D-глюкопиранозил-5-(2-хлорэтил)аминоурацил

(не обладает противоопухолевой активностью)



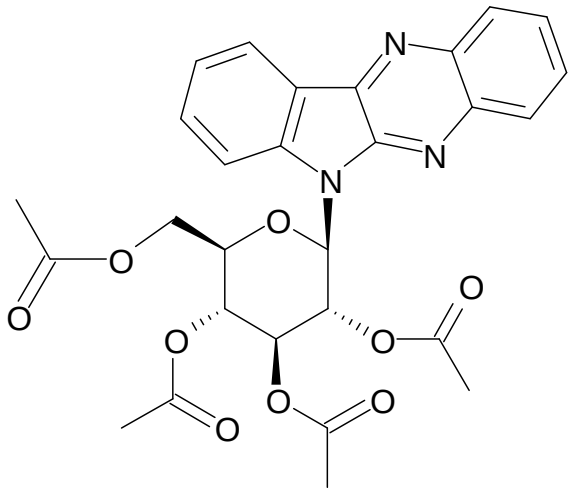
1-(бета-D-рибофуранозил)-4-метилмеркаптопиразоло[3,4-b]пиридин

(не обладает противоопухолевой активностью)



1-(альфа-L-арабинопиранозил)-5-нитроиндол

(обладает противоопухолевой активностью)



1-(2,3,4,6-тетра-O-ацетил-бета-D-глюкопиранозил)индоло[2,3-b]хиноксалин

(не обладает противоопухолевой активностью)

Рисунок 13. 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 неуглеводного 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). Далее приводятся результаты работы методов на первых 10 матрицах.

В качестве методов адаптации описывающего отображения выступали отбор значимых дескрипторов (значимыми считались дескрипторы, имеющие ненулевое значение хотя бы для 5% соединений выборки) и эволюционный отбор дескрипторов параметрами $sel = 3,\ nevol = 10$ (параметры эволюционного отбора описаны в **подразделе 2.2.2**).

В **таблице 2** приведена общая информация по количеству использованных дескрипторов, числу значимых дескрипторов и качеству классификации методами регрессии и SVM.

Таблица 2. Прогнозирование без использования правил отказа

Матрица	Число дескрипторов	Значимых дескрипторов	Регрессия	SVM
1	2100	546	0,8421	0,9737
2	2604	557	0,8684	0,947368
3	2625	819	0,8553	0,934211
4	3255	831	0,8684	0,947368
5	2100	546	0,8421	0,828947
6	2604	774	0,8553	0,947368
7	2625	819	0,8684	0,868421
8	3255	1049	0,8684	0,973684
9	2100	409	0,8684	0,973684
10	2604	546	0,8553	0,8289

Для построения нечеткого классификатора использовалась стандартная реализация алгоритма иерархической кластеризации в среде MATLAB – функция *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 выступала линейная регрессия. **Таблица 3** показывает преимущества предложенного подхода по сравнению с регрессией для предсказания данной химической активности.

Таблица 3. Результат работы нечёткого классификатора на выборке гликозидов

Матрица	Регрессия	Нечеткий классификатор	Отказы
1	0,8421	0,9231	11
2	0,8684	0,9275	7
3	0,8553	0,9868	0
4	0,8684	0,9737	0
5	0,8421	0,9677	14
6	0,8553	0,971	7
7	0,8684	0,9545	32
8	0,8684	0,9868	0
9	0,8684	1	1
10	0,8553	1	5

Ниже (**Таблица 4**) представлены результаты прогнозирования с использованием двухфазной схемы на базе классификатора SVM (при использовании SVM применялся Multilayer Perceptron kernel [62]). Как и ранее, столбцы D1 и D2 содержат количество дескрипторов, отобранных для решения задач классификации первого и второго уровня соответственно.

Таблица 4. Прогнозирование активности гликозидов с использованием двухфазной схемы

Матрица	φ_1	φ_2	Отказы	φ_0	D1	D2
1	0,9737	0,9868	3	1	1	1
2	0,947368	1	4	1	1	3
3	0,934211	0,960526	2	0,959459	1	1
4	0,947368	0,986842	5	1	3	1
5	0,828947	0,921053	13	0,952381	1	3
6	0,947368	1	4	1	1	3
7	0,868421	0,921053	8	0,941176	2	4
8	0,973684	1	2	1	4	2
9	0,973684	1	2	1	1	4
10	0,828947	0,973684	15	1	1	4

В **таблице 5** приведено сравнение качество прогнозирования при использовании нечеткого классификатора, двухфазной схемы на базе регрессии и двухфазной схемы на базе метода опорных векторов. В скобках указано количество отказов от прогноза.

Таблица 5. Сравнение различных подходов к классификации с отказами.

Матрица	2Ф (Регрессия)	2Ф (SVM)	Нечеткий классификатор
1	0,876712 (3)	1 (3)	0,9231 (11)
2	0,8684 (0)	1 (4)	0,9275 (7)
3	0,878378 (2)	0,959459 (2)	0,9868 (0)
4	0,88 (1)	1 (5)	0,9737 (0)
5	0,8421 (0)	0,952381 (13)	0,9677 (14)
6	0,8904 (3)	1 (4)	0,9710 (7)
7	0,8684 (0)	0,941176 (8)	0,9545 (32)
8	0,88 (1)	1 (2)	0,9868 (0)
9	0,8684 (0)	1 (2)	1 (1)
10	0,8667 (1)	1 (15)	1 (5)

3.3 Прогнозирование противоопухолевой активности соединений разных химических классов

Подробные результаты данного исследования публиковались в [57], а также обсуждались на XX российском национальном конгрессе Человек и Лекарство [58]. Для анализа был представлен набор химических соединений, состоящий из двух групп: обучающая выборка (соединения с известной активностью) и тестовые соединения (соединения, для которых активность предполагается неизвестной). Обучающая выборка включает 359 активных соединений и 423 неактивных соединения, количество тестовых соединений 101.

Целью работы являлась оценка применимости методов поиска количественных соотношений «структура – свойства» для прогнозирования проти-

воопухолевой активности. В эксперименте принимали участие исследователи-исполнители (реализующие моделирование структура – свойство) и контролирующая организация (РОНЦ), которая обладала информацией об активности тестовых соединений.

Рассматривались две основные задачи.

Первая, наиболее общая, состоит в том, чтобы построить по обучающей выборке модели «структура – свойство», а затем осуществить с помощью набора этих моделей согласованный прогноз активности тестовых соединений. При этом в качестве результирующих данных выступает набор значений активности для всех тестовых соединений. Контролирующая организация имеет возможность оценить расхождение между предсказанной активностью и действительными значениями активности, которыми она располагает. Целью исследователей является минимизировать такое расхождение. Данную задачу далее в тексте будем обозначать как общую задачу прогноза.

Другая задача носит более практический характер, и состоит в следующем. По ответам построенных моделей исследователь готовит список наиболее перспективных соединений, которые с наибольшей вероятностью обладают изучаемой активностью. Эта задача напрямую связана, во-первых, с задачей виртуального скрининга, а во-вторых с задачей моделирования новых соединений, обладающих заданным свойством (методы прогнозирования могут дать первичную оценку успешности моделирования и показать перспективы дальнейшего синтеза соединения).

При решении второй задачи исследователь направляет в качестве результирующих данных список предполагаемо активных соединений – своего рода предложений (кандидатов). При этом контролирующая организация оценивает, сколько из присланных соединений действительно обладают рассматриваемой активностью. Целью исследователя в данном случае является

минимизация ошибочных предложений. Такую задачу далее в тексте будем обозначать как специальную задачу прогноза.

В данном случае обе задачи решались практически одинаково, и их решение разбивалось на следующие этапы.

Традиционно первым этапом анализа обучающей выборки является её описание в виде векторов дескрипторов, описывающих структуру химических соединений. В данной работе на этапе описания использовались фрагментарные дескрипторы на базе цепочек атомов с различными маркерами химических связей между атомами. Из-за того, что для описания могут быть использованы различные наборы параметров алгоритмов, на выходе в настоящем исследовании было предложено 24 различных описания обучающей выборки.

Каждое такое описание задает две МД-матрицы, первая из которых соответствует обучающей выборке, а вторая описанию тестовых соединений. Каждая матрица содержит по строкам векторы дескрипторов химических соединений. Таким образом, число строк в матрицах равно числу соединений в соответствующих наборах, то есть 782 для первой матрицы и 101 для второй.

Далее, по каждой матрице, соответствующей обучающей выборке, можно построить прогнозирующую модель одним из методов машинного обучения (так как известны значения активности) и оценить её качество. В качестве оценки качества моделей в настоящей работе применялся коэффициент скользящего контроля R^2_{cv} [10]. Отметим, что по одной МД-матрице строилось множество различных прогнозирующих моделей, соответствующих различным параметрам эволюционного отбора дескрипторов, различным методам машинного обучения и различным наборам параметров этих методов.

Теперь на вход каждой построенной модели подавалась матрица, содержащая описание соединений с неизвестной активностью, а на выходе получался прогноз активности каждого из этих соединений.

Таким образом, в нашем случае имелся набор моделей, для каждой из которых известен показатель качества и вектор прогнозов активности тестовых соединений. Длина указанного вектора соответствует количеству тестовых соединений и равна 101. Значения активности принимают значения «1» и «–1», соответствующие активным и неактивным соединениям.

Далее проводился предварительный отбор распознающих моделей для конечного анализа.

Предметом исследований, помимо прочего, являлось также сравнение методов согласованного прогноза по множеству распознающих моделей.

В качестве методов согласованного прогноза рассматривались следующие: методы простого, взвешенного голосования и голосования сильнейших, метод положительных оценок, метод победителя, а также вероятностная оценка активности (подробнее о методах согласованного прогнозирования в **разделе 2.1**).

Таким образом, контролирующей организации отправлялись различные решения рассматриваемых задач (общей и специальной задачи прогнозирования) с целью сравнения указанных методологий по эффективности и сравнения различных наборов моделей, а также эффективности рассматриваемых алгоритмов машинного обучения.

Обучающая часть выборки содержала 359 активных и 423 неактивных соединений. Прогноз осуществлялся для 101 соединения контрольной выборки.

Структурные формулы и данные о противоопухолевой активности извлечены из корпоративной Базы данных по противоопухолевым веществам РОНЦ им. Н.Н. Блохина РАМН [59, 60, 61].

Использованная выборка веществ была неоднородной по химической структуре. В нее входили вещества из таких классов, как азиридины, алкалоиды, азулены, акридины, амиды карбоновых кислот, производные аминокислот, производные аскорбиновой кислоты, аценафтены, различные галоидалкиламины, преимущественно хлорэтиламины, гликозиды, изатины, имидазолы, индазолы, кремнийорганические соединения, органические соединения металлов, лактоны, морфолины, нитрозосоединения, нуклеозиды, пептиды, пиразолы, пирролы, полисахариды, спирты, стероидные гликозиды, тиосемикарбазолы, триазины, триазолы.

На **рисунке 14** представлены примеры неактивных и активных соединений из контрольной выборки.

Неактивные Активные

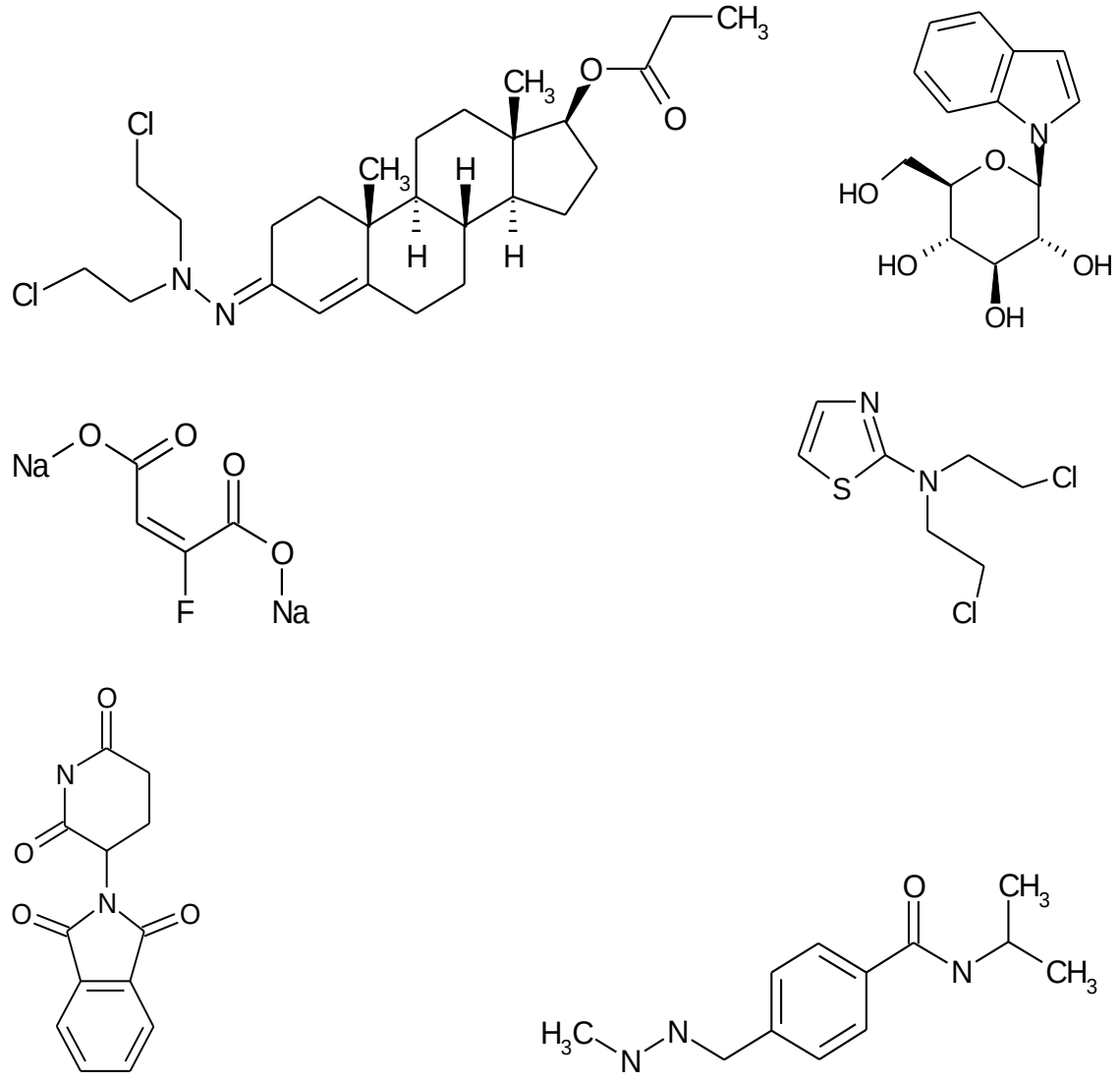


Рисунок 14. Структуры, предложенные для анализа

В ходе проводимого эксперимента было сделано предположение о том, что модели с низким качеством прогноза (R^2_{cv}<0,7) не стоит учитывать при согласованном прогнозе активности тестовых соединений.

Таким образом, сначала были получены результаты для моделей, использующих метод опорных векторов в качестве алгоритма машинного обучения на этапе поиска функциональной зависимости, так как именно с помо-

щью метода опорных векторов удалось построить наиболее качественные модели.

Параметрами описания выступали следующие величины: длина линейных фрагментов (варьировалась от 2 до 4) и маркеры, участвующие в описании (d –степень вершины молекулярного графа, b – информация о наличии химических связей, r – положение в кольце). При использовании метода опорных векторов длина линейных фрагментов была фиксирована и равнялась 2. Это связано с тем, что при использовании более длинных линейных фрагментов нарастает мощность описания (количество дескрипторов), что влечет за собой рост времени работы алгоритма эволюционного отбора, который в случае использования метода опорных векторов занимает весьма продолжительное время.

При получении результатов была использована реализации алгоритма SVM в системе MATLAB с ядром (kernel) mlp (Multilayer Perceptron kernel [62]). Параметрами эволюционного отбора дескрипторов выступало количество отбираемых селекций (в нашем случае равнялось 3) и число итераций алгоритма отбора (не превосходило 10). Ниже приведена **таблица 6**, содержащая параметры описания выборки, количество отобранных дескрипторов и показатель качества модели.

Таблица 6. Построенные SVM-модели

Параметры описания	R^2_{cv}	Дескрипторы
[___]	0,829923	4
[__r]	0,961637	1
[_b_]	0,741688	2
[d__]	0,892583	2
[_br]	0,966752	1
[d_r]	0,970588	1
[db_]	0,956522	2
[dbr]	0,970588	1

Как видно из **таблицы 6**, для большинства описаний обучающей выборки удалось построить модели с качеством прогноза выше 0,9, что говорит об их высокой прогностической способности. Однако эксперимент показал, что использование таких моделей не ведет к успешному прогнозированию активности тестовых соединений.

В **таблицах 7** и **8** ниже приведено качество прогноза активности тестовых соединений при использовании различных методологий согласованного прогнозирования. В **таблице 7** дано количество верных прогнозов при решении общей задачи прогнозирования. В **таблице 8** указан процент активных соединений среди набора предложенных кандидатов при решении специальной задачи прогнозирования. В данных таблицах приняты следующие обозначения:

- Vote – метод голосования;
- PosVote – метод положительных оценок для простого голосования;
- WeightVote – метод взвешенного голосования;
- PosWeightVote – метод положительных оценок для взвешенного голосования;
- 1LidVote – метод победителя;
- 3LidVote – метод голосования сильнейших (учитываются 3 лучших модели);
- ProbVote – выход из вероятностной оценки активности на базе метода положительных оценок.

Таблица 7. Количество успешных прогнозов на тестовой выборке (общая задача прогнозирования)

Vote	WeightVote	1LidVote	3LidVote	ProbeVote
45	45	38	42	45

Таблица 8. Процент успешных прогнозов на тестовой выборке (специальная задача прогнозирования)

PosWeightVote	55,56%
PosVote	60%

Всего, напомним, в контрольной выборке находилось 101 соединение (из которых 40 обладают исследуемой активностью и 61 неактивно). Таким образом, использование качественных на обучающей выборке моделей SVM не показало эффективности при предсказании активности соединений из тестовой выборки.

Скорее всего, это связано с тем, что, несмотря на использование показателя качества со скользящим контролем, в ходе построения распознающей модели всё равно происходит переобучение (устанавливаются зависимости «структура – свойство», характерные именно для данной обучающей выборки, а не для всех химических соединений). Это происходит по причине применения эволюционного отбора дескрипторов, который уже использует значение качества со скользящим контролем. Таким образом, отбираются дескрипторы, оптимизирующие значение показателя качества именно для данной выборки.

Выходом из ситуации может стать использование для оценки качества распознающих моделей процедуры двойного скользящего контроля (при удалении каждого нового соединения из обучающей выборки эволюционный отбор дескрипторов происходит заново), однако, эта процедура увеличит время работы алгоритма пропорционально количеству соединений в обучающей выборке, что является огромным минусом для решения практических задач.

В данном исследовании рассмотрен другой подход – использование для прогнозирования активности соединений тестовой выборки менее качественных моделей или включение в процесс согласованного прогноза всех моде-

лей. Ниже в **таблице 9** представлены результаты использования различных наборов распознающих моделей.

Таблица 9. Процент успешных прогнозов на тестовой выборке разными наборами моделей

Модель	SVM	KNN	REGRESS	ALL	KNN+REGRESS
Общая задача прогнозирования	44,55%	64,36%	60%	62,38%	62,38%
Специальная задача прогнозирования	60%	50%	68,75%	70%	71,43%

Как видно из **таблицы 9** лучшие результаты наблюдаются именно при использовании моделей низкого качества (KNN при решении общей задачи прогнозирования и KNN совместно с регрессией при решении специальной задачи прогнозирования). При этом согласованное прогнозирование по множеству всех моделей также является приемлемым.

Отметим также перспективы использования регрессионных моделей. В **таблице 10** приведено качество построенных регрессионных моделей. Здесь так же, как и для метода опорных векторов, использовались описания с длиной линейного фрагмента, равной 2. Параметры эволюционного отбора также оставались прежними: 3 отбираемых селекций и ограничение на число итераций 10.

Таблица 10. Регрессионные модели

Параметры описания	R^2_{cv}	Дескрипторы
[___]	0,648338	5
[__r]	0,663683	6
[_b_]	0,647059	6
[_br]	0,641944	10
[d__]	0,664962	8
[d_r]	0,681586	10
[db_]	0,689258	10
[dbr]	0,682864	10

Несмотря на то, что качество прогнозирования регрессионных моделей на обучающей выборке значительно уступает качеству более «продвинутых» распознающих моделей, качество согласованного прогноза при использовании только регрессионных моделей сравнимо с аналогичным показателем при использовании всех моделей. Это говорит о том, что регрессионные модели при грамотном использовании могут составлять конкуренцию более «продвинутым» алгоритмам при поиске потенциально активных соединений.

Также важной характеристикой рассматриваемых подходов к прогнозированию активности неизученных соединений является количество ложноположительных и ложноотрицательных прогнозов. Важно понять, чего в дальнейших экспериментальных исследованиях веществ, отобранных с помощью таких подходов как перспективные, будет больше – потерь времени и средств на исследование неактивных веществ, которые были спрогнозированы как активные, или потери заведомо активных соединений, которые модели характеризуют как неперспективные, и они из исследований отбрасываются еще до экспериментального изучения. Вторая ситуация также является негативной, так как если таких исключений будет много, маловероятно обнаружение действительно эффективных лекарств при скрининге.

Таблица 11 содержит проценты ложноположительных и ложноотрицательных прогнозов от общего числа соединений в контрольной выборке.

Третий столбец таблицы характеризует качество прогноза на тестовой выборке (как и ранее, для общей задачи прогнозирования – процент верно предсказанных соединений тестовой выборки, для специальной задачи прогнозирования – процент активных соединений среди набора предложенных). Таким образом, для строк, соответствующих решению общей задачи прогнозирования, сумма в трех последних столбцах равна 100%, для строк, соответствующих решению специальной задачи прогнозирования, это свойство не выполняется.

Таблица 11. Процент ложноположительных и ложноотрицательных прогнозов при использовании различных наборов моделей

Модель	Тип задачи	Качество	Ложноположительные	Ложноотрицательные
SVM	Общая	44,55%	36,63%	18,81%
	Специальная	60,00%	5,94%	30,69%
KNN	Общая	64,36%	11,88%	23,76%
	Специальная	50,00%	10,89%	28,71%
REGRESS	Общая	59,41%	17,82%	22,77%
	Специальная	68,75%	4,95%	28,71%
ALL	Общая	62,38%	16,83%	20,79%
	Специальная	70,00%	2,97%	32,67%
KNN+REGRESS	Общая	62,38%	14,85%	22,77%
	Специальная	71,43%	3,96%	29,70%

Как видим, методология решения специальной задачи прогнозирования подразумевает минимизацию ложноположительных прогнозов, минимальное значение достигается при использовании согласованного прогноза по всем построенным моделям. Минимум же ложноотрицательных прогнозов достигается при решении общей задачи прогнозирования с помощью моделей высокого качества (на базе метода опорных векторов).

Приведем сравнение полученных результатов с прогнозами, рассчитанными с помощью системы прогнозирования биологической активности PASS (Prediction of Activity Spectra for Substances) [63], разработанной в НИИ биомедицинской химии им. В.Н. Ореховича РАМН. В настоящее время система предсказывает более тысячи видов важнейших биологических активностей по структурной формуле химического вещества, включая основные и побочные фармакологические эффекты, механизмы действия, мутагенность, канцерогенность, тератогенность и эмбриотоксичность. Обучающая выборка в системе PASS содержит огромное количество соединений (сотни тысяч) и постоянно пополняется новой информацией о биологически активных веществах, отбираемой как из публикаций в научно-технической литературе, так и из многочисленных баз данных. В качестве описания структур PASS использует оригинальные MNA дескрипторы (Mulilevel Neighbourhoods of Atoms). В качестве метода поиска функциональной зависимости для PASS предложено использование различных статистических методов [63, 64].

Таблица 12. Сравнение использованных подходов с результатами работы системы PASS

Прогнозатор	Качество	Ложноположительные	Ложноотрицательные
KNN	64,36%	11,88%	23,76%
ALL	62,38%	16,83%	20,79%
PASS	51,49%	21,78%	26,73%

Из **таблицы 12** заключаем, что предложенный в работе подход на данной контрольной выборке соединений демонстрирует результаты лучшие, нежели система PASS, как по числу верно предсказанных соединений, так и по количеству ложноположительных и ложноотрицательных прогнозов.

Обсуждение результатов

Как уже было изложено выше, на данной обучающей выборке удалось построить довольно качественные на обучающей выборке модели, однако,

использование их в явном виде для предсказания активности неизученных соединений не так эффективно. Среди причин можно выделить основные:

1. переобучение модели вследствие использования эволюционного отбора дескрипторов (эволюционный отбор происходил по значениям качества со скользящим контролем, таким образом, отбирались оптимизирующие это значение дескрипторы, однако, далеко не факт что они же лучшим образом оптимизируют качество на контрольной выборке);
2. неоднородность обучающей и контрольной выборок (в выборках находились вещества нескольких десятков химических классов, что затрудняло построение единой прогнозирующей модели для всех веществ).

Первая причина была обсуждена выше. Здесь же сосредоточимся на рассмотрении методов преодоления второй причины.

Чисто методологически гипотеза об улучшение качества прогнозирования при увеличении степени однородности выборки будет проверена на соответствующих выборках соединений одного класса, которые будут подготовлены в РОНЦ им. Н.Н. Блохина РАН в ближайшее время. Помимо этого рассмотрим несколько подходов к анализу неоднородных выборок.

Ниже в **таблице 11** и **таблице 12** показано число отказов при использовании метода, основанного на минимальных и максимальных значениях используемых дескрипторов при прогнозировании некоторыми моделями, построенными в ходе проведенного исследования. Столбец «Отказы1» содержит количество отказов при использовании эволюционно отобранных дескрипторов, столбец «Отказы2» -– всех дескрипторов. Столбцы «Дескрипторы1» и «Дескрипторы2» содержат соответствующее число дескрипторов, использованных для расчета количества отказов. По понятным причинам значения двух последних столбцов **таблиц 13** и **14** совпадают.

Таблица 13. Отказы при использовании SVM

Параметры описания	Дескрипторы1	Отказы1	Дескрипторы2	Отказы2
[___]	4	2	40	2
[__r]	1	0	104	3
[_b_]	2	1	89	3
[_br]	1	0	223	5
[d__]	2	1	114	3
[d_r]	1	0	254	4
[db_]	2	1	190	4
[dbr]	1	0	425	4

Таблица 14. Отказы при использовании REGRESS

Параметры описания	Дескрипторы1	Отказы1	Дескрипторы2	Отказы2
[___]	5	2	40	2
[__r]	6	2	104	3
[_b_]	6	0	89	3
[_br]	10	2	223	5
[d__]	8	1	114	3
[d_r]	10	1	254	4
[db_]	10	0	190	4
[dbr]	10	0	425	4

Рассмотренные ограничения допустимости могут быть полезны при скрининге больших баз соединений, но в проведенном исследовании они не влияют принципиально на результат, так как из таблиц видно, что отказы от прогнозирования единичны. Кроме того, они не носят системного характера (различные соединения попадают в отказы по одной модели, но другими предсказываются).

Рассмотрим теперь правила отказа на базе кластерной структуры обучающей выборки. Ниже приведен пример кластеризации алгоритмом k-means. В качестве метрики использовались полученные при эволюционном отборе на базе SVM дескрипторы. Также приведено оригинальное разбиение на

классы активных и неактивных соединений и разделяющие «плоскости», построенные SVM-алгоритмом.

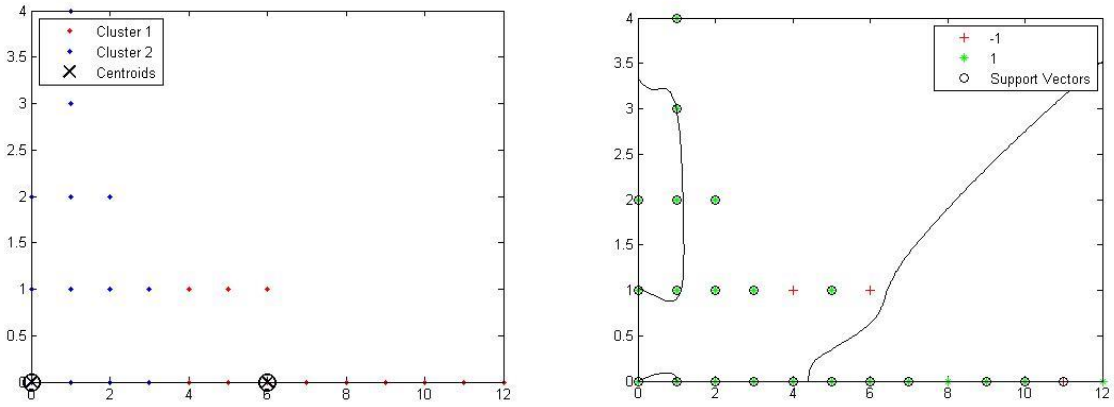


Рисунок 15. Кластеризация и SVM-mlp, параметры описания [_b_]

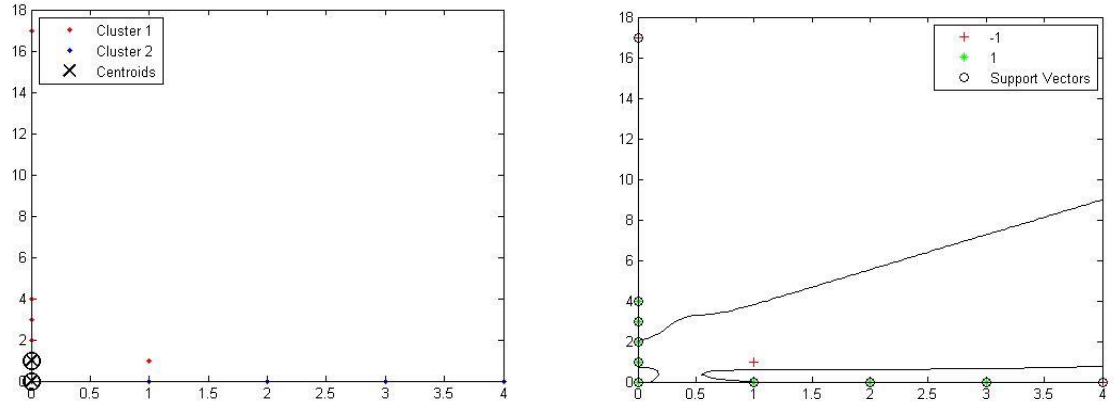


Рисунок 16. Кластеризация и SVM-mlp, параметры описания [db_]

Для SVM-моделей улучшение качества за счет кластеризации не наблюдается. И, анализируя кластерную структуру на **рисунке 15** и **рисунке 16**, можно сделать заключение о том, что правила отказа на базе таких кластеров не будут принципиально отличаться от приведенных ранее правил на основе максимальных значений дескрипторов.

В случае использования регрессионных моделей (число дескрипторов 5 – 10) кластерная структура ещё менее наглядна. Областей для построения локальных моделей в данном случае обнаружено не было.

3.4 Прогнозирование способности ингибировать активность поли-(АДФ-рибоза)-полимеразы-1

Расчеты проведены для ингибиторов поли-(АДФ-рибоза)-полимеразы-1 (ПАРП) – это фермент, который локализуется в клеточном ядре и катализирует поли-АДФ-рибозилирование различных белков [65]. Ингибирование PARP [65], необходимое для подавления механизмов репарации и выживания клеток при химио- и радио- терапии , рассматривается как многообещающая стратегия лечения различных видов рака. Моделирование биологической активности в рамках решения задачи «структура – свойство» представляется на сегодняшний день актуальной задачей молекулярной биологии.

В качестве исходных данных была взята выборка химических соединений, для которой рассматривалась активность – ингибирование фермента деления клеток. В выборке представлено 120 соединений с известной активностью – 86 активных и 34 неактивных. Также представлены 196 молекул с неизвестной активностью.

Задача заключалась в том, чтобы 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 [9].

Регулируемые *параметры описания*:

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

2) маркеры, участвующие в описании (d –степень вершины молекулярного графа, b – информация о наличии химических связей, r – положение в кольце).

В соответствии с выбором параметров построено 24 матрицы «молекулярный граф – дескриптор» – 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.

Для оценки качества моделей использовался показатель качества со скользящим контролем [10].

Ниже в таблицах приведены результаты использования двухфазной схемы решения на регрессионных моделях и моделях на базе метода опорных векторов (SVM) [40]. При использовании SVM применялся Multilayer Perceptron kernel [61]. В таблицах ниже длина линейных фрагментов на этапе описания была фиксирована и равнялась 2. Также для задач классификации первого и второго уровня эволюционный отбор дескрипторов применялся независимо. В таблицах ниже столбцы D1 и D2 содержат количество дескрипторов, отобранных для решения задач классификации первого и второго уровня соответственно.

Таблица 15. Качество прогноза двухфазной схемы решения с использованием метода опорных векторов

Маркеры	φ_1	φ_2	Отказы	φ_0	D1	D2
***	0,941667	1	7	1	2	1
**r	0,891667	0,883333	1	0,890756	5	1
b	0,908333	0,933333	13	0,971963	3	3
*br	0,941667	0,925	9	0,981982	4	1
d**	0,883333	0,891667	9	0,918919	1	1
d*r	0,875	0,875	8	0,910714	3	1
db*	0,908333	0,925	8	0,946429	2	1
dbr	0,966667	0,925	9	0,981982	4	1

Таблица 16. Качество прогноза двухфазной схемы решения с использованием регрессии

Маркеры	φ_1	φ_2	Отказы	φ_0	D1	D2
***	0,841667	0,841667	0	0,841667	3	1
**r	0,808333	0,816667	1	0,815126	4	2
b	0,916667	0,916667	0	0,916667	8	1
*br	0,95	0,958333	1	0,957983	7	2
d**	0,9	0,9	0	0,9	7	1
d*r	0,758333	0,908333	24	0,916667	2	5
db*	0,933333	0,933333	0	0,933333	6	1
dbr	0,95	0,95	0	0,95	5	1

Построенные модели прогнозирования были применены к выборке соединений с неизвестными значениями активности. Результаты были объединены в сводную таблицу и прогноз активности соединения оценивался по сумме значений активности, полученных на каждой модели. Напомним, что допустимые значения активности в нашем случае «+1» и «–1», соответственно, активное и неактивное соединение.

В результате применения построенных моделей к имеющейся выборке из 196 соединений с неизвестным значением активности можно выделить несколько соединений, предположительно, обладающих активностью.

Наиболее перспективные с точки зрения построенных моделей соединения прошли экспериментальные испытания, результатом которых являлась оценка ингибирующей концентрации. Одно из соединений (на **рисунке 17**) показало активность. Уточнение результатов дало ингибирующую концентрацию на данном соединении в 200 микроМ (0,2 мМ), то есть соединение является достаточно эффективным ингибитором ПАРП.

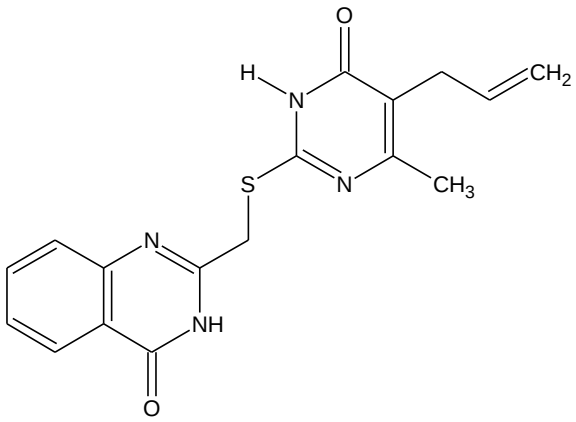


Рисунок 17. Структура соединения, показавшего активность на экспериментальных испытаниях

Для оценки эффективности разработанного метода было проведено его сравнение с наиболее эффективными на сегодняшний день методами моделирования биологической активности химических соединений. Для этого выборка из 196 соединений была также проанализирована методами молекулярного докинга, реализованными в пакете программ Lead Finder [66]. Было показано, что во многих случаях предсказания QSAR-моделей согласуется с предсказаниями молекулярного докинга [67], однако для нескольких соединений, наиболее перспективных с точки зрения построенных моделей методами молекулярного докинга, методом QSAR было предсказано отсутствие биологической активности. Имела место и обратная ситуация.

Таким образом, прогнозирование активности химических соединений на базе QSAR-моделей позволяет устранить существующие ошибки молекуляр-

ного моделирования и исключить ложноотрицательные предсказания. Из чего следует однозначные выводы:

1) метод моделирования, описанный в настоящей работе, показывает свою эффективность;

2) два основных метода моделирования биологической активности – QSAR и докинг – дополняют друг друга.

3.5 Выводы

Представленные в настоящей главе результаты подтверждают практическую значимость разработанных подходов.

Показана содержательность использования ограничений допустимости и результативность предложенного подхода. Проведена серия практических испытаний, в ходе которых подтверждены теоретические оценки качества прогнозирования с использованием двухфазной схемы решения задачи «структура – свойство».

Предложенные методы позволили получить прогнозирующие модели высокого качества. В большинстве случаев наблюдается значительное улучшение качества прогноза по сравнению с классическими методами.

Описано применение предложенной методологии прогнозирования активности неизученных соединений для практических исследований в области современной органической химии и фармацевтики. Проведено сравнение классических методов виртуального скрининга и предложенных автором методов моделирования «структура – свойство», в ходе чего показана значимость последних.

Заключение

В рамках работы над настоящей диссертацией получены следующие основные результаты.

1. Теоретически обоснован и разработан универсальный подход к прогнозированию свойств М-графов, включающий в себя метод адаптации дескрипторного описания, методы построения ограничений допустимости для моделей «структура – свойство», а также методы согласованного прогнозирования по множествам и семействам моделей «структура – свойство».

2. Разработан нечеткий классификатор на базе кластерной структуры обучающей выборки, а также алгоритм оптимизации параметров кластеризации для выбора локально лучшей модели в некотором классе классифицирующих функций.

3. Разработана двухфазная схема решения задачи «структура – свойство». Получены теоретические условия, при которых двухфазная схема гарантирует улучшение качества прогноза на обучающей выборке за счет осуществления отказа от прогноза. Приведены необходимые оценки вычислительной сложности.

4. Разработаны методы независимой адаптации описывающих отображений под задачи классификации и задачи построения ограничений допустимости, что позволяет с одной стороны повысить качество моделей «структура – свойство», а с другой – снизить вычислительную сложность прогнозирования свойств неизученных М-графов.

5. Предложен комплексный метод оптимизации сложности дескрипторного описания обучающей выборки, понижающий вычислительную сложность обработки неоднородных выборок М-графов.

6. Создана программная реализация разработанных алгоритмов в среде MATLAB. В серии вычислительных экспериментов подтверждены получен-

ные теоретические результаты, в том числе оценки качества прогнозирования с использованием двухфазной схемы решения, а также оценки вычислительной сложности.

7. Проведены совместные прикладные исследования с учеными из Института Органической Химии им. Н.Д. Зелинского РАН и Российского Онкологического Научного Центра им. Н.Н. Блохина РАМН по прогнозированию активности на реальных выборках химических соединений, которые показали практическую значимость предложенных в работе подходов и перспективность описанных методов.

8. Построены модели для прогнозирования наличия у химических соединений различных классов противоопухолевой активности и способности ингибировать активность поли-(АДФ-рибоза)-полимеразы-1.

В качестве перспектив развития предложенного подхода можно отметить два направления.

1. Как показано в **разделе 2.4.5**, двухфазная схема решения позволяет использовать её для последовательного вычерпывания ошибки при анализе сложных неоднородных выборок молекулярных графов, а также для осуществления многоуровневой классификации. Разработка алгоритмов для данных методов и их тестирование на практике являются направлением будущей работы. Также предстоит провести масштабные исследования возможностей предложенных методов при проведении виртуального скрининга на больших базах соединений с использованием параллельных вычислений.

2. Кроме того, существенной проблемой в настоящий момент является вычислительная сложность методов адаптации дескрипторного описания при построении распознающих моделей. Перспективным является метод частичного эволюционного отбора дескрипторов по случайным подмножествам множества дескрипторов. Разработка такого метода, а также оценка его эффективности также представляют интерес для дальнейших исследований.

Список литературы

1. Gasteiger, Johann (ed.) Handbook of Chemoinformatics. From Data to Knowledge. Wiley-VCH, Weinheim, 2003, in 4 volumes.

2. Varnek A., Tropsha, A. Chemoinformatics Approaches to Virtual Screening, RSCPublishing, 2008.

3. Brown, Frank (2005). «Editorial Opinion: Chemoinformatics – a ten year update». Current Opinion in Drug Discovery & Development 8 (3): 296–302.

4. Н.И. Жохова, И.И. Баскин, А.Н. Зефиров, В.А. Палюлин, Н.С. Зефиров Псевдофрагментные дескрипторы на основе комбинаций свойств атомов во фрагментах в исследованиях количественных соотношений “структура-свойство” при прогнозировании физических свойств полимеров // Докл. АН, сер. химия, 2010, Т.430, N 5, с. 635-638.

5. Kier L.B., Hall L.H. Molecular connectivity in structure-activity analysis. / Wiley, London, 1986.

6. Nantasenamat C., Isarankura-Na-Ayudhya C., Naenna T., Prachayasittikul V. A practical overview of quantitative structure-activity relationship // Excli J. (2009) 8: 74–88.

7 J. Alvarez, B. Shoichet. Virtual Screening in Drug Discovery. — CRC Press, Taylor & Francis Group, 2005.

8 REACH – European Community Regulation on chemicals and their safe use. (URL: http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm [Электронный ресурс] дата обращения 14.06.2012).

9. 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.

10. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, B, 36, pp. 111–147, 1974.

11. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

12. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Множественная регрессия — 3-е изд. — М.: «Диалектика», 2007. — С. 912.

13. Richard O. Duda, Peter E. Hart, David G. Stork Pattern classification (2nd edition), Wiley, – 2001. – New York.

14. Финн В.К. О возможностях формализации правдоподобных рассуждений средствами многозначных логик // Всесоюз. симпозиум по логике и методологии науки.— Киев: Наукова думка, 1976.— С. 82–83.

15. Vert, J-P, Schölkopf B, Tsuda K (2004). *Kernel methods in computational biology*. Cambridge, Mass: MIT Press.

16. Maggiora, G.; Shanmugasundaram, V., Molecular Similarity Measures, in: Bajorath, J. (Ed.), Chemoinformatics, Humana Press, 2004.

17. P. Mahé, L. Ralaivola, V. Stoven, and J-P Vert.The pharmacophore kernel for virtual screening with SVM. J. Chem. Inf. Model., 46(5):2003-2014, 2006.

18. B. Ganter, P.A. Grigoriev, S.O. Kuznetsov, and M.V. Samokhin, Concept-based Data Mining with Scaled Labeled Graphs. In: K.E. Wolff, H. D. Pfeiffer, H. S. Delugach, Eds., Proc. 12th International Conference on Conceptual Structures (ICCS 2004), Lecture Notes in Artificial Intelligence (Springer), Vol. 3127, pp. 94-108, 2004.

19. S.O. Kuznetsov and M.V. Samokhin, Learning Closed Sets of Labeled Graphs for Chemical Applications. In: Proc. 15th Conference on Inductive Logic

Programming (ILP 2005), Lecture Notes in Artificial Intelligence (Springer), Vol.3625, pp.190-208., 2005.

20. V.G. Blinova, D.A. Dobrynin, V.K. Finn, S.O. Kuznetsov, and E.S. Pankratova, Toxicology analysis by means of the JSM-method. Bioinformatics, vol. 19(10), pp. 1201-1207, 2003.

21. 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.

22. Worth, A.P.; Bassan, A.; Gallegos, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. ECB Report EUR 21866 EN, European Commission, Joint Research Centre; Ispra, Italy, 2005; p. 95.

23. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. Altern. Lab. Anim. 2005, 33, 445–459.

24. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O.A. Stepwise approach for defining the applicability domain of SAR and QSAR models. J. Chem. Inf. Model. 2005, 45, 839–49.

25. Sheridan, R.; Feuston, R.P.; Maiorov, V.N.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J. Chem. Inf. Comp. Sci. 2004, 44, 1912–1928.

26. Preparata, F.P.; Shamos, M.I. Convex hulls: Basic Algorithms. In Computational Geometry: An Introduction; Preparata, F.P., Shamos, M.I., Eds.; Springer-Verlag: New York, NY, USA, 1991; pp. 95–148

27. Jouan-Rimbaud, D.; Bouveresse, E.; Massart, D.L.; de Noord O.E. Detection of prediction outliers and inliers in multivariate calibration. Anal. Chim. Acta 1999, 388, 283–301.

28. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Kovalishyn, V.V.; Prokopenko, V.V.; Tetko, I.V. Applicability domain for in silico models to achieve accuracy of experimental measurements J. Chemometrics., 2010, 24(3-4), 202-208.

29. I. Baskin, N. Kireeva and A. Varnek The One-Class Classification Approach to Data Description and to Models Applicability Domain // Molecular Informatics, Volume 29, Issue 8-9, pages 581–587, 2010.

30. Tong, W.; Hong, H.; Fang, H.; Xie, Q. Perkins, R. Decision forest: Combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci. 2003, 43, 525–531.

31. D. Horvath, G. Marcou, A. Varnek Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models // J. Chem. Inf. Mod., 49, 1762–1776 (2009).

32. H.-J. Bohm, G. Schneider. Virtual Screening for Bioactive Molecules. — Wiley-VCH, 2000.

33. Walters WP, Stahl MT, Murcko MA (1998). «Virtual screening – an overview». Drug Discov. Today 3 (4): 160–178.

34. Eckert H, Bajorath J (2007). «Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches». Drug Discov. Today **12** (5-6): 225–33.

35. 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. – 2010. – № 3. – С. 39-45.

36. Прохоров Е. И. «Нечеткое» прогнозирование свойств химических соединений: Использование нечеткой функции классификации на кластерах обучающего множества в задаче «структура – свойство», Saarbrucken, Germany: LAP Lambert Academic Publishing, 2012, – 80 с.

37. Прохоров Е.И. Нейронные сети для построения ограничений допустимости в задаче «структура – свойство» // Нейрокомпьютеры: разработка, применение. – 2012. ¬– № 10. – С. 46–56.

38. R. Todeschini, V. Consonni: Handbook of Molecular Descriptors. WILEY-WCH Publishers, Weinheim, 2000. ISBN 3-527-29913-0

39. Ю. А. Овчинников Биоорганическая химия. — Москва: Просвещение, 1987. — С. 24—26.

40. Vapnik, V. N. The nature of statistical learning theory / V. N. Vapnik. New York; London : Springer, 1998. 189 p.

41. К.В. Воронцов Машинное обучение. Курс лекций (URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов) [Электронный ресурс] дата обращения 14.06.2012).

42. J. Neyman; E. S. Pearson On the Problem of the Most Efficient Tests of Statistical Hypothese // Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 231. (1933), pp. 289-337.

43. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974.

44. Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. MIT Press,

Cambridge, USA, 1998.

45. Shawe-Taylor J., Cristianini N. Robust bounds on generalization from the margin distribution: Tech. Rep. NC2–TR–1998–029: Royal Holloway, University of London, 1998.

46. Cover, Thomas M. and Joy A. Thomas (1991). Elements of Information Theory. New York: Wiley.

47 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.

48. A.V. Bekker, A.A. Suleimanov, G.N. Apryshko, M.I. Kumskov, and R.B. Pugacheva. Multilevel adaptive description of molecular graphs in the “structure-property” problem. *Pattern Recognition and Image Analysis*, 23(1):44–50, 2013.

49. Прохоров Е.И., Беккер А.В., Перевозников А.В., Свитанько И.В., Захаренко А.Л., Суханова М.В., Кумсков М.И. Приложения метода эволюционного отбора дескрипторов в математическом моделировании зависимости биологической активности соединения от его структуры // Прогнозирование свойств химических соединений. Унифицированный Репозиторий моделей «структура – свойство»: – Сборник научных работ. – М.: МАКС Пресс, 2012. – С. 3-24.

50. Химические приложения топологии и теории графов, под ред. Р. Кинга \\ Chemical Applications of Topology and Graph Theory, ed. by R. B. King. — М.: Мир, 1987. — 560 с.

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

52. A. Likas, N. Vlassis, and J. J. Verbeek, The global k-means clustering algorithm, Pattern Recognition, vol. 36, no. 2, pp. 451-461, Feb. 2003.

53. Prokhorov E.I., Ponomareva L.A., Permyakov E.A., Kumskov M.I. Fuzzy classification and fast rejection rules in the structure-property problem // Pattern Recognition and Image Analysis, 2013, Volume 23, Number 1, Pp. 130–138.

54. J. C. Bezdek Pattern Recognition with Fuzzy Objective Function Algoritms, Plenum Press, New York, 1981.

55. Rousseeuw, P. and Leroy, A.: 1996, Robust Regression and Outlier Detection. John Wiley & Sons., 3rd edition.

56. Трещалина Е.М., Жукова О.С., Герасимова Г.К., Андронова Н.В., Гарин А.М. Методические указания по изучению противоопухолевой активности фармакологических веществ // Руководство по экспериментальному (доклиническому) изучению новых фармакологических веществ. – М., 2005. – С. 637–651.

57. Прохоров Е.И., Кумсков М.И., Беккер А.В., Перевозников А.В., Пугачева Р.Б., Апрышко Г.Н. Согласованное прогнозирование противоопухолевой активности по семейству моделей «структура-свойство» // Прогнозирование свойств химических соединений. Унифицированный Репозиторий моделей «структура – свойство»: – Сборник научных работ. – М.: МАКС Пресс, 2012. – С. 25–56.

58. Е.И. Прохоров, Г.Н. Апрышко, Р.Б. Пугачева, А.В. Беккер, А.В. Перевозников, М.И. Кумсков Математические методы прогнозирования противоопухолевой активности // XX российский национальный конгресс Человек и Лекарство: Сборник материалов конгресса. – ЗАО РИЦ Человек и лекарство. – Москва, 2013. – С. 415–415.

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

60. Апрышко Г.Н. Биологическая информация в электронной базе данных по противоопухолевым веществам НИИ ЭДИТО РОНЦ РАМН // Вестник РОНЦ. – 2007. – № 2. – С. 25–31.

61. Апрышко Г.Н. База данных по противоопухолевым веществам НИИ ЭДиТО Онкологического научного центра им. Н.Н. Блохина РАМН. Российский биотерапевтический журнал. – 2008. – № 2. – С. 49–53.

62. Thomas R., Karsten B. Multilayer Perceptron kernel. Proceedings of the 24th SIBGRAPI Conference on Graphics, Patterns and Images, Maceió, Alagoas, Brazil, 2011. P. 337–343.

63. Филимонов Д.А., Поройков В.В. Прогноз спектров биологической активности органических соединений // Российский химический журнал. – 2006. – Т. 50. – № 2. – С 66–75.

64. Filimonov D. A., Zakharov A. V., Lagunin A. A., Poroikov V. V. 'QNA-based 'Star Track' QSAR approach' // SAR and QSAR in Environmental Research. – 2009. – V. 20. – № 7. – P. 679–709.

65. D.D' Amours, S. Desnoyers, I. D'Silva and G. G. Poirier Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions. // Biochem. J. 1999. 342 (Pt 2). 249–268.

66. Stroganov O.V., Novikov F.N., Stroylov V.S., Kulkov V., Chilov G.G. Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening // J Chem Inf Model. 2008 Dec; 48(12):2371–85.

67. Leonid V. Romashov, Alexey A. Zeifman, Alexandra L. Zakharenko, Fedor N. Novikov, Viktor S. Stroilov, Oleg V. Stroganov, Germes G. Chilov, Svetlana N.

Khodyreva, Olga I. Lavrik, Ilya Yu. Titov and Igor V. Svitan'ko. Rational design and synthesis of new PARP1 inhibitors. Mendeleev Communications, 22(1), 15-17 (2012).