

**УТВЕРЖДАЮ**

Проректор по научной работе и информатизации ФБГОУ ВПО «Воронежский государственный университет»,

д.б.н. профессор

Попов В.Н.



«    » 2014 г.

**ОТЗЫВ**

ведущей организации на диссертационную работу Прохорова Евгения Игоревича «Адаптивная двухфазная схема решения задачи «структура – свойство», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 – Теоретические основы информатики

**Актуальность темы диссертации.** Диссертация Е.И. Прохорова посвящена вопросам поиска количественных отношений между структурами химических соединений и их физико-химическими свойствами и биологической активностью.

Рассматриваемая в работе задача «структура – свойство» является актуальной прикладной математической задачей теории распознавания и классификации, находящей применение в области химии. Методы решения данной задачи имеют существенное значение для компьютерного моделирования активности соединений и для поиска потенциально активных веществ в больших базах химических соединений. Обозначенное направление представляется одним из наиболее важных в области хемоинформатики, относительно новой научной дисциплины, возникшей в последнее время в связи с широким применением методов информатики для решения химических проблем.

Исследование автора посвящено устранению ряда недостатков, которыми обладают классические решения задачи «структура – свойство», таких как ограниченность обучающей выборки и высокая вычислительная сложность дескрипторного описания структур соединений. Актуальность исследований такого характера обусловлена тем обстоятельством, что для применения моделей «структура – свойство» на практике с одной стороны чрезвычайно важно получить достоверный прогноз активности неизученных соединений, чему препятствует ограниченный объем обучающей выборки, а с

другой – организовать вычисления максимально эффективно, так как объемы обрабатываемых баз достигают сотен тысяч и миллионов соединений.

**Общая оценка работы.** Диссертация состоит из введения, трех глав, заключения и списка литературы.

Во введении дан краткий обзор тематики исследования и обоснована ее актуальность, перечислены цели и задачи, которые стояли в процессе работы над диссертацией. Введение содержит формулировки новых результатов, полученных автором, показана научная и практическая значимость данных результатов.

В первой главе обсуждается общая постановка задачи «структурасвойство», этапы ее решения и различные особенности задачи, выделяющие ее в ряду общих задач классификации. Особенности, которые отмечает автор, существенно зависят от выбранного подхода к решению задачи, подразумевающего использование многоуровневого признакового описания М-графов на базе фрагментных дескрипторов особых точек. На основе данных особенностей приводятся постановки задач, решение которых и рассмотрено во второй главе.

Вторая глава диссертации посвящена разработке метода построения распознающих моделей, реализующих собственные ограничения допустимости. При этом автор преследует цель построить эффективные правила отказа в смысле определения, данного в первой главе. Часть методов, описанных во второй главе, является вспомогательной и на защиту не выносится. В качестве основного результата данная глава содержит изложение двухфазной схемы решения задачи классификации. Для предложенной схемы доказана оценка качества результирующей модели при использовании ограничений допустимости. Также во второй главе описан метод понижения вычислительной сложности при обработке неоднородных выборок.

Третья глава диссертации содержит результаты практического тестирования предложенных методов и алгоритмов. Глава содержит описание программного комплекса, разработанного автором и его коллегами, а также обзор результатов, полученных в ходе обработки выборок химических соединений, предоставленных профильными институтами РАН и РАМН.

В работе автор использует различные методы теоретической информатики, теории машинного обучения и теории сложности.

Основные научные положения диссертации в достаточной степени апробированы и отражены в научной печати, что подтверждается 14 публикациями, включающими 4 статьи в журналах Перечня ВАК РФ.

**Научная новизна.** В работе предложен новый подход к решению задачи «структура – свойство» на базе описания структур молекулярных графов фрагментными дескрипторами особых точек. В рамках данного подхода автором предложено использование ограничений допустимости для моделей «структура – свойство», а также метод построения таких ограничений. Представленные в диссертации оценки качества прогнозирования для моделей «структура – свойство» позволяют говорить об эффективности данных ограничений допустимости в смысле, formalизованном автором.

Также в рамках предложенного подхода предложен метод, позволяющий осуществлять адаптацию описания структур химических соединений, с целью оптимизировать вычислительные ресурсы, требующиеся для расчета дескрипторов. Все вышеуказанные методы являются оригинальными, теоретические оценки разработанных методов получены автором лично.

Предложенный автором подход позволяет избавиться от основных недостатков существующего решения на базе фрагментных дескрипторов особых точек, связанных с особенностями прикладных задач прогнозирования свойств химических соединений.

**Значимость полученных автором результатов** заключается в том, что использование предложенных ограничений допустимости не только гарантирует в некотором смысле более достоверный прогноз свойств химических соединений, но и ускоряет сам процесс просмотра больших баз соединений за счет осуществления «быстрых» отказов от прогноза. Кроме того, предложенный в работе подход применим для решения общих задач классификации, а не только и не столько для прогнозирования свойств химических соединений. Практическая значимость работы заключается в реализации разработанных методов и алгоритмов на ЭВМ в виде программного комплекса, обеспечивающего построение моделей «структура – свойство» и использовании этих моделей для прогнозирования свойств неизученных ранее химических соединений.

**Достоверность и обоснованность** представленных результатов обеспечиваются математическими формулировками и доказательствами, а также тестами и экспериментальными исследованиями.

**Результаты диссертации могут быть использованы** в научных центрах, занимающихся исследованиями в области органической химии и хемоинформатики. Кроме того, результаты могут использоваться для решения широкого круга задач классификации, возникающих при обработке изображений, цифровой обработке сигналов и в других актуальных областях прикладной математики.

### **Замечания и пожелания по диссертации.**

1. В работе неоднократно затрагиваются аспекты вычислительной сложности рассматриваемых алгоритмов. Утверждается, что автору удалось снизить вычислительную сложность решения некоторых задач. Однако нет строгих оценок, показывающих, насколько снижается вычислительная сложность прогнозирования. Помимо этого, в тексте диссертации общие рассуждения о вычислительной сложности алгоритмов порой носят некорректный характер. Так, например, на с. 79 смешиваются понятия комбинаторной, экспоненциальной и NP-сложности.

2. Принятый в главе 2 стиль изложения затрудняет понимание границы между личными результатами автора диссертации и результатами его предшественников.

3. Несмотря на то, что текст работы содержит иллюстрации к предложенной двухфазной схеме решения на примере метода опорных векторов, обоснование двухфазной процедуры классификации могло бы только выиграть, если бы автор на каком-либо простом классическом примере, например, на примере различения статистических распределений, показал, к чему приводит реализация предлагаемого подхода, как в асимптотике ведут себя частоты ошибок первого и второго рода, а также частота отказа.

4. Трудно оценить обоснованность выбранных методов машинного обучения для прогнозирования свойств M-графов. На с. 45 перечислены и описаны реализуемые методы: линейная регрессия и машина опорных методов. В главе 3 представлены три метода: линейная регрессия, машина опорных методов, нечеткий классификатор. Однако в общем случае интерес представляют и другие методы, стандартные реализации которых доступны в среде Matlab.

5. Результаты эксперимента описаны довольно некорректно. Например, качество прогнозирования в таблице 2 (с. 107) оценивается величинами, вычисляемыми с точностью до 6 знака после запятой, в то время как количество экспериментов (соединений обучающей выборки) равно 76. Таким образом, количество экспериментов не соответствует точности предъявленных оценок.

Тем не менее, вышеперечисленные недостатки существенно не влияют на итоговую ценность достигнутых результатов и общую оценку научно-квалификационного уровня соискателя.

### **Заключение.**

Диссертация Прохорова Е.И. «Адаптивная двухфазная схема решения задачи «структура – свойство»» представляет собой завершенную, доведенную

