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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2014

Работа выполнена на кафедре вычислительной математики механико-математического факультета ФГБОУ ВПО «Московский государственный университет имени М.В. Ломоносова».

Научный руководитель:	Кумсков Михаил Иванович доктор физико-математических наук, профессор
Официальные оппоненты:	Кузнецов Сергей Олегович доктор физико-математических наук, профессор, заведующий отделением прикладной математики и информатики ФГАОУ ВПО «Национальный исследовательский университет «Высшая школа экономики»»
	Филимонов Дмитрий Алексеевич кандидат физико-математических наук, ведущий научный сотрудник ФГБУ «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» Российской академии медицинских наук
Ведущая организация:	ФГБОУ ВПО «Воронежский государственный университет»

Защита состоится 28 мая 2014 года в 16 час. 45 мин. на заседании диссертационного совета Д 501.002.16, созданного на базе ФГБОУ ВПО «Московский государственный университет имени М.В. Ломоносова», по адресу: 119991, Москва, ГСП-1, Ленинские горы, д.1, ФГБОУ ВПО МГУ имени М.В. Ломоносова, механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в Фундаментальной библиотеке ФГБОУ ВПО «Московский государственный университет имени М.В. Ломоносова» (Москва, Ломоносовский проспект, д.27, сектор А, 8 этаж).

Автореферат разослан «____»_____2014 г.

Ученый секретарь диссертационного совета
Д 501.002.16, созданного на базе ФГБОУ ВПО МГУ,
доктор физико-математических наук,
профессор Корнев Андрей Алексеевич

Общая характеристика работы

Актуальность

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].

Одной из ключевых задач хемоинформатики является задача поиска количественных соотношений «структура – свойство». 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[2].

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

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

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

рованный просмотр базы данных химических соединений и отбор тех из них, для которых прогнозируется наличие желаемых свойств, носит название виртуальный скрининг[3]. Использование виртуального скрининга позволяет существенно сократить объем длительных и дорогостоящих экспериментальных исследований в области химии, медицины и биологии.

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.

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

В настоящей диссертационной работе рассматривается подход к описанию структур М-графов на базе фрагментных дескрипторов (различных уровней) особых точек М-графов[4]. Особыми точками выступали 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 $x = (x_1, \ldots x_M) \in \mathbb{R}^M$ 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 $x = (x_1, \ldots x_M) \in \mathbb{R}^M$ 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 прогноза на обучающей выборке. В настоящей работе для определения качества модели используется процент верно классифицированных М-графов в процессе выполнения процедуры скользящего контроля.

[4] 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.

Процедура скользящего контроля (leave-one-out cross-validation[5]) 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, построенная с помощью фиксированного алгоритма обучения по обучающей выборке, обозначается $RM(LS)$. 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, обозначается как $\varphi(RM, CS) = \varphi(RM(LS), CS)$. При этом $\varphi(RM) := \varphi(RM, LS)$.

Модель, построенная с помощью фиксированного алгоритма обучения по обучающей выборке, обозначается $RM(LS)$.

В случае, когда заранее задано дескрипторное описание молекулярных графов, задача «структура – свойство» сводится к задаче классификации[6]. В свою очередь задача «структура – активность» сводится к задаче регрессии[7]. Для обеих задач могут быть применены математические методы теории распознавания образов и методы машинного обучения[8]. Одним из оригинальных подходов к решению задачи можно считать предложенный В.К.Финном ДСМ-метод автоматического порождения гипотез[9]. Наряду с методами, использую-

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

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

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

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

[9] 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.

щими дескрипторное описание, существуют также многочисленные беспризнаковые подходы[10].

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 конкретного

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

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 фраг-

ментных дескрипторов особых точек, связанных с особенностями прикладных задач прогнозирования свойств химических соединений.

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

Практическая значимость

Разработанные алгоритмы решения задачи «структура – свойство» могут быть использованы для решения прикладных задач предсказания физико-химических свойств или биологической активности веществ по их структуре. Это позволяет отказаться от дорогостоящего и длительного экспериментального скрининга 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.

Апробация работы

Материалы работы докладывались и обсуждались на 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 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 ру-

ководством 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 информатики 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.).

Публикации по теме диссертации

По материалам диссертации опубликовано 14 научных работ [1–14]. Из них: одна монография [5], четыре работы [1, 2, 3, 4] представлены в журналах из перечня ведущих научных журналов и изданий, рекомендованных ВАК РФ.

Структура и объем диссертации

Диссертационная работа состоит из введения, 3 глав, заключения и списка литературы. Общий объем диссертации – 137 страниц. Список литературы содержит 67 названий.

Краткое содержание работы

Во **введении** дано общее описание полученных результатов, сформулирована научная новизна работы, показана практическая значимость диссертации.

В первой главе приведена общая постановка задачи «структура – свойство». Кратко рассматриваются этапы решения задачи. Приводятся ключевые особенности задачи «структура – свойство». Также глава содержит основные определения и постановки задач, используемые для формулирования теоретической части работы.

Раздел 1.1 содержит описание основных этапов решения задачи – этапа описания структуры молекулярного графа и этапа поиска функциональной зависимости. **Раздел 1.2** посвящен ключевым особенностям задачи «структура – свойство», выделяющим её среди общих задач классификации. **Раздел 1.3** содержит основные определения и постановки задач, используемые для формулирования теоретической части работы. Постановки задач учитывают особенности задачи «структура – свойство», описанные в **разделе 1.2** диссертации. Предлагаемые решения поставленных задач, а также их теоретические оценки, содержатся в **главе 2**.

В главе 2 приведены 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. Описаны практические методы согласованного прогнозирования свойств неизученных молекулярных графов по множествам моделей. Приведен метод понижения сложности дескрипторного описания для неоднородных выборок. Также обсуждаются перспективы предложенного подхода.

Раздел 2.1 описывает общую методологию прогнозирования свойств неизученных М-графов по множествам распознающих моделей с заданными ограничениями допустимости $RM_1, RM_2, \ldots RM_K$. Модель RM_i называется допустимой для М-графа $\hat{x}$, если $g_i(\hat{x}) = 0$. Таким образом, М-графу $\hat{x}$ можно поставить в соответствие набор допустимых моделей $SRM_x = \{RM_i \in SRM \mid g_i(\hat{x}) = 0\}$. В качестве методов согласованного прогнози-

рования приводятся методы голосования и взвешенного голосования, метод положительных оценок, голосование сильнейших, метод победителя, а также вероятностная оценка.

В **разделе 2.2** описан метод эволюционного отбора дескрипторов, используемый для адаптации дескрипторного описания. **Раздел 2.3** посвящен методам построения моделей «структура – свойство» на базе кластерной структуры обучающей выборки. В **подразделе 2.3.1** приводится пример построения ограничений допустимости на базе кластерной структуры обучающей выборки. Подробно описана процедура построения ограничений допустимости с использованием алгоритмов кластеризации *k*-средних с ядрами и минимального покрывающего дерева. Алгоритм минимального покрывающего дерева предлагается использовать для определения числа кластеров. В то время как *k*-средних с ядрами – для определения точной формы кластеров и задания их центров и радиусов.

В **подразделе 2.3.2** описан нечеткий классификатор на базе нечеткой кластерной структуры обучающей выборки. Нечёткие кластеры описываются *матрицей нечёткого разбиения* $S=[\mu_{ij}], \mu_{ij}\in[0,1],\ i\in\{1,\dots N\},\ j\in\{1,\dots k\}$, в которой *i*-ая строчка содержит степени принадлежности М-графа $(x_{i1},\dots x_{iM})$ к кластерам $S_1,\dots S_k$.Предлагается следующий способ параметризации нечеткого разбиения. Параметрами выступает пара чисел $\lambda_1,\lambda_2\in\mathbb{R},\ \lambda_1\le 1,\lambda_2\ge 1$. 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:

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

Для нового М-графа $\tilde{x} = (\tilde{x}_1, \ldots \tilde{x}_M)$ рассматривается k предсказаний целевого 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. В **подразделе 2.3.3** обсуждается также оптимизация нечеткой кластерной структуры по её параметрам.

В **разделе 2.4** представлена двухфазная схема решения задачи «структура – свойство». Пусть обучающая выборка LS состоит из N М-графов x_i, $i = 1, \ldots, N$, каждому из которых поставлено в соответствие одно из значений: «1» или «–1». Значение «1» 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. 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 с помощью модели первого уровня М-графов. Через W_1 обозначим множество ошибочно классифицированных с помощью модели первого уровня М-графов: $W_1 = \{x_i \in LS \mid RM_1(x_i) \neq y_i\}$. Таким

образом, показатель качества 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 с помощью модели первого уровня получен верный прогноз (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, но 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 целевого свойства М-графа x_i.

Для $x_i \in LS$ определим:

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

Таким образом, результирующая модель осуществляет отказ от прогноза тогда, когда модель второго уровня предсказывает, что модель первого уровня ошибается, и осуществляет прогноз целевого свойства с помощью модели первого уровня в противном случае.

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

В **подразделе 2.4.2** доказан основной результат.

Теорема. Верна следующая оценка качества результирующей модели.

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \mathrm{R}eject}{2(N - \mathrm{R}eject)}.$$

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

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

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

Таким образом, доказано, что применение двухфазной схемы решения задача «структура – свойство» позволяет улучшить качество прогноза на обучающей выборке за счет осуществления отказа от прогноза. Это подтверждено также в серии практических испытаний метода, описанных в **главе 3**. Кроме того, преимуществами подхода являются его универсальность – описанная схема не зависит от конкретных алгоритмов классификации и предоставляет исследователю широкую свободу в выборе методов построения распознающих моделей, а также возможность независимой адаптации дескрипторного описания обучающей выборки под каждую из задач классификации, что в свою очередь позволяет, накладывая определенные ограничения на отбор дескрипторов

для модели второго уровня, добиться низкой вычислительной сложности для правил отказа.

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.

В подразделе 2.4.3 дается интерпретация предложенной двухфазной схемы на примере классификации методом опорных векторов. **Подраздел 2.4.4** содержит изложение модифицированной двухфазной схемы, реализующей классификацию без отказов. В качестве приложений двухфазной схемы в **подразделе 2.4.5** рассмотрены метод последовательного вычерпывания ошибки и многоуровневая классификация.

Оценки вычислительной сложности предложенных в работе алгоритмов анализа обучающей выборки, а также алгоритмов прогнозирования свойств неизученных М-графов приведены в **разделе 2.6**. В частности, получены следующие результаты. 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. Сложность построения ограничений допустимости при использовании двухфазной схемы решения задачи «структура – свойство» не превосходит $O(CRM(N,M)\cdot N\cdot M)$. Вычислительная сложность прогнозирования целевого свойства молекулярного графа с использованием двухфазной схемы равна $O(CD\cdot\max(M_1,M_2))$, где CD – средняя сложность вычисления одного дескриптора, M_1 – количество дескрипторов, эволюционно отобранных для решения задачи классификации первого уровня, а M_2 – количество дескрипторов, отобранных для решения задачи второго уровня. Отказ от прогноза для недопустимых М-графов при этом осуществляется за $O(CD\cdot M_2)$.

Раздел 2.7 содержит описание метода построения модели «структура – свойства», позволяющего понизить вычислительную сложность дескрипторного описания неоднородных выборок. Обучающая выборка LS называется *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.

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 достичь на неоднородной обучающей выборке, обозначим это значение φ_c. Первым этапом обработки выборки будет выделение общих кластеров с помощью наиболее простого дескрипторного описания. Наиболее простые с вычислительной точки зрения дескрипторы используются 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 (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 показатель 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 $x_i \in LS$: $M(x_i) = \sum_{j=1}^{k} \mu_j(x_i) \cdot RM_j(x_i)$. Качество обобщающей модели обозначим через θ. Доказан промежуточный результат.

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

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

числялись 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. 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 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).

Глава 3 содержит результаты практического тестирования подхода. В **разделе 3.1** описана программная реализация разработанных алгоритмов. Далее приведены результаты ее использования на реальных обучающих выборках соединений. Приведено сравнение результатов использования разработанного подхода с классическими аналогами. Экспериментально подтверждены полученные теоретические оценки. В **разделах 3.2 – 3.4** предлагаются 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.

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

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.

Список опубликованных работ по теме диссертации

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).