

ОТЗЫВ

официального оппонента на диссертацию М.И.Забейайло
«Комбинаторные средства формализации эмпирической индукции»,
представленную на соискание ученой степени
доктора физико-математических наук по специальности
05.13.17 – Теоретические основы информатики

Задача восстановления зависимостей по прецедентам – одна из классических задач компьютерного анализа данных. Большинство имеющихся на сегодняшний день исследований по этой тематике посвящено обработке числовых данных. Вместе с тем актуальной и сравнительно мало изученной областью является проблематика извлечения зависимостей из данных (описаний прецедентов), представленных *нечисловыми* объектами. Традиционные подходы, основанные на метрических соображениях, как правило, не дают здесь надежных результатов (адекватное определение соответствующих метрик или же мер близости в конкретных ситуациях оказывается скорее удачей исследователя, а вовсе не объективным следствием продуктивности используемых подходов и математических формализмов). Дополнительные трудности возникают также в ситуациях, когда исследуемые нечисловые объекты, представляемые как реляционные структуры - множества признаков с теми или иными отношениями на этих признаках (например, мультимножества, графы, цепочки символов конечного алфавита и т.п.), в своих описаниях содержат также и числовые значения тех или иных существенных параметров. Не менее важными оказываются возможности оперировать малыми (статистически незначимыми) обучающими выборками примеров и контрпримеров. Наконец, принципиально значимым для целого ряда приложений является возможность восстанавливать по заданным прецедентам исчерпывающий набор причинных «влияний», формирующих соответствующие каузальные зависимости в имеющихся исходных эмпирических данных.

Диссертационная работа М.И.Забейайло посвящена именно этой проблематике – разработке комплекса математических моделей и методов интеллектуального анализа данных (ИАД) для эффективного решения задач обучения на прецедентах, которые описаны как имеющие ту или иную внутреннюю структуру *нечисловые* объекты (дополнительно характеризуемые числовыми значениями существенных параметров) и наделены определенными целевыми свойствами. Математический инструментарий анализа связей вида ОБЪЕКТ=>СВОЙСТВА актуален сегодня для целого ряда дисциплин – от наук о жизни (в частности, анализа физиологических активностей химических соединений), медицинской и технической диагностики, до задач компьютерной обработки больших массивов полнотекстовых документов или же, например, геологического прогнозирования. Предложенные автором диссертационного исследования возможности оптимизации (а в ряде случаев - существенного сокращения) объемов вычислений при порождении зависимостей из структурных эмпирических данных актуальны также и для проблематики обработки больших данных (Big Data).

Структура диссертационной работы соответствует традиционной для исследований в области искусственного интеллекта схеме:

- выделен целевой класс задач (подлежащих детальному исследованию),
- определен набор эвристик, которыми специалист-исследователь обычно пользуется при поиске решения задач этого типа,
- предложена математическая формализация этого комплекса эвристик, что позволяет создавать на ее основе компьютерные системы ИАД,
- продемонстрирована работоспособность созданного инструментария анализа данных (комплекса математических моделей, методов и алгоритмов) для решения прикладных задач в нескольких значимых предметных областях.

В первой главе дана достаточно полная характеристика особенностей таких задач обучения на прецедентах, представленных как объекты нечислового характера (для которых «внутренняя структура» играет существенную роль в анализе данных и принятии финального решения). Рассмотрены примеры важных прикладных проблем, формализуемых как задачи эмпирической индукции на нечисловых объектах: анализ свойств химических соединений, автоматическая обработка больших коллекций текстов на естественном языке и др. .

Функциональные возможности класса эвристик, ассоциируемых с понятием эмпирической индукции (индуктивного обучения на примерах и контрпримерах, рассуждений по аналогии и порождения абдуктивных объяснений для формируемых гипотез, эмпирических зависимостей и закономерностей) детально рассмотрены в соответствующих разделах второй и третьей глав.

Собственно математическая часть диссертационной работы посвящена исследованию четырех (формирующих единый комплекс) взаимосвязанных классов проблем, которые автор идентифицирует и решает в рамках выполненного им процесса формализации эмпирической индукции:

- описанию анализируемой процедурной конструкции алгебраическими средствами и выделению области ее математически корректного использования;
- определению операции сходства для используемых типов данных и доказательству формальной корректности каждого из таких определений;
- построению оценок вычислительной сложности переборных задач, возникающих при порождении эмпирических зависимостей рассматриваемого типа, и
- разработке средств оптимизации комбинаторного перебора вариантов (метода последовательных приближений) при расширениях неявно заданных примерами и контрпримерами зависимостей на новые объекты. Итак:

1. Сначала в главе 2 строится алгебраическая формализация процедурной схемы эмпирической индукции (интеграции индуктивного обучения на примерах и контрпримерах, переноса в рамках рассуждений по аналогии результатов индуктивного обучения на новые объекты и контроля достаточности оснований для принятия порождаемых результатов путем формирования абдуктивных объяснений исходно заданных прецедентов построенными эмпирическими зависимостями). В ее основе – уточнение сходства прецедентов как бинарной алгебраической операции. Базовая схема формализации, – стартовав с имеющихся прецедентов и уточнения операции сходства, перейти к классам сходства и далее – к восстановлению покрывающих их классов эквивалентности прецедентов каждого знака (отдельно –

примеров и контрпримеров). Вместе с обосновывающими ее математическими утверждениями эта схема позволяет сформировать ясную процедурную конструкцию, корректность использования которой легко контролируется выполнимостью специальных формальных условий вхождения новых объектов в уже построенные классы эквивалентности. Для порождения восстанавливаемых классов эквивалентности на множествах прецедентов вводится оператор замыкания, позволяющий единообразным путем работать со всеми рассматриваемыми типами структурных данных. Результатом восстановления классов эквивалентности прецедентов является особый тип комбинаторных объектов – диаграммы взаимной вложенности таких классов. Показано что структурные свойства диаграмм вложенности противоположных знаков – диаграмм примеров и диаграмм контрпримеров – полностью определяют корректность извлечения эмпирических зависимостей из исходно заданного множества прецедентов. Изучены возможности дополнения таких диаграмм до решеток (в общем случае - недистрибутивных).

2. В третьей главе, предложены варианты математически корректной (т.е. формирующей в каждом конкретном случае соответствующую нижнюю полурешетку) формализации сходства на описаниях прецедентов различных типов (множествах, мультимножествах, множествах с отношениями на их элементах, векторах числовых значений параметров и др.). При этом доказательство корректности определения операции сходства на векторах числовых значений (т.е. демонстрация того факта, что порождаемая на прецедентах - числовых векторах - алгебра действительно является нижней полурешеткой и при этом адекватно отражает содержательную семантику решаемой задачи извлечения зависимостей из данных) оказывается технически достаточно сложной математической конструкцией. Тем не менее, именно таким путем удастся организовать распространение уже представленной ранее процедурной схемы обучения на прецедентах также и на кортежи вида *<структурное описание; вектор числовых значений существенных параметров>*.

3. Четвертая глава полностью посвящена анализу комбинаторных характеристик диаграмм взаимной вложенности классов эквивалентности прецедентов - оценкам вычислительной сложности возникающих при построении таких диаграмм переборных задач. Здесь автором диссертационной работы получен ряд новых интересных математических результатов, характеризующих размеры (емкость) подобных диаграмм, сложность восстановления их верхних и нижних границ, а также сложность порождения таких элементов рассматриваемых диаграмм, которые удовлетворяли бы определенным ограничениям как на структуру порождающего их результата выполнения операции сходства, так и на состав формирующего их подмножества исходных прецедентов. Показано, что экспоненциально быстро растущие оценки сложности вычислений для основных переборных задач здесь сохраняются не только для собственно структурных данных (множеств, графов и т.п.), но и для дополняющих эти структуры векторов числовых значений параметров. Таким образом показана однородная (общая для различных типов данных) комбинаторная природа задач формализации структурной эмпирической индукции. Получен ряд технически сложных математических результатов (в частности, о NP-полноте и перечислительной полноте соответствующих комбинаторных задач). Выявлен (и использован при доказательстве ряда важных утверждений в теорети-

ко-множественном случае) эффект двойственности, фактически аналогичный хорошо известному эффекту двойственности в линейном программировании.

4. В пятой главе построена эффективная процедурная схема восстановления приближенных описаний изучаемых диаграмм вложимости – метод последовательных приближений, позволяющий целенаправленно управлять последовательностью порождения изучаемых диаграмм (выделяя в первую очередь специальные быстро-разрешимые частные случаи рассмотренных в главе 4 общих переборных задач). Процедурная схема восстановления диаграмм вложимости стартует с построения полиномиально быстро восстанавливаемых поддиаграмм специального вида – так называемых каркасов, а на последующих шагах целенаправленно формируются и экспоненциально сложные структурные фрагменты. Предложены средства управления выбором той или иной конкретной последовательности реконструкции подобных фрагментов, а также «глубиной» восстановления каждого из них. Доказаны математические утверждения, определяющие корректность предложенного метода последовательных приближений. Отдельно считаю необходимым отметить предложенный автором вариант реализации этого метода в режиме параллельных вычислений. Параллельная версия предлагаемого математического инструментария ИАД позволяет существенным образом расширить область его применения в практически значимых приложениях.

В Главе 6 представлены два дополнительных математических результата, отражающих специфику формализации структурной эмпирической индукции. Это – доказательство функциональности используемого отношения причинности, и демонстрация эффектов немонотонности (ненаследуемости) порождения соответствующих эмпирических зависимостей при расширении исходно заданной выборки прецедентов новыми объектами.

Представленный в главе 8 подробный обзор приложений демонстрирует работоспособность созданного комплекса математических моделей, методов и алгоритмов на примерах решения прикладных задач ИАД в нескольких значимых предметных областях. Здесь, наряду с приложениями в области анализа физиологических активностей химических соединений или же компьютерной обработки больших массивов полнотекстовых документов, интересными и перспективными представляются приложения развиваемого в диссертационной работе подхода в задачах управления потоками данных (быстрой адресации пакетов) и обеспечения информационной безопасности (быстрого анализа вложимости вредоносного кода в пакеты сообщений) в компьютерных сетях.

Новизна, теоретическая и практическая значимость результатов выполненного М.И. Забежайло исследования позволяет квалифицировать его диссертационную работу как важное научное достижение в области теоретической информатики.

Оформление диссертационной работы соответствует требованиям ВАК РФ. Автореферат полностью отражает содержание диссертации. Все положения диссертационной работы достоверны, выводы и заключения обоснованы. Используя строгие математические методы при проведении теоретического исследования, автор приходит к важным выводам относительно комбинаторных свойств задач обучения на прецедентах нечислового характера. Доказанные в работе утверждения

позволяют построить целостное описание процедурной конструкции извлечения эмпирических зависимостей из данных, наделенных сложной внутренней структурой. Работоспособность сформированной конструкции подтверждена ее успешными применениями при решении прикладных задач ИАД в ряде (различных по своей природе) предметных областей. Результаты диссертационного исследования опубликованы в 37 статьях, в том числе – в 17 работах в изданиях, рекомендованных ВАК РФ для публикации основных научных результатов диссертаций на соискание ученой степени доктора наук. Полученные автором результаты были представлены на отечественных и международных конференциях, научных семинарах.

Автор хорошо знаком с литературой по искусственному интеллекту и ИАД, диссертационная работа содержит весьма обширную библиографию по этой тематике. Материал хорошо структурирован, изложение характеризуется логической последовательностью. Текст написан, в основном, хорошим языком, стиль изложения четкий и понятный.

О недостатках работы. Столь подробный вариант изложения полученных результатов подходит, на мой взгляд, скорее для монографии: так значительную часть содержания Главы 8 можно было бы представить более компактно (фокусируясь лишь на постановках прикладных задач и комментариях полученных экспериментальных результатов), или же, например, оформить (вместе с Главой 7) как Приложения к основному тексту диссертации.

Перечисленные недостатки не умаляют значимости представленных в рассматриваемой диссертационной работе результатов и выводов и не оказывают влияния на ее общую положительную оценку.

Диссертационная работа М.И.Забейайло «Комбинаторные средства формализации эмпирической индукции» полностью удовлетворяет критериям, установленным «Положением о присуждении ученых степеней», утвержденным Постановлением Правительства РФ от 24 сентября 2013 г. № 842. Данная работа соответствует всем требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени доктора физико-математических наук, а ее автор – Забейайло Михаил Иванович - заслуживает присуждения ему ученой степени доктора физико-математических наук по специальности 05.13.17 – Теоретические основы информатики.

Член-корреспондент РАН,
д.т.н., профессор

« 9 » июня 2015 года



Желтов Сергей Юрьевич

Генеральный директор ФГУП «Государственный научно-исследовательский институт авиационных систем».

Телефон: (499) 157-94-08

Адрес электронной почты: zhl@gosniias.ru

Подпись д.т.н., проф. Желтова С.Ю. заверяю

Ученый секретарь ФГУП «ГосНИИАС»

д.т.н., проф.



Мужичек С.М.