

«УТВЕРЖДАЮ»

И.о. директора
Федерального государственного
бюджетного учреждения науки
Института Проблем Управления
им. В.А. Трапезникова
Российской Академии Наук



Проф. РАН

М. В. Губко

24 сентября 2018 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

о диссертационной работе Юрия Савельевича Кашницкого «Методы замкнутых описаний в задаче классификации данных со сложной структурой», представленной к защите на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики»

Диссертация Ю.С. Кашницкого посвящена интерпретируемой классификации данных со сложной структурой, такой как последовательности и графы, на основе аппарата Анализа Формальных Понятий и, в частности, узорных структур и их проекций.

Тема диссертации Ю.С. Кашницкого, несомненно, актуальна поскольку она посвящена описанию принципиально новых методов решения описания в классификационных задачах, получивших чрезвычайно широкое практическое распространение при том, что специалистам в этой области хорошо известны издержки и, порою, весьма, ограниченные возможности уже известных методов.

Диссертация содержит 111 страниц и состоит из введения, 4 глав, заключения и списка литературы из 107 наименований.

Во введении обосновывается актуальность работы, формулируются метод и предмет исследования, характеризуется ее научная новизна, перечисляются теоретические и практические достоинства работы. Также перечислены 12 конференций и семинаров, на которых автор с 2014 по 2018 годы докладывал результаты, излагаемые в диссертации. К этому следует добавить, что ведущей организации – ИПУ РАН им. В.А. Трапезникова – многие положения рецензируемого диссертационного исследования знакомы по

выступлению его автора с докладом на тему «Методы замкнутых описаний в задачах классификации данных со сложной структурой» на научном семинаре в сентябре 2018 года.

В первой главе дается обзор базовых понятий теории решеток и Анализа Формальных Понятий (математического аппарата, на основе которого происходит дальнейшее изложение), методов классификации в машинном обучении, основанных на ассоциативных правилах, а также рассматриваются критерии отбора классифицирующих правил. При этом уделяется внимание не только качеству классификации (обычно понимаемому как средняя доля верных ответов алгоритма на кросс-валидации или отложенной выборке), но и интерпретируемости правил, участвующих в классификации тестовых примеров. Также обсуждается, как Анализ Формальных Понятий до этого применялся в машинном обучении и, конкретно, в задаче классификации. Деревья решений интерпретируются в терминах АФП и показывается, как с помощью алгоритма построения решетки формальных понятий предложить отбор правил в задаче классификации, при котором гарантируется, что каждый объект тестовой выборки классифицируется правилом не хуже (в терминах выбранного критерия, такого как неопределенность Джини или прирост информации), чем при классификации на основе дерева решений.

Во второй главе рассматривается аппарат узорных структур (Pattern Structures) который позволяет расширить методы Анализа Формальных Понятий на случай, когда объекты задаются не бинарными признаками, а сложными описаниями, такими как интервалы числовых значений, множества последовательностей, строк или графов. Рассматриваются подходы к классификации данных со сложной структурой на основе ядерных функций и метода опорных векторов, а также на основе узорных структур и их проекций. Также автор вводит проекции интервальных узорных структур, позволяющие интерпретировать деревья решений в терминах АФП также и в случае количественных признаков. С помощью этих дискретизирующих проекций интервальных узорных структур показано, как в случае данных с количественными признаками находить классифицирующие правила не хуже, чем правила, построенные деревом решений, по критерию информативности типа прироста информации или неопределенности Джини.

В первой и второй главах показано, что Анализ Формальных Понятий предлагает удобный формализм для того чтобы, с одной стороны, выразить на этом языке многие алгоритмы, основанные на классифицирующих ассоциативных правилах, а с другой, чтобы обобщить эти алгоритмы на случай данных со сложной структурой. В третьей главе предлагается алгоритм классификации произвольных данных со сложной структурой, для которых можно ввести полурешёточную операцию сходства. Это позволяет решать задачу

классификации для данных разного типа, при этом придерживаясь единого подхода. Отдельно и с подробными примерами рассматриваются частные случаи, когда данные представлены бинарными, количественными и интервальными признаками, а также помеченными графами.

В четвертой главе описывается разработанный программный комплекс, реализующий алгоритмы, описанные в третьей главе, затем приводятся результаты вычислительных экспериментов. В силу того, что предлагаемые алгоритмы приспособлены для работы с данными различной природы, демонстрируются результаты решения задачи классификации для 13 наборов данных с количественными и бинарными и категориальными признаками (репозиторий UCI), для 14 наборов данных с количественными признаками (также UCI), а также приводятся результаты экспериментов в задачах классификации различных видов последовательностей (транскрипции австралийских видеозаписей языка жестов глухонемых людей, видео взаимодействия руки человека с предметом, предобработанные признаки временного ряда мускульной активности и позиции ног профессиональных лыжников и проч.) и прогнозирования свойств химических веществ (токсичности для мышей и крыс, данные Predictive Toxicology Challenge) и прочие). Помимо этого, приводятся результаты применения алгоритма в задаче прогнозирования оттока клиентов телеком-оператора. Анализируется как качество решения задачи классификации (средняя доля верных ответов на кросс-валидации), так и прочие свойства алгоритма, в частности, длина посылок правил, которыми классифицировались тестовые примеры (в работе для этого вводится термин «локальная интерпретируемость», подчеркивающий то наблюдение, что более простые и «короткие» правила легче интерпретируются человеком).

Диссертация Ю.С. Кашницкого обладает некоторыми особенностями, которые заметно отличают ее от других работ, посвященных АФП и классификации данных со сложной структурой. А именно:

- 1). В большинстве работ анализируется либо самый популярный формат данных в реальных задачах — с категориальными и количественными признаками, либо специфичные структуры данных, например, которые можно свести к помеченным графам. Предлагаемый в работе подход универсален, показано, что он применим как к данным с категориальными и количественными признаками, так и к последовательностям и графам. Более того, предлагаемые алгоритмы могут быть применены (с некоторыми ограничениями) для классификации произвольных сложно структурированных данных, например, в ситуации, когда значения количественных признаков известны не точно, а с некоторой погрешностью

2). Автор исследует 4 свойства алгоритмов классификации: качество решения задачи (как правило, доля верных ответов, в большинстве работ исследуют только это свойство), быстродействие, интерпретируемость и возможность работы с данными со сложной структурой. Отметим интересные результаты автора в исследовании интерпретируемости классифицирующих правил, когда без потери точности задачу можно решить с помощью более простых правил.

Практическая ценность работы подтверждается ее внедрением в проекте компании «Гемотест» при исследовании прогнозирования концентрации химических веществ в венозной крови.

К недостаткам работы можно отнести следующие:

1). Рассмотрено большое количество подходов к классификации данных, представленных графами, с помощью ядерных методов, однако сравнение проводится только с одним из них.

2). Введённый термин «локальная интерпретируемость» понятен, однако знак в определяющей формуле выбран неверно: если указанный индекс имеет большое значение, это означает, что правила «длинные» и плохо интерпретируемые, однако при этом формально «локальная интерпретируемость» оказывается выше.

3). В серии экспериментов с графовой классификацией нет четкого указания на преимущества предлагаемого подхода перед СВА.

4). Не во всех экспериментах указано время работы алгоритмов.

5). В тексте диссертации и автореферата встречаются отдельные стилистические несуразиды (помечены в тексте работы).

Впрочем, отмеченные недостатки не носят принципиального характера и не оказывают влияния на положительную характеристику диссертационной работы в целом.

Результаты рецензируемого научного исследования соответствуют паспорту специальности 05.13.17 «Теоретические основы информатики» в пунктах 5 – «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечение, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» и 7 – «Разработка методов распознавания образов, фильтрации, распознавания и синтеза изображений, решающих правил. Моделирование формирования эмпирического знания».

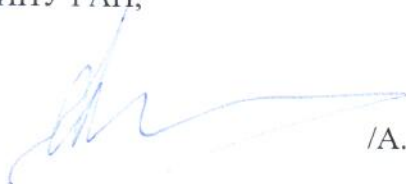
Диссертация содержит важные результаты, обладающие научной новизной,

теоретической значимостью и практической ценностью.

Таким образом, диссертационная работа Ю.С. Кашницкого «Методы замкнутых описаний в задаче классификации данных со сложной структурой» отвечает требованиям пунктов 9-14 Положения о защите ученых степеней, утвержденного Постановлением Правительства Российской Федерации от 24.09.2013 г. №842 и требованиям ВАК, Минобрнауки РФ, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики», а ее автор Ю.С. Кашницкий заслуживает присвоения ему ученой степени кандидата технических наук по указанной специальности.

24 сентября 2018 г., г. Москва.

главный научный сотрудник лаб. № 25 ИПУ РАН,
доктор технических наук



/А. С. Мандель/

Диссертация Ю.С. Кашницкого заслушана на объединенном семинаре лаб. №№ 7, 11, 25, 33, 38, 57 и 77 ИПУ РАН им. В.А. Трапезникова 24 сентября 2018 года и получила положительную оценку.

Председатель семинара
Зав. лаб. № 77 ИПУ РАН,
кандидат технических наук



/А. В. Макаренко/

Сведения о ведущей организации: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук (ИПУ РАН).

Адрес: 117997, Россия, Москва, ул. Профсоюзная, д. 65, телефон: +7 495 334-89-10

E-mail: dan@ipu.ru



А.С. Мандель
А.В. Макаренко