

ОТЗЫВ

**официального оппонента на диссертационную работу Виноградова
Дмитрия Вячеславовича «Вероятностно-комбинаторный формальный
метод обучения, основанный на теории решеток», представленную на
соискание ученой степени доктора физико-математических наук по
специальности 05.13.17 «Теоретические основы информатики»**

Анализ экспериментальных данных является составной частью исследовательского арсенала практически всех естественно-научных дисциплин. Среди соответствующих методов выделяются методы интеллектуального анализа данных (ИАД), особенностью которых является использование логических и семантических отношений на элементах и структурах данных. В таких случаях требуется привлечение комбинаторных средств, опирающихся на технику эмпирической индукции. Применение логико-комбинаторных методов (например, индуктивного логического программирования) не требует репрезентативных обучающих выборок. Однако такие методы обладают высокой вычислительной сложностью и не могут справиться с обучающими выборками большого объема (так называемый феномен Big Data – «большие данные»).

Актуальность темы диссертационного исследования определяется необходимостью расширения комбинаторных методов ИАД на большие выборки. В обсуждаемой диссертации эта задача решается путем применения вероятностного подхода на основе предлагаемого в работе вероятностно-комбинаторного формального метода (ВКФ-метода) к машинному обучению.

Научная новизна предложенного подхода и соответствующего формального аппарата состоит в следующем.

- Использование методов теории решеток для представления объектов, описываемых признаками со значениями в нижних полурешетках, битовыми строками.
- Анализ феномена переобучения – появления нежелательных комбинаций признаков как сходства обучающих примеров, имеющих различные механизмы появления целевого свойства.
- Создание и исследование нескольких вероятностных алгоритмов для нахождения кандидатов в гипотезы о причинах появления целевого свойства.
- Реализация синтеза вероятностных познавательных процедур (индукции, аналогии и абдукции).

Достоверность полученных результатов. Все математические результаты диссертационной работы строго доказаны, выводы и заключения обоснованы. Достоверность полученных результатов подтверждается также тестированием созданной автором программной системы, реализующей предложенные в диссертационной работе алгоритмы, на массивах данных из репозитория данных Университета Калифорнии в г. Ирвайн.

Практическая значимость результатов работы состоит в том, что они создают основу для создания интеллектуальных систем для ИАД с возможностью анализа «больших данных». Следует отметить, что большинство теорем имеют практический смысл: устанавливают свойства предложенных алгоритмов (остановку с вероятностью единица, ускорение в «ленивой» схеме вычислений) или определяют численные значения необходимых параметров (число необходимых гипотез для правильного предсказания, число предварительно учитываемых траекторий для верхней границы склеивания). Работоспособность предложенного подхода подтверждена применением его программной реализации к реальным массивам из репозитория данных для тестирования алгоритмов машинного обучения. Эти эксперименты показали преимущества перед некоторыми другими комбинаторными методами машинного обучения (обучение покрытию с помощью целочисленного программирования) и дают обоснованную надежду на возможность применения развитого автором подхода к массивам данных большого объема.

Основные результаты диссертационного исследования опубликованы в 22 научных публикациях (20 из которых написаны без соавторов), в том числе, 17 работ – в изданиях, рекомендованных ВАК для публикации основных научных результатов диссертаций на соискание ученых степеней кандидата и доктора наук. Полученные результаты докладывались на многочисленных научных семинарах, отечественных и международных конференциях по искусственному интеллекту и интеллектуальному анализу данных.

Содержание диссертации. Диссертация (общим объемом 131 с.) включает Введение, 4 главы, Заключение и список используемых источников из 80 наименований.

Введение описывает постановку задачи и современное состояние исследований в области ИАД. Главы 1-4 диссертации посвящены предлагаемому автором оригинальному подходу к машинному обучению, опираясь на математический аппарат теории решеток:

В Главе 1 вводятся необходимые определения и факты из области анализа формальных понятий – современного раздела теории решеток. Затем определяются ключевые операции «замыкай-по-одному», посредством которых можно осуществить случайное блуждание на решетке кандидатов в гипотезы. Для этих операций доказываются их основные свойства и теорема об эффективности использования «ленивой» схемы вычислений на длинных случайных сериях их применений. В результате предложен алгоритм (Algorithm 1) кодирования битовыми строками элементов нижней полурешетки (с операцией побитового умножения), применяя который можно анализировать объекты, описываемые признаками, на множествах значений которых задаются нижние полурешетки (с добавленным тривиальным сходством, если его нет).

В Главе 2 исследуется феномен переобучения. Он возникает при наличии «фантомных» сходств, когда комбинация случайно совпавших признаков у обучающих примеров, каждый из которых обладает своим собственным механизмом возникновения целевого свойства, объявляется еще одной причиной появления целевого свойства. Сначала устанавливаются негативные результаты о том, что защитные механизмы классического ДСМ-метода (ограничение на число родителей и запрет контрпримеров) недостаточны, чтобы полностью устранить возникновение фантомных сходств. Затем выводятся явные формулы для вычисления вероятности порождения сходства с запретом какого-то числа контрпримеров.

В Главе 3 исследуются цепи Маркова, задающие случайные блуждания на решетке кандидатов в гипотезы, являющиеся вероятностной процедурой порождения представительного множества гипотез о причинах целевого свойства. Формально представлены и исследованы алгоритмы для немонотонной (Algorithm 2), монотонной (Algorithm 3) и спаривающей (Algorithm 4) цепей Маркова. Для класса спаривающих цепей Маркова из классической теории счетных дискретных цепей Маркова выводится результат об остановке соответствующего алгоритма с вероятностью единица. Для случая Булевой алгебры доказаны теоремы о средней длине траекторий и сильной концентрации длин траекторий вокруг своего среднего. В общем случае, используя технику субмультипликативности, доказывается теорема с оценкой того, насколько отличаются вероятности результатов обычной и остановленной спаривающих цепей Маркова в метрике тотальной вариации. Остановку следует осуществлять, если длина текущей траектории превосходит сумму длин предварительно выполненных траекторий.

В Главе 4 рассматриваются алгоритмы машинного обучения – предсказания по аналогии (Algorithm 5), индуктивного обобщения (Algorithm 6), абдуктивного обобщения (Algorithm 7), их программная реализация и

результаты применения к массивам данных. Доказана теорема о необходимом числе гипотез для того, чтобы с заданной вероятностью (надежностью) все важные (доопределяемые с вероятностью не ниже заданной) тестовые примеры были предсказаны правильно.

Диссертация характеризуется высоким научным уровнем. Текст написан хорошим языком, стиль изложения четкий и понятный. Материал диссертации хорошо структурирован, изложение характеризуется логической последовательностью и математической строгостью. Достоинствами диссертационного исследования является нацеленность предложенного формального аппарата (ВКФ-метода в целом и соответствующих теорем) на повышение эффективности машинного обучения («ленивая» схема вычислений, учет предварительных запусков цепи Маркова для остановки слишком длинных траекторий) и ИАД (включая «большие данные»), а также возможность сокращения труда экспериментатора, что следует из теоремы о необходимом числе гипотез, чтобы с заданной надежностью достичь требуемой точности прогноза.

Замечания по диссертации

1. В названии диссертации вместо «обучения» следовало написать «машинного обучения», так как исследование связано именно с машинным обучением.

2. Недостаточное описание массивов данных, на которых проводилось сравнение с другими методами машинного обучения и тестировалась возможность работы на выборках данных большого объема.

3. В работе отмечается хорошее распараллеливание предложенных алгоритмов, что позволяет существенно сократить время их выполнения, однако какие-либо формальные оценки не приводятся.

4. Практическая ценность работы была бы усилена, если бы полученные результаты (разработанные ВКФ-метод и ВКФ-система на его основе) были не только протестированы на двух массивах из репозитория данных UCI Machine Learning Repository для тестирования алгоритмов машинного обучения, но и применены для решения какой-либо прикладной задачи в некоторой предметной области (что отмечено автором относительно ДСМ-метода).

5. Имеются некоторые погрешности по тексту диссертации и автореферата, в частности: символ F на стр. 16 и 17 диссертации означает наибольший (максимальный) элемент, а чуть ниже (на стр.17) и далее под F понимается множество признаков; в автореферате: при формулировке теоремы 1 (стр. 13) не определены переменные n и k (далее на стр. 13 указано, что n обозначает количество сопутствующих признаков, а на стр. 14 n используется в контексте "последовательность n испытаний Бернулли"); в

определении 4 (стр. 14) следовало указать, что $s \in S$, и не определен параметр z ; при формулировке теоремы 8 (стр. 17) не определен параметр γ ; сокращение ВКФ используется на стр. 15 («ВКФ-кандидатов»), стр. 17 («ВКФ-системой»), а расшифровывается только на стр. 19 в тексте «вероятностно-комбинаторный формальный метод (ВКФ-метод) машинного обучения».

6. Следует отметить, что автор, прекрасно владея математическим аппаратом, излишне часто использует выражения типа «легко видеть (доказать, установить)», «очевидно», опуская соответствующие доказательства и пояснения, что не всегда оправдано и усложняет восприятие материала. Текст диссертации довольно компактен (131 стр.) и после основного текста вполне можно было бы добавить Приложение с дополнительными пояснениями (доказательствами).

Отмеченные замечания (касающиеся в основном изложения, а не собственно исследований) не умаляют значимости и общей положительной оценки диссертационной работы. Автореферат и публикации автора полностью отражают основное содержание работы.

Новизна, теоретическая и практическая значимость результатов выполненного исследования позволяют квалифицировать диссертационную работу «Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток» как открытие нового научного направления в теоретической информатике – «машинное обучение, основанное на теории решеток». Диссертационная работа удовлетворяет критериям, установленным «Положением о присуждении ученых степеней», утвержденным Постановлением Правительства РФ от 24 сентября 2013 г. № 842, полностью соответствует всем требованиям ВАК, предъявляемым к диссертациям на соискание ученой степени доктора физико-математических наук, а ее автор – Виноградов Дмитрий Вячеславович – заслуживает присуждения ему искомой ученой степени доктора физико-математических наук по специальности 05.13.17 «Теоретические основы информатики».

Официальный оппонент:

Зав. кафедрой прикладной математики ФГБОУ ВО НИУ «МЭИ»

д.т.н., профессор

«7» февраля 2019 года

А.П. Еремеев

Организация – место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Национальный исследовательский университет «МЭИ» (ФГБОУ ВО НИУ «МЭИ»)

Почтовый адрес: 111250, Россия, г. Москва, Красноказарменная ул., дом 14. Тел.: 8 (495) 362 79 62. Адрес электронной почты: mpi@mpei.ru. Веб-сайт: <http://www.mpei.ru/>



История достоверно 5

ЗАМЕСТИТЕЛЬ НАЧАЛЬНИКА
ПРАВЛЕНИЯ ПО РАБОТЕ С ПЕРСОНАЛОМ
Л.И. ПОЛЕВАЯ