

## ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Прохорова Е.И. «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 выбранной тематики.** Диссертационная работа Е.И. Прохорова посвящена совершенствованию методов классификации в прикладной задаче поиска количественных отношений «структура – свойство». Актуальность тематики с одной стороны состоит в том, что изучение свойств химических соединений в рамках моделей «структура – свойство» и «структура – активность» имеет широкое применение для оценки эколого-гигиенической безопасности химических веществ, прогнозирования токсичности и канцерогенности соединений различных химических классов, а с другой определяется ростом интереса к проблематике приобретения знаний из данных, заданных мечеными графами, в области машинного обучения и разработки данных. Об актуальности темы диссертации дополнительно свидетельствуют большое количество публикаций, семинаров и конференций на эту тему.

Работа рассматривает решение задачи «структура – свойство» с позиций подхода, использующего в качестве представлений молекулярных графов фрагментные дескрипторы. Данные дескрипторы являются широко известными и неплохо изученными. Автором предложен метод, позволяющий дополнительно развить подход, обогатив его такими идеями общей теории машинного обучения, как классификация с отказами и области применимости.

**Общая оценка содержания работы.** Представленная Е.И. Прохоровым диссертационная работа содержит 137 страниц текста, список использованной литературы из 67 наименований, в том числе 14 публикаций соискателя. Диссертация состоит из введения, трех глав основного текста и заключения.

К работе приложен автореферат на 22 страницах, достаточно полно и точно отражающий основные положения диссертации. Диссертация представляет собой законченную научно-квалификационную работу, выполненную на актуальную тему. Идеи и теоретические изыскания автора воплощены в формальных определениях, теореме, следствиях и утверждениях. Предложенные методы и алгоритмы доведены до программной реализации.

Во введении автор формулирует проблематику исследования, показывает его актуальность, излагает цели и задачи работы, дает обзор полученных результатов и формулирует научную новизну работы.

Первая глава диссертации посвящена обсуждению исходной задачи «структура – свойство», а также подхода к ее решению на базе описания фрагментными дескрипторами. Автор приводит описание практических приложений моделей «структура – свойство» и связанных с ними трудностей. На основании данного обзора формулируется частная задача построения эффективных ограничений допустимости для моделирования «структура – свойство», а также формулируются общие требования к разрабатываемым методам и алгоритмам.

Вторая глава целиком посвящена методам решения поставленных задач. Излагается подход к прогнозированию свойств химических соединений на базе множеств распознающих моделей, методы адаптации дескрипторного описания молекулярных графов, методы классификации с использованием нечеткой кластеризации и, наконец, двухфазная схема решения задачи «структура – свойство». Для предложенного подхода в рамках формализованных понятий сформулирована оценка качества моделирования (теорема на стр.67 и следствия 1,2 на стр. 68). Также глава содержит оценки вычислительной сложности рассмотренных алгоритмов и описание метода, в некотором смысле оптимизирующего вычислительные ресурсы в процессе обработки неоднородных выборок химических соединений.

Третья глава посвящена практической реализации предложенных методов. Автором описан разработанный программный комплекс, а также результаты тестирования на различных выборках химических соединений.

В заключении подводятся итоги и обсуждаются некоторые направления дальнейших исследований.

**Обоснованность и достоверность** основных положений диссертационной работы обеспечиваются их представлением в виде строгих утверждений, сформулированных на математическом языке и снабженных

необходимыми доказательствами, а также наличием научных публикаций по теме работы; в том числе в реферируемых изданиях, апробацией на российских и международных конференциях, апробацией на исследовательских семинарах ведущих научных учреждений. В своем исследовании автор применяет методы теоретической информатики, теории машинного обучения и теории сложности.

Выносимые на защиту результаты диссертации являются новыми. Они четко сформулированы, обоснованы в виде математических доказательств, алгоритмов, и получены автором самостоятельно. Для каждой публикации, выполненной в соавторстве, указана та часть работы, которая принадлежит соискателю.

**Научная и практическая значимость работы.** Значимость работы заключается в том, что предложенные автором методы обладают рядом преимуществ перед стандартным решением задачи «структура – свойство» на базе фрагментных дескрипторов особых точек, они позволяют повысить качество и достоверность прогнозирования свойств неизученных соединений. Практическая значимость состоит в том, что исследования автора доведены до программной реализации, построены модели «структура – свойство», которые могут быть использованы для прогнозирования целевых свойств в различных профильных организациях химического и биологического профиля.

Вместе с тем, можно отметить следующие недостатки работы и сформулировать общие **замечания**.

1. Высказывание автора (стр.18-19) "*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])*" не соответствует сути ДСМ-метода. Во-первых, ДСМ-метод сформулирован на специальном логическом языке слабой логики предикатов второго порядка, и его формулировка на языке АФП не является обязательной (хотя и весьма удобна для понимания и вычислений). Во-вторых, в рамках ДСМ-метода использование дескрипторных языков возможно и применялось в десятках экспериментах, в основном в сочетании с дескрипторным языком ФКСП. Правда, ДСМ-метод можно применять на графовом представлении "напрямую", без создания дескрипторов, а используя операцию пересечения на множествах графов.

2. Недостаточно внимания уделено сравнению с другими методами, допускающими отказы от классификации, как например ДСМ-метод (при противоречивой классификации из-за наличия положительных и отрицательных гипотез).
3. В качестве кросс-валидации (см. стр. 40) используется только выброс одного элемента выборки. Стандартными значениями параметров кросс-валидации являются 10% с повтором в 10 раз.
4. "Метод эволюционного отбора признаков" (стр. 52-54) весьма близок методам построения деревьев решений, без уточнения критерия отбора следующих признаков. Отсутствует указание на данный факт.
5. В тоже время, модель обучения второго уровня (стр. 63-78) – это разновидность бустинга, о чем практически не сказано и не сделано сравнения с другими методами бустинга.
6. Диссертация обладает также некоторыми огрехами изложения:
  - стр. 28 "*набор вершин*" – множество вершин;
  - стр. 38, внизу: "*не может быть осуществлен*" – может быть осуществлен;
  - стр. 71–73: вместо отказа можно было бы предложить классифицировать отброшенные точки отдельно.

Несмотря на отмеченные недостатки и сделанные замечания, диссертационная работа Прохорова Е.И. производит положительное впечатление. В целом работа написана хорошим научным языком, содержит лишь небольшое количество опечаток и языковых ошибок. В постановках задач и в методах решения наблюдается баланс теории и практики. Математические формулировки и доказательства сочетаются с программными реализациями и экспериментальным тестированием. Ценность и значимость достигнутых результатов не вызывают сомнения.

Диссертация представляет собой законченную, доведенную до практической реализации, научно-исследовательскую работу, выполненную на актуальную тему. Автореферат диссертации достаточно полно и точно отражает содержание диссертации. Результаты диссертации являются новыми, в достаточной степени обоснованными и апробированными.

Таким образом, диссертационная работа «Адаптивная двухфазная схема решения задачи «структура – свойство»» удовлетворяет требованиям «Положения о присуждении ученых степеней» Правительства РФ, а ее автор, Прохоров Евгений Игоревич, заслуживает присуждения ему ученой степени

кандидата физико-математических наук по специальности 05.13.17 – теоретические основы информатики.

Отзыв подготовил доктор физико-математических наук, профессор, заведующий отделением прикладной математики и информатики ФГАОУ ВПО «Национальный исследовательский университет «Высшая школа экономики»», Кузнецов Сергей Олегович.

11 «мая» 2014 года

Зав. отделением прикладной математики и информатики НИУ ВШЭ, профессор НИУ ВШЭ, д.ф.-м.н., Кузнецов Сергей Олегович

Адрес: Москва, Большой Трёхсвятительский переулок, 3, ауд. 420,

телефон: +7 (495) 772-95-90 \* 22670,

e-mail: [skuznetsov@hse.ru](mailto:skuznetsov@hse.ru).