ОТЗЫВ

официального оппонента доктора технических наук Спицына Владимира Григорьевича на диссертацию Кузьмича Романа Ивановича на тему «Модифицированный метод логического анализа данных для задач классификации», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка информации (информатика, вычислительная техника и управление)

Актуальность избранной темы работы. В настоящее время при решении практических задач классификации часто требуется знать, почему новое наблюдение принадлежит определенному классу, насколько это решение обоснованно. В таких случаях нужен метод классификации данных, который помимо самого решения представит в явном виде решающее правило, то есть выявляет знания из имеющихся данных. Это справедливо для метода логического анализа данных, принцип работы которого состоит в выявлении закономерностей в данных и формализации их в виде набора правил, то есть набора закономерностей, описываемых простой логической формулой. В предлагаемом методе несомненную важность и актуальность имеют проблемы, связанные с построением оптимизационных моделей для формирования информативных закономерностей и получением интерпретируемого классификатора с высокой обобщающей способностью. Представленная диссертационная работа посвящена решению поставленных проблем путем разработки модификаций для метода логического анализа данных, позволяющих повысить интерпретируемость классификатора и качество классификации новых наблюдений.

Степень достоверности и обоснованности научных положений, выводов и рекомендаций. Обоснованность научных результатов, полученных в диссертации, подтверждается корректным использованием теорий системного анализа, комбинаторики, методов оптимизации. Все утверждения обоснованы и подтверждены надлежащими аргументами, исходные утверждения подтверждены ссылками на источники.

Достоверность результатов работы подтверждается теоретическими и экспериментальными данными, опубликованными в 15 работах, в том числе 5 – в изданиях из перечня ВАК, по программной реализации теоретических результатов получено свидетельство о государственной регистрации. Результаты диссертационного исследования обсуждались на всероссийских и международных научных конференциях и семинарах.

Структура и содержание. Диссертационная работа состоит из введения, трех глав, заключения, списка литературы из 115 источников и 2 приложений. Основной текст диссертации содержит 121 страницу, 10 рисунков и 19 таблиц.

Первая глава диссертации посвящена обзору наиболее распространенных логических алгоритмов классификации, алгоритмов поиска закономерностей в массивах данных, а также обзору основных программных систем, решающих задачи обнаружения закономерностей в данных. Рассматриваются такие алгоритмы как решающие списки, решающие деревья, алгоритмы простого и взвешенного голосования правил. Обсуждаются их преимущества и недостатки. Также в главе приводится анализ программных систем для решения обсуждаемых задач. Отмечены два направления развития соответствующих программных средств: узкоспециализированные пакеты, которые направлены на небольшой круг практических задач, и программные средства широкого назначения, использующие разнообразные методы и подходы.

Вторая глава посвящена описанию основных этапов метода логического анализа данных, созданию оптимизационных моделей для формирования закономерностей и разработке алгоритмических процедур, позволяющих улучшить интерпретируемость классификатора.

Разработаны алгоритмические процедуры наращивания закономерностей, выбора базовых наблюдений для формирования закономерностей, построения классификатора как композиции информативных закономерностей.

На основе разработанных алгоритмических процедур предложены модификации для метода логического анализа данных, позволяющие повысить интерпретируемость классификатора за счет сокращения числа правил в нем,

сохраняя при этом высокую точность классификации при решении практических задач.

Представлена модель оптимизации для формирования закономерностей с покрытием существенно различных подмножеств наблюдений обучающей выборки, которая позволяет повысить обобщающие способности классификатора, получаемого на базе данных правил.

Третья глава посвящена программной реализации предложенного метода с демонстрацией его работоспособности на реальных задачах. С помощью метода логического анализа данных в диссертационной работе решены следующие задачи классификации:

– выявление спама;

– классификация результатов радарного сканирования ионосферы;

– прогнозирование осложнений инфаркта миокарда (фибрилляция предсердий, фибрилляция желудочков, разрыв сердца, отек легких, летальный исход).

Показано, что разработанные модификации для метода логического анализа данных доказали свою эффективность при решении практических задач. Приведено сравнение по точности различных алгоритмов классификации с методом логического анализа данных.

Содержание автореферата соответствует содержанию диссертации.

Новизна научных результатов.

В ходе разработки метода автором получены следующие новые научные результаты:

1) на основе алгоритма «k-средних» разработана алгоритмическая процедура целенаправленного выбора исходных наблюдений для выявления закономерностей в данных;

2) разработана алгоритмическая процедура наращивания закономерностей с максимальным покрытием наблюдений обучающей выборки;

3) решена задача формирования закономерностей на основе постановки и решения задачи оптимизации специального вида, отличающейся наличием в целевой функции весового коэффициента покрываемого наблюдения (с возможностью захвата наблюдений другого класса);

4) разработана алгоритмическая процедура построения классификатора в виде композиции информативных закономерностей за счет совместного использования критерия бустинга для оценки информативности закономерностей и итеративной процедуры выбора порога информативности;

5) модифицирован метод логического анализа данных на основе разработанных алгоритмических процедур.

Теоретическая и практическая значимость диссертации. На основе результатов диссертационного исследования автором исследован и модифицирован метод логического анализа данных, основанный на использовании оптимизационных моделей для формирования информативных закономерностей и алгоритмических процедур сокращения количества правил в классификаторе, что является существенным вкладом в теорию интеллектуальных технологий и представления знаний, практику их применения в системах обработки информации и интеллектуального анализа данных. Метод реализован в программной системе поддержки принятия решений, которая позволяет эффективно решать практические задачи классификации в различных областях человеческой деятельности. В ходе выполнения работы успешно решены задачи: выявление спама, радарное сканирование ионосферы, прогнозирование осложнений инфаркта миокарда.

Замечания по диссертационной работе. По диссертационной работе и автореферату можно выделить следующие замечания:

1. При тестировании разработанного алгоритма использовались данные из репозитория машинного обучения UCI, который содержит 255 различных наборов данных для задачи классификации. В связи с этим возникает вопрос, почему автор при проведении тестирования использовал только два набора данных из 255?

2. При описании сравнения предложенного алгоритма с методом, основанным на применении искусственной нейронной сети (ИНС), не описаны тип используемой ИНС, количество слоев, нейронов, алгоритм обучения и другие параметры ИНС (с. 86-87). Для решения задачи разделения объектов на два класса достаточно успешно используют метод опорных векторов, сравнение результатов предложенного алгоритма с результатами применения метода опорных векторов позволило бы наиболее четко выразить преимущества и недостатки предложенного подхода.

3. При сравнении предложенного в работе алгоритма LDA с существующими логическими алгоритмами классификации, автор использует результаты работы алгоритмов, реализованных в системе WEKA, игнорируя результаты полученные другими исследователями на этих же наборах данных, опубликованные в научных статьях. В частности, не упоминается алгоритм SLIPPER (W. Cohen, Y. Singer. A Simple, Fast and Effective Rule Learner // Proceedings of the sixteenth national conference on Artificial intelligence, AAAI '99/IAAI '99, 1999, pp. 335-342), который при тестировании на 32 наборах данных из репозитория UCI показал значительно лучшее качество классификации на большинстве тестов по сравнению с алгоритмами RIPPER и C4.5, с которыми сравнивался предложенный автором алгоритм LDA.

4. Не приведена оценка скорости работы предлагаемого алгоритма. В процессе работы алгоритм многократно решает задачу условной псевдобулевой оптимизации для минимизации опорного множества. При этом размерность пространства поиска равна количеству признаков, а сложность вычисления ограничивающей функции зависит от размера обучающей выборки. В связи с этим возникает вопрос о применимости разработанного алгоритма для решения задач классификации векторов из тысяч признаков на выборках большого объема.

Отмеченные замечания не снижают общей положительной оценки диссертации.

Заключение. Диссертационная работа Кузьмича Р. И. «Модифицированный метод логического анализа данных для задач классификации» является завершенной научно-квалификационной работой, в которой на основании выполненных автором исследований содержится решение задачи разработки модификаций для метода логического анализа данных, имеющей значение для развития теории и практики интеллектуального анализа данных. В целом диссертационная работа соответствует требованиям п.9 «Положения о порядке присуждения учёных степеней» постановления Правительства Российской Федерации от 24.09.2013 г. № 842, а ее автор Кузьмич Роман Иванович заслуживает присуждения степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка информации (информатика, вычислительная техника и управление).

Официальный оппонент,
профессор кафедры вычислительной техники
Федерального государственного автономного
образовательного учреждения высшего
образования «Национальный исследовательский
Томский политехнический университет»,
доктор технических наук, профессор Спицын Владимир Григорьевич

Подпись В.Г. Спицына заверяю
Ученый секретарь
Федерального государственного автономного
образовательного учреждения высшего
образования «Национальный исследовательский
Томский политехнический университет» О.А. Ананьева

25 марта 2016 года

Почтовый адрес: 634050, г. Томск, проспект Ленина, дом 30, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский политехнический университет». Телефон: (3822) 701-609, e-mail: spvg@tpu.ru.