

УТВЕРЖДАЮ

Проректор по научной работе и
стратегическому развитию
Московского физико-технического
института (государственного универ-
ситета),
член-корреспондент РАН



Аушев Тагир
Абдул - Хамидович

2016 г.

ОТЗЫВ

ведущей организации на диссертацию **Забейайло Михаила Ивановича**
«Комбинаторные средства формализации эмпирической индукции»,
представленную на соискание ученой степени доктора
физико-математических наук по специальности 05.13.17
– «Теоретические основы информатики»,

Диссертационная работа М.И.Забейайло посвящена разработке математических моделей, методов и алгоритмов обучения на прецедентах, развиваемых в рамках одного из наиболее актуальных разделов современной теоретической и прикладной информатики – интеллектуального анализа данных (ИАД). Основным объектом исследования выбран сравнительно мало изученный вариант задачи обучения на примерах и контрпримерах, представленных как нечисловые объекты (внутренняя структура которых существенна с точки зрения порождения извлекаемых из них эмпирических зависимостей).

Математический инструментарий анализа данных и поддержки принятия решений важен для целого ряда областей приложений ИАД, в том числе – интенсивно развивающихся в последнее время задач компьютерной обработки эмпирических данных в области наук о жизни (создания новых лекарственных средств, выявления потенциально опасных химических соединений в окружающей человека среде, доказательной медицине, медицинской диагностике и др.), в задачах автоматической обработки больших коллекций полнотекстовых документов, задачах поддержания работоспособности сложных технических систем (в т.ч. - идентификации компьютерных атак и противодействия их последствиям) и др.. В отличие от традиционно используемых в рассматриваемой области метрических уточнений понятия сходства анализируемых прецедентов, автор развивает альтернативный подход – формализацию сходства как алгебраической операции (что позволяет эффективно оперировать с «внутренними структурами» анализируемых прецедентов). Допол-

нительные возможности открываются здесь и за счет расширения предлагаемой алгебраической формализации сходства на операции с векторами числовых значений релевантных параметров, дополняющих структурное описание анализируемых прецедентов. Использование таких нетрадиционных математических инструментов анализа данных открывает автору рассматриваемой работы возможности оперировать в том числе и *малыми* (статистически незначимыми) *коллекциями прецедентов*, а также вести *каузальный анализ* - выделять структурные *причины* (комбинации актуальных «факторов влияния»), которые и определяют в каждом конкретном случае формирование эмпирических зависимостей со свойственной именно ему внутренней «архитектурой». Таким образом, актуальность темы данного исследования, посвященного разработке адекватного математического инструментария для формализации, анализа и эффективного решения (в том числе – в режиме параллельных вычислений) соответствующих задач автоматического извлечения «структурных» зависимостей из исходно заданных эмпирических данных, представляется бесспорной.

Диссертационная работа состоит из введения, шести глав, заключения и восьми приложений.

Глава 1 носит вводный характер и представляет общие характеристики области ИАД а также позиционирование в ней проблематики Эмпирической Индукции (ЭИ) - математических моделей и методов восстановления зависимостей из накапливаемых эмпирических данных. Задача формализации ЭИ рассматривается в контексте современных исследований в области искусственного интеллекта – как задача адекватной (т.е. позволяющей контролировать *доказуемую корректность* использования порождаемых результатов) формализации ассоциируемого с понятием ЭИ комплекса эвристик (индуктивного обучения, рассуждений по аналогии и абдуктивных объяснений).

В Главе 2 предложена унифицированная (позволяющая оперировать различными типами описаний исходных прецедентов) схема формализации процедурной конструкции ЭИ алгебраическими средствами. Показано, что уточнение сходства прецедентов как бинарной алгебраической операции позволяет порождать классы сходства и, далее, классы эквивалентности прецедентов, а задача восстановления зависимостей может быть сведена к задаче восстановления подобных классов эквивалентности и анализу пересечений таких классов эквивалентности на примерах и на контрпримерах. Демонстрируется, что диаграммы взаимной вложенности классов эквивалентности прецедентов (отдельно - примеров и контрпримеров) полностью определяют область корректности извлекаемых из исходных данных эмпирических зависимостей (в том числе – корректность их распространения на новые, ранее не изученные прецеденты). Для формирования классов эквивалентности на различных типах описаний исходных прецедентов вводится специальный оператор *Closure*, который, как показано, оказывается оператором замыкания на множестве прецедентов. В этой же Главе рассмотрены функциональные свойства и особенности семантики формализуемого комплекса эвристик Эмпирической Индукции.

Глава 3 посвящена формальному уточнению представленных в Главе 2 базовых элементов развиваемого подхода:

- формализации сходства как алгебраической операции при работе с различными вариантами описаний элементов из исходной выборки прецедентов вида *<множества признаков и отношения на них; релевантные числовые данные>*. Доказаны соответствующие утверждения о корректности определения сходства для работы с числовыми значениями параметров,
- детализации процедурной схемы использования формализуемого комплекса эвристик (эмпирической индукции в стиле Ф.Бэкона-Д.С.Милля, выводов по аналогии, порождения абдуктивных объяснений в стиле Ч.С.Пирса и др.),
- обсуждению возможностей так называемых квази-аксиоматических теорий как средства представления знаний в компьютерных системах ИАД, ориентированных на решение задач Эмпирической Индукции,
- детальному описанию используемых инструментов анализа данных: так называемых решающих предикатов, правил правдоподобного вывода и стратегий формализованных рассуждений.

В Главе 4 представлен комплекс утверждений, которые определяют вычислительную сложность основных комбинаторных задач, характеризующих диаграммы вложенности классов эквивалентности прецедентов для рассматриваемых в работе типов данных. Установлено, что емкость диаграмм обсуждаемого типа растет экспоненциально быстро с линейным ростом объема исходной выборки прецедентов, а задача о числе элементов в таких диаграммах принадлежит классу перечислительно полных задач. Показано, что верхняя и нижняя границы анализируемых диаграмм могут быть порождены полиномиально сложными алгоритмами. Получены оценки вычислительной сложности поиска в таких диаграммах элементов с заданными ограничениями на структуру и емкость порождаемых классов эквивалентности (в т.ч. – на размеры классов и параметры порождающего их сходства). В частности, показано, что задача поиска класса эквивалентности, который порождается заданным числом прецедентов из исходной выборки, принадлежит классу *NP*-полных задач. В конечном итоге, продемонстрировано, что комбинаторно-трудная природа рассматриваемых здесь основных переборных задач – *свойство*, характерное даже не для тех или иных определенных типов описания исходных данных, а собственно для формализуемой задачи Эмпирической Индукции.

В Главе 5 предложен метод дискретной оптимизации комбинаторного перебора вариантов при восстановлении диаграмм взаимной вложенности классов эквивалентности прецедентов, описываемых множествами признаков. Показано, что каждая диаграмма может быть представлена объединением поддиаграмм специального вида - так называемых псевдо-деревьев. Предложена технология порождения приближенных описаний таких псевдо-деревьев с помощью полиномиально быстро формируемых поддиаграмм – каркасов – и (диагностируемых при порождении каркасов) экспоненциально-сложных архитектурных фрагментов. Для восстановления последних разработан метод последовательных приближений (целенаправленного порождения все более сложных описаний архитектурных фрагментов подобного типа), оперирующий специальным сжимающим отображением на множестве исходно заданных прецедентов (которое позволяет последовательно переходить к целенаправленно выбираемым подзадачам при восстановлении каждого анализируемого псевдо-дерева), а также специальным оператором, монотонно расширяющим

текущее приближенное описание реконструируемого псевдо-дерева вновь порождаемыми при анализе подзадач архитектурными фрагментами.

Показано, что этот *метод последовательных приближений* позволяет формировать целевые диаграммы вложимости в режиме *параллельных* вычислений. Дано строгое обоснование корректности разработанного метода последовательных приближений. Описан механизм целенаправленного управления порождением последовательных приближений *заданной точности*. Эти результаты имеют центральное значение для всей диссертационной работы, т.к. определяют алгоритмически корректную и строго обоснованную процедурную конструкцию, которая позволяет (не взирая на ряд имеющихся ограничений в области сложности вычислений – см. выше - Главу 4) оперировать при порождении эмпирических зависимостей из данных и прогнозировании свойств новых объектов выборками прецедентов не ограничиваемого размера.

В Главе 6 рассмотрен ряд дополнительных особенностей предложенного инструментария ИАД (характерные для задач Эмпирической Индукции эффекты логико-математической *немонотонности* формализованных рассуждений, *функциональность* используемого отношения причинности и др.).

В Приложениях представлены как детальные примеры использования, так и подробный обзор применения разработанных в диссертационной работе моделей и средств ИАД в компьютерном анализе свойств физиологически активных веществ, автоматической обработке полнотекстовых документов, геологическом прогнозировании и др., подтверждающий практическую значимость и эффективность созданных математических моделей, методов, алгоритмов и инструментов ИАД.

Принципиальных замечаний по сути представленных в диссертационной работе результатов и выводов нет. Однако, следует отметить, что:

1. В тексте используется сквозная нумерация утверждений – Определений, Теорем, Лемм, Следствий, Замечаний и т.п. (например: 1.2.3.4 Утверждение, 4.3.2.1 Теорема и т.п.), ведущая в ряде случаев к коллизиям, когда в тексте имеется, например, *Теорема X.X.X.X*, однако отсутствует *Теорема X.X.X.X-1*.
2. При обсуждении сходства на различных типах данных автор иллюстрирует определения примерами. По-видимому, выбранный стиль изложения лишь выиграл бы, если бы автор не ограничился рассмотрением лишь простейших примеров (в частности, если бы при определении сходства на числовых векторах рассматривались также примеры *m*-матриц как с пересекающимися множествами столбцов, так и с разделенными дополнительными множествами столбцов матрицы-универсума).

В качестве пожелания по дальнейшим действиям, представляется целесообразным подготовить на базе текста диссертационной работы монографию и учебное пособие по проблематике использования современных математических моделей и методов искусственного интеллекта в ИАД и его приложениях.

В целом, диссертация М.И.Забейхайло является завершенным научным исследованием в области интеллектуального анализа данных, полученные в ней результаты характеризуются новизной и практической значимостью. Все положения диссертационной работы достоверны, сформулированные в ней выводы и заключения соответствующим образом обоснованы. Опираясь на строгие математические ме-

тоды, автор приходит к важным выводам, определяющим общие для различных типов данных комбинаторные свойства задач обучения на прецедентах нечислового характера. Доказанные в работе утверждения позволяют сформировать целостное описание комплекса моделей и алгоритмов извлечения эмпирических зависимостей из данных, наделенных сложной внутренней структурой. Работоспособность сформированной алгоритмической конструкции подтверждена ее успешными применениями при решении прикладных задач ИАД.

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 научных семинарах, отечественных и международных конференциях по искусственному интеллекту и интеллектуальному анализу данных.

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

Диссертация и отзыв обсуждены и одобрены на заседании семинара Кафедры информатики и вычислительной математики Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (государственный университет)». Протокол № 6 от 21 октября 2016 года.

Член-корреспондент РАН, д.ф.-м.н, профессор,
заведующий кафедрой информатики
и вычислительной математики



И.Б. Петров

«25» ноября 2016 года

Почтовый адрес:

141700, Московская область, г. Долгопрудный, Институтский пер., 9. МФТИ

Телефон: +7 (495) 408-45-54; +7 (495) 408-66-95

Организация – место работы: федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (государственный университет)», кафедра информатики и вычислительной математики

Должность: заведующий кафедрой

Адрес электронной почты: petrov@mipt.ru

Web-site: <http://mipt.ru/>