На правах рукописи

Кузьмич Роман Иванович

МОДИФИЦИРОВАННЫЙ МЕТОД ЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ

05.13.01 – Системный анализ, управление и обработка информации
(информатика, вычислительная техника и управление)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Красноярск 2016

Работа выполнена в ФБГОУ ВО «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева» (СибГАУ), г. Красноярск

Научный руководитель: кандидат физико-математических наук, доцент
Масич Игорь Сергеевич

Официальные оппоненты: **Аршинский Леонид Вадимович**
доктор технических наук, доцент
ФГБОУ ВО «Иркутский государственный университет путей сообщения», кафедра информационных систем и защиты информации, заведующий кафедрой

Спицын Владимир Григорьевич
доктор технических наук, профессор
ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», кафедра вычислительной техники, профессор

Ведущая организация: ФГБОУ ВО «Воронежский государственный технический университет» (г. Воронеж)

Защита состоится «21» апреля 2016 года в 14:00 часов на заседании диссертационного совета Д 999.007.02 на базе Сибирского федерального университета и Института вычислительного моделирования СО РАН по адресу: 660074, г. Красноярск, ул. Академика Киренского, 26, УЛК 115.

С диссертацией можно ознакомиться в библиотеке и на сайте Сибирского федерального университета www.sfu-kras.ru.

Автореферат разослан «___» марта 2016 года.

Ученый секретарь
диссертационного совета Бронов Сергей Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. В настоящее время при решении задач распознавания образов, помимо требования высокой точности, часто возникает необходимость в интерпретируемости и обоснованности получаемых решений. Особенно интерпретируемость и обоснованность являются ключевыми факторами при решении тех практических задач, в которых потери от принятия неверного решения могут быть велики. Поэтому система поддержки принятия решений, используемая для таких задач, должна обосновывать возможные решения и интерпретировать результат.

Для создания такой системы потребуются алгоритмы классификации данных, которые помимо самого решения предоставляют в явном виде решающее правило, то есть выявляют знания из имеющихся данных. Это справедливо для логических алгоритмов классификации, принцип работы которых состоит в выявлении закономерностей в данных и формализации их в виде набора правил, т.е. набора закономерностей, описываемых простой логической формулой.

Процесс формирования логических правил сопровождается решением задач выбора наилучших альтернатив в соответствии с некоторым критерием. В предлагаемом методе логического анализа данных формализация процесса формирования логических правил осуществляется в виде ряда задач комбинаторной оптимизации, что формирует гибкий и эффективный алгоритм логического анализа для классификации данных. Объединив некоторое количество закономерностей в композицию, получаем классификатор, который решает поставленную задачу.

Однако в настоящее время существует ряд проблем, связанных с применением метода логического анализа данных при решении практических задач классификации. Одной из них является построение оптимизационных моделей для формирования информативных закономерностей. При рассмотрении данного вопроса, прежде всего, необходимо определиться с теми критериями и ограничениями, которые лежат в основе этих оптимизационных моделей. Другой проблемой исследуемого метода является построение классификатора, который смог бы верно отнести новое наблюдение, т.е. наблюдение, не принимавшее участие при его построении, к тому или иному классу. Основной задачей на данном этапе метода является повышение интерпретируемости классификатора и качества классификации новых наблюдений, т. е. улучшение обобщающих способностей классификатора.

Таким образом, разработка модификаций для метода логического анализа данных, позволяющих улучшить интерпретируемость и обобщающие способности классификатора, является актуальной научно-технической задачей.

Следует отметить, что большой вклад в развитие логических алгоритмов классификации внесли следующие ученые: Ю. И. Журавлев,

К. В. Рудаков, К. В. Воронцов, Н. Г. Загоруйко, Г. С. Лбов, Е. В. Дюкова, О. В. Сенько, В. И. Донской, P. L. Hammer, G. Alexe, S. Alexe, Y. Freund, R. E. Schapire.

Цель диссертационной работы состоит в повышении точности решения задач классификации и улучшении интерпретируемости классификатора, основанного на логических закономерностях.

Поставленная цель определила необходимость решения следующих задач:

1. Провести анализ существующих логических алгоритмов классификации, алгоритмов поиска информативных закономерностей для них, и основных программных систем, решающих практические задачи классификации.
2. Разработать алгоритмическую процедуру выбора базовых наблюдений для формирования закономерностей в методе логического анализа данных.
3. Разработать алгоритмическую процедуру улучшения закономерностей для повышения их информативности и усиления обобщающих способностей классификатора, построенного на базе данных закономерностей.
4. Создать модель оптимизации для формирования закономерностей, покрывающих существенно различные подмножества наблюдений обучающей выборки в методе логического анализа данных.
5. Разработать алгоритмическую процедуру построения классификатора, учитывающую информативность закономерностей, для метода логического анализа данных.
6. Модифицировать метод логического анализа данных на основе разработанных алгоритмических процедур.
7. Алгоритмизировать и реализовать метод логического анализа данных в виде программной системы, провести его апробацию и сравнительный анализ по точности с другими алгоритмами классификации на практических задачах.

Методы исследования. В диссертационной работе использовались методы системного анализа, теория множеств, теория вероятностей, комбинаторика, методы оптимизации.

Новые научные результаты, выносимые на защиту:

1. Разработана алгоритмическая процедура выбора базовых наблюдений для формирования закономерностей, отличающаяся от известных целенаправленным выбором базовых наблюдений, получаемых путем применения алгоритма «k-средних» к множеству наблюдений обучающей выборки, позволяющая сократить количество правил в классификаторе и снизить трудоемкость его построения при сохранении высокой точности.
2. Разработана алгоритмическая процедура наращивания закономерностей, полученных на базе оптимизационной модели с максимальным покрытием наблюдений обучающейся выборки,

позволяющая повысить информативность правил, тем самым, способствуя увеличению точности принимаемых классификатором решений.

3. Создана модель оптимизации для формирования закономерностей, отличающаяся от известных наличием в целевой функции весового коэффициента покрываемого наблюдения, а также возможностью захвата наблюдений другого класса, позволяющая формировать правила, которые выделяют существенно различные подмножества наблюдений обучающей выборки.

4. Разработана алгоритмическая процедура построения классификатора как композиции информативных закономерностей, отличающаяся от известных совместным использованием критерия бустинга для оценки информативности закономерностей и новой итеративной процедуры выбора порога информативности, позволяющая сократить количество правил в классификаторе при сохранении высокой точности.

5. Модифицирован метод логического анализа данных на основе разработанных алгоритмических процедур, позволяющих повысить интерпретируемость классификатора, сокращая количество правил в нем, и сохранить при этом высокую точность при решении практических задач классификации.

Теоретическая значимость результатов диссертационного исследования состоит в разработке и исследовании модификаций для метода логического анализа данных, основанных на создании 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, практики их применения в системах обработки информации и интеллектуального анализа данных.

Практическая значимость. На основе метода логического анализа данных реализована программная система поддержки принятия решений, которая позволяет, используя рекомендации по настройке ее параметров, широкому кругу специалистов эффективно решать практические задачи классификации.

Материалы диссертационного исследования и разработанная программная система использованы для решения следующих практических задач: классификация результатов радарного сканирования, выявление спама, прогнозирование осложнений инфаркта миокарда.

Достоверность и обоснованность результатов диссертации подтверждается: исследованием существующих логических алгоритмов классификации и алгоритмов поиска информативных закономерностей для них, корректным обоснованием постановок задач, результатами применения предложенных моделей, методов и алгоритмических процедур, сравнительным анализом по точности с существующими алгоритмами классификации на практических задачах.

Реализация результатов работы. Диссертационная работа поддержана Фондом содействия развития малых форм предприятий в научно-технической сфере по программе «У.М.Н.И.К.» («Участник молодежного научно-инновационного конкурса») в рамках НИОКР «Разработка программной системы на базе логических алгоритмов классификации для решения задач медицинской диагностики и прогнозирования» на 2011-2013 гг. Результаты диссертации использовались в гранте Президента РФ МК-463.2010.9 «Комбинаторная оптимизация в задачах распознавания при диагностике и прогнозировании». Разработанная программная система «Логические анализ данных в задачах классификации» зарегистрирована в Реестре программ для ЭВМ 17 марта 2011 г. (свидетельство № 2011612265).

Апробация работы. Основные положения и результаты диссертации докладывались и обсуждались на XIV, XV Международной научной конференции «Решетневские чтения» (г. Красноярск 2010, 2011, 2014); XLIX Международной научной студенческой конференции «Студент и научно-технический прогресс» (г. Новосибирск 2011); III Общероссийской молодежной научно-технической конференции «Молодежь. Техника. Космос» (г. Санкт-Петербург 2011); XIV Международной научно-технической конференции «Фундаментальные и прикладные проблемы приборостроения и информатики» (г. Москва 2011); Всероссийской молодежной научной конференции с международным участием «Современные проблемы фундаментальных и прикладных наук» (г. Кемерово 2011); Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР–2013» (г. Томск 2013).

Публикации. По теме диссертационной работы опубликовано 15 работ, из них 5 в изданиях из перечня ВАК, зарегистрирована программная система в Реестре программ для ЭВМ.

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

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулирована цель и поставлены задачи исследования, рассмотрены вопросы теоретической и практической ценности проведенных исследований, изложены новые научные результаты, выносимые на защиту.

Первая глава посвящена обзору наиболее распространенных логических алгоритмов классификации, алгоритмов поиска закономерностей в форме конъюнкций для них, а также анализу основных программных систем, решающих задачи классификации.

Пусть φ: X $\rightarrow$ {0, 1} – некоторый предикат, определённый на множестве наблюдений X. Предикат φ покрывает наблюдение x, если $\varphi(x) = 1$. Предикат называют закономерностью, если он покрывает

достаточно много наблюдений одного класса, и практически не покрывает наблюдения других классов. Любая закономерность классифицирует лишь часть наблюдений. Объединив определённое количество закономерностей в композицию, можно получить классификатор, способный классифицировать любые наблюдения множества. Логическими алгоритмами классификации будем называть композиции, состоящие из легко интерпретируемых закономерностей. Следует отметить, что чем больше наблюдений своего класса по сравнению с наблюдениями всех других классов покрывает закономерность, тем она более информативна.

Как правило, закономерности строятся в виде конъюнкций, так как они легко интерпретируются. Поиск наиболее информативных конъюнкций в общем случае требует полного перебора, хотя на практике используют различные эвристики для сокращённого целенаправленного поиска конъюнкций, близких к оптимальным. Среди алгоритмов синтеза конъюнкций можно выделить: «градиентный» алгоритм, жадный алгоритм, случайный локальный поиск, генетический алгоритм. Каждый из алгоритмов имеет свои особенности, преимущества и недостатки.

В диссертации рассматриваются основные алгоритмы логической классификации: решающие списки, решающие деревья, алгоритмы простого и взвешенного голосования правил. Приводятся их преимущества и недостатки, способы построения. Отмечено, что для классификации наиболее приемлем алгоритм, основанный на голосовании правил.

В работе приводится анализ программных систем для решения задач классификации. Определено два пути развития программных средств: узкоспециализированные пакеты, которые направлены на небольшой круг практических задач, а их алгоритмической базой является какой-либо один из альтернативных подходов к классификации, и программные средства, основанные на включении основных существующих подходов.

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

В диссертационной работе рассматривается задача классификации следующего вида. Имеется выборка данных, которая состоит из двух непересекающихся множеств Ω^+ и Ω^- n-мерных векторов, принадлежащих соответственно положительному или отрицательному классу. Компоненты вектора, называемые также признаками, могут быть как численными или номинальными, так и бинарными. Задача состоит в том, чтобы для некоторого нового наблюдения, являющегося также вектором n переменных, определить, к какому классу он принадлежит.

В основе предлагаемого подхода к классификации данных лежит метод, происходящий из теории комбинаторной оптимизации и называемый логическим анализом данных. Последовательные элементы метода:

1) Для исключения избыточных переменных в исходной выборке данных во множестве переменных определяется некоторое подмножество *S*, используя которое можно отличать положительные наблюдения от отрицательных. Далее для работы метода используются проекции Ω_s^+ и Ω_s^- множеств Ω^+ и Ω^- на *S*.

2) Множество Ω_s^+ покрывается семейством однотипных подмножеств уменьшенного пространства, каждое из которых имеет значительное пересечение с Ω_s^+, но не пересекается с Ω_s^-, либо допускается небольшое пересечение с Ω_s^- для большего увеличения пересечения с Ω_s^+. Такие подмножества называются «положительными закономерностями». Аналогично множество Ω_s^- покрывается «отрицательными закономерностями».

3) Определяется подмножество положительных закономерностей, объединение которых покрывает все наблюдения Ω_s^+, и подмножество отрицательных закономерностей, объединение которых покрывает все наблюдения Ω_s^-.

4) Положительный или отрицательный характер некоторого наблюдения, покрываемого объединением двух подмножеств модели, определяется с помощью классификатора, основанного на этих подмножествах.

Предлагаемый метод предназначен для работы с выборками данных, где признаки принимают бинарные значения. Так как исходная выборка может состоять из разнотипных признаков, необходимо их бинаризовать.

После бинаризации необходимо осуществить поиск опорного множества признаков, т.е. признаков которые могут отделить с высокой точностью положительные от отрицательных наблюдений. Для этого предлагается подход, базирующийся на отборе подмножества признаков путем построения модели в виде задачи комбинаторной оптимизации.

Множество *S* признаков называется опорным множеством, если проекция Ω_s^+ множества Ω^+ на *S* не пересекается с проекцией Ω_s^- множества Ω^- на *S*. Множество всех признаков является опорным, так как изначально Ω^+ и Ω^- не пересекаются. Опорное множество будем называть минимальным, если, исключив из него любую оставшуюся переменную, получим выборку, в которой некоторые положительное и отрицательное наблюдения совпадают.

Чтобы выявить минимальное опорное множество, в соответствие каждому признаку x_i, i=1,…,t бинарной выборки ставится новая бинарная переменная y_i, которая равна 1, если x_i принадлежит опорному множеству, и равна 0, если нет. Следует обозначить $U=(u_1,u_2,\ldots,u_t)$ – бинарный вектор, ассоциированный с положительным наблюдением, и $V=(v_1, v_2,\ldots, v_t)$ – с отрицательным наблюдением. Вводится переменная:

$$w_i(U,V)=\begin{cases}1, u_i \neq v_i,\\ 0, u_i = v_i.\end{cases}$$

Условие раздельности множеств Ω_s^+ и Ω_s^- эквивалентно требованию выполнения неравенства $\sum w_i(U,V)y_i \geq 1$ для любых $U \in \Omega_S^+$ и $V \in \Omega_S^-$.

Для того чтобы выборка данных была более устойчива к ошибкам измерений, условие следует усилить, заменив число 1 в правой части неравенства на некоторое целое число d. Это означает, что положительное и отрицательное наблюдения должны отличаться не менее чем d признаками.

Таким образом, задача минимизации опорного множества может быть сформулирована как задача условной псевдобулевой оптимизации:

$$\sum_{j=1}^{t} y_j \to \min,$$

$$\sum_{i=1}^{t} w_i(U,V)y_i \geq d \text{ для любых } U \in \Omega_S^+ \text{ и } V \in \Omega_S^-,$$

где $y \in \{0,1\}^t$.

Следующим этапом метода является формирование закономерностей.

1) *Модель формирования закономерностей с максимальным покрытием наблюдений обучающей выборки.*

Положительной закономерностью называется подкуб пространства булевых переменных B_2^t, который пересекается с множеством Ω_s^+ и не имеет общих элементов с множеством Ω_s^-. Отрицательная закономерность задается аналогично. Положительная ω-закономерность для $\omega \in \{0,1\}^t$ – это закономерность, содержащая в себе точку ω. Для каждой точки $\omega \in \Omega_s^+$ найдем максимальную ω-закономерность, то есть покрывающую наибольшее число точек Ω_s^+.

Соответствующий подкуб задается с помощью переменных y_j:

$$y_j = \begin{cases} 1, \text{ если } i\text{-ый признак зафиксирован в подкубе,} \\ 0, \text{ в противном случае.} \end{cases}$$

Условие, говорящее о том, что положительная закономерность не должна содержать ни одной точки Ω_s^-, требует, чтобы для каждого наблюдения $\rho \in \Omega_s^-$ переменная y_j принимала значение 1 по меньшей мере для одного j, для которых $\rho_j \neq \omega_j$:

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^{t} y_j \geq 1 \text{ для любого } \rho \in \Omega_s^-.$$

Усиление ограничения для повышения устойчивости к ошибкам производится путем замены числа 1 в правой части неравенства на целое положительное число d.

С другой стороны, позитивное наблюдение $\sigma \in \Omega_s^+$ будет тогда входить в рассматриваемый подкуб, когда переменная y_j принимает значение 0 для всех индексов j, для которых $\sigma_j \neq \omega_j$. Число положительных наблюдений, покрываемых ω-закономерностью, может быть вычислено как:

$$\sum_{\sigma\in\Omega_s^+}\prod_{\substack{j=1\\ \sigma_j\neq\omega_j}}^{t}(1-y_j).$$

Таким образом, для формирования закономерностей получается задача условной псевдобулевой оптимизации с алгоритмически заданными функциями:

$$\sum_{\sigma\in\Omega_S^+}\prod_{\substack{j=1\\ \sigma_j\neq\omega_j}}^{t}(1-y_j)\to\max \tag{1}$$

$$\sum_{\substack{j=1\\ \rho_j\neq\omega_j}}^{t}y_j\geq d \text{ для любого } \rho\in\Omega_s^-,\ y\in\{0,1\}^t. \tag{2}$$

Целевая функция (1) и функция ограничения (2) в этой задаче являются унимодальными монотонными псевдобулевыми функциями.

Аналогично формулируется задача нахождения максимальных отрицательных закономерностей.

Каждая закономерность характеризуется двумя показателями: покрытием, т.е. числом захватываемых наблюдений своего класса, и степенью, т.е. числом участвующих переменных при ее формировании.

Специфика задач классификации, встречающихся на практике, состоит в том, что база данных имеет большое число неизмеренных значений (пропущенных данных), а сделанные измерения могут быть неточны либо ошибочны. Такие данные не позволяют построить классификатор с «хорошо интерпретируемыми» правилами и высокой точностью решения задач. Для повышения устойчивости метода к выбросам следует ослабить ограничение (2) – сделать возможным, чтобы закономерность захватывала некоторое малое число наблюдений другого класса. В этом случае степень вычисляемых закономерностей уменьшится, а покрытие увеличится.

Ограничение оптимизационной модели будет выглядеть следующим образом:

$$\sum_{\rho\in\Omega_S^-}z_\rho\leq D,\ \text{где } z_\rho=\begin{cases}0, \text{если } \sum\limits_{\substack{j=1\\ \rho_j\neq\omega_j}}^{t}y_j\geq d,\\ 1, \text{в противном случае;}\end{cases} \tag{3}$$

D – число наблюдений другого класса, которым допускается быть покрытыми закономерностью (целое неотрицательное число).

Функции (1)–(3) построенной модели оптимизации задаются алгоритмически. Для решения задачи оптимизации используются алгоритмы оптимизации, основанные на поиске граничных точек допустимой области. Эти алгоритмы разработаны специально для этого класса задач и основаны на поведении монотонных функций модели оптимизации в пространстве булевых переменных.

Согласно модели (1,3) наиболее предпочтительными являются закономерности с наибольшим покрытием. Следствием этого является то, что формируемые закономерности имеют маленькую степень, т.е. состоят из небольшого числа термов и используют лишь малую часть признаков. Закономерности с маленькой степенью соответствуют большим областям в пространстве признаков. Это приводит к возможному покрытию наблюдений другого класса (отсутствующих в обучающей выборке) и повышению количества неверно классифицированных наблюдений. Данная особенность влияет на информативность закономерности, уменьшая ее. Поэтому с целью повышения информативности предлагается алгоритмическая процедура наращивания закономерностей. Она применяется к каждой построенной закономерности и заключается в максимальном увеличении степени данных закономерностей при условии сохранения покрытия:

$$\sum_{j=1}^{t} y_j \to \max$$
$$fc(Y) = fc'(Y),$$

где $fc(Y)$ – значение целевой функции (покрытие) для закономерности до процедуры наращивания, $fc'(Y)$ – значение целевой функции для закономерности после процедуры наращивания.

2) *Модель формирования закономерностей с покрытием существенно различных подмножеств наблюдений обучающей выборки.*

Предлагается подход для задания целевой функции модели оптимизации при построении закономерностей, базирующийся на модификации целевой функции (1) для увеличения различности правил в классификаторе.

Согласно целевой функции (1) каждая формируемая закономерность максимизирует свое покрытие, захватывая наблюдения, которые являются типичными представителями класса, а нетипичные наблюдения класса остаются непокрытыми и в классификаторе отсутствуют закономерности, учитывающие их. Поэтому мы получаем набор сходных закономерностей для класса, тем самым, снижая качество классификации. Для получения классификатора с более высокой различностью правил, которая позволяет выделять существенно различные подмножества наблюдений, предлагается модифицировать целевую функцию (1) для нахождения положительных закономерностей следующим образом:

$$\sum_{\sigma \in \Omega_S^+} K_\sigma \cdot \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^{t} (1 - y_j) \to \max, \qquad (4)$$

где K_σ – вес позитивного наблюдения $\sigma \in \Omega_s^+$, уменьшаемый при покрытии данного наблюдения, понижая свой приоритет участия в формировании следующей закономерности в пользу непокрытых наблюдений.

Аналогично формируется целевая функция модели оптимизации для нахождения отрицательных закономерностей.

При использовании модели оптимизации с целевой функцией (4) для формирования закономерностей необходимо задать начальные веса для всех наблюдений и правило изменения весов для наблюдений, которые приняли участие при формировании текущей закономерности. Начальные веса предлагается выбрать равными 1 для каждого наблюдения в обучающей выборке. Правило изменения веса для наблюдения, которое приняло участие при формировании текущей закономерности:

$$K_{i+1} = \max\left[0, K_i - \frac{1}{N_{\max}}\right],$$

где K_i, K_{i+1} – веса покрываемого наблюдения при формировании текущей и следующей закономерностей, $N_{\max}$ – параметр, задаваемый исследователем, означающий максимальное количество закономерностей, покрывающих наблюдение обучающей выборки в классификаторе.

Таким образом, используя оптимизационную модель с целевой функцией (4) и ограничением (3) для построения закономерностей, получаются различные правила, из которых в дальнейшем формируется классификатор.

Когда все закономерности сформированы, переходим к следующему этапу метода – построение классификатора. Результатом предыдущего этапа является семейство закономерностей, число которых ограничено мощностью выборки данных $|\Omega^+ \cup \Omega^-|$. Классификатор состоит из полного набора положительных и отрицательных закономерностей.

Чтобы классифицировать новое наблюдение, воспользуемся следующим решающим правилом:

1) Если наблюдение удовлетворяет условиям одной или нескольких положительных закономерностей и не удовлетворяет условиям ни одной из отрицательных, то оно классифицируется как положительное.

2) Если наблюдение удовлетворяет условиям одной или нескольких отрицательных закономерностей и не удовлетворяет условиям ни одной из положительных, то оно классифицируется как отрицательное.

3) Если наблюдение удовлетворяет условиям p' из p положительных закономерностей и q' из q отрицательных, то знак наблюдения определяется как $p'/p - q'/q$.

4) В случае, если наблюдение не удовлетворяет условиям ни одной закономерности, положительной или отрицательной, то оно относится к классу, имеющему наименьшую цену ошибки.

Ввиду того, что объем выборки данных может быть значителен, встает вопрос о сокращении числа закономерностей, так как это число равно в исходном классификаторе мощности обучающей выборки данных $|\Omega^+ \cup \Omega^-|$. Иными словами, необходимо определить классификатор, состоящий из некоторого числа закономерностей, таким образом, чтобы он был способен классифицировать те же наблюдения, которые можно классифицировать с помощью полной системы закономерностей. С этой

целью в диссертации предлагаются две алгоритмические процедуры сокращения количества закономерностей в исходном классификаторе.

Чтобы реализовать алгоритмическую процедуру выбора базовых наблюдений для формирования закономерностей, необходимо выполнить ряд последовательных действий. Во-первых, на основе наблюдений обучающей выборки получить центроиды для каждого класса, используя алгоритм «k-средних». Во-вторых, добавить полученные наборы центроидов к наблюдениям обучающей выборки. В-третьих, использовать центроиды в качестве базовых наблюдений для формирования закономерностей.

Предлагается другой подход для сокращения количества закономерностей в исходном классификаторе. Необходимо построить классификатор, количество закономерностей которого равно мощности обучающей выборки данных, и сократить данное количество правил при сохранении высокой точности классификации. Для реализации такого подхода предлагается процедура построения классификатора как композиции информативных закономерностей.

В работе предлагается использовать критерий бустинга для измерения информативности закономерности, т.к. он адекватно оценивает информативность закономерности и прост для вычисления:

$$H(p,n)=\sqrt{p}-\sqrt{n}, \qquad (5)$$

где p – количество наблюдений своего класса, которые захватывает построенная закономерность; n – количество наблюдений другого класса, которые захватывает построенная закономерность.

Предлагается формировать классификатор только из информативных закономерностей, т.е. информативность которых выше некоторого порога информативности (H_0), задаваемого исследователем. В результате это приведет к сокращению числа закономерностей в классификаторе без потери его точности или при незначительном ее изменении в положительную или отрицательную сторону.

При решении данной задачи возникает проблема выбора порога информативности. Для решения данной проблемы в работе разработана следующая итеративная процедура. На первом шаге порог информативности выбрать равным нулю для положительного и для отрицательного набора закономерностей, тем самым получается исходный классификатор, состоящий из максимального числа закономерностей. На следующем шаге процедуры предлагается выбрать порог информативности для отрицательных (положительных) закономерностей равный значению средней информативности ($H_{ср}$) по всем отрицательным (положительным) закономерностям:

$$H_{ср}=\frac{1}{\mathrm{q}}\cdot\sum_{i=1}^{q}H_i,$$

где q – количество отрицательных (положительных) закономерностей в классификаторе, H_i – информативность отрицательной (положительной) i-й закономерности, рассчитанная по формуле (5).

Для получения нового классификатора, состоящего из более информативных закономерностей, удаляем из исходного классификатора все отрицательные (положительные) закономерности, значения информативности которых ниже найденного значения порога информативности для них. Рассчитав значения средней информативности для отрицательных и положительных закономерностей текущего классификатора, будем их использовать для построения последующего классификатора, состоящего из закономерностей, информативность которых превышает значения средней информативности текущего классификатора. Таким образом, строим каждый последующий классификатор, используя значения средней информативности текущего. При этом количество закономерностей сокращается, а значения средней информативности возрастают для каждого последующего классификатора. Условием остановки следует считать момент увеличения количества неклассифицированных (непокрытых) наблюдений при классификации, т.е. закономерности, входящие в текущий классификатор, не покрывают некоторые наблюдения, входящие в экзаменующую выборку. Поэтому необходимо вернуться либо к предыдущему классификатору, поменяв значение двух порогов информативности на предыдущие их значения, либо поменять значение только одного порога информативности по отрицательным (положительным) закономерностям.

На основе разработанных алгоритмических процедур предлагаются модификации для метода логического анализа данных с целью усиления обобщающих способностей классификатора и повышения его интерпретируемости за счет сокращения числа правил, используемых в нем:

– применение целевой функции (4) и ограничения (3) для формирования закономерностей и построение классификатора только из тех правил, значение целевой функции для которых больше нуля;

– использование алгоритмической процедуры выбора базовых наблюдений для формирования закономерностей и применение к полученным правилам процедуры наращивания;

– применение алгоритмической процедуры построения классификатора как композиции информативных закономерностей на базе оптимизационной модели (1, 3) с процедурой наращивания.

Предлагаемые модификации для метода логического анализа данных позволяют повысить качество классификации новых наблюдений.

Третья глава посвящена программной реализации метода логического анализа данных и экспериментальным исследованиям на практических задачах.

Метод логического анализа данных реализован в виде программной системы, с помощью которой решены следующие задачи классификации:

выявление спама, классификация результатов радарного сканирования ионосферы, прогнозирование осложнений инфаркта миокарда.

Для нахождения правил в каждой задаче использовались четыре оптимизационные модели: «жесткая» модель, не допускающая, чтобы построенные правила покрывали наблюдения другого класса; модифицированная модель, позволяющая, чтобы правила покрывали некоторое ограниченное число наблюдений другого класса; модифицированная модель с процедурой наращивания закономерностей; модель для формирования закономерностей с покрытием существенно различных подмножеств наблюдений обучающей выборки.

В таблице 1 приведены результаты классификации для одной из решаемых задач – выявление спама. В испытании участвовало 279 отрицательных (не спам) и 181 положительных наблюдений (спам), 20% выборки используется для тестирования. Проведено 20 экспериментов, результаты экспериментов усреднены.

Таблица 1 – Результаты классификации для задачи выявления спама

Задача оптимизации	Мн-во правил	Кол-во правил	Покрытие отрицательных наблюдений	Покрытие положительных наблюдений	Степень правила	Точность классификации, %
Целевая функция (1), ограничение (2)	отр.	234	49	0	4	98
	пол.	134	0	29	4	68
Целевая функция (1), ограничение (3)	отр.	234	96	5	5	98
	пол.	134	5	50	4	79
Целевая функция (1), ограничение (3) с процедурой наращивания	отр.	234	96	4	7	98
	пол.	134	4	50	5	87
Целевая функция (4), ограничение (3)	отр.	49	69	5	4	96
	пол.	59	5	31	4	72

В результате применения процедуры наращивания закономерностей получаются закономерности с максимальным покрытием и с более высокой степенью, повышая надежность принимаемых классификатором решений. Модификация для метода логического анализа данных, связанная с применением целевой функции (4), позволяет упростить классификатор, значительно сокращая количество закономерностей в нем.

Выполняется проверка процедуры выбора базовых наблюдений для формирования закономерностей. Для задачи классификации результатов радарного сканирования ионосферы генерируются по 15 центроидов для каждого класса, используя алгоритм «k-средних» в программе WEKA. Добавляются в исходную обучающую выборку сгенерированные центроиды, строятся на их базе закономерности. В итоге, в данной задаче для тестирования используется 20% выборки, состоящей из 240

положительных и 141 отрицательных наблюдений. Результаты классификации приведены в таблице 2.

Таблица 2 – Точность для задачи классификации результатов радарного сканирования ионосферы

Множество правил	Покрытие отр. наблюдений в новом / исходном классификаторах	Покрытие пол. наблюдений в новом / исходном классификаторах	Степень правила в новом / исходном классификаторах	Кол-во правил в новом / исходном классификаторах	Точность нового классификатора, %	Точность исходного классификатора, %
отр.	45 / 36	15 / 15	2 / 2	15 / 95	74	68
пол.	15 / 15	139 / 130	3 / 3	15 / 186	96	98

Согласно результатам (таблица 2), получили небольшое изменение точности классификации для решаемой задачи и сокращение количества правил классификатора в 9 раз.

Выполняется проверка алгоритмической процедуры построения классификатора как композиции информативных закономерностей на задаче выявления спама. Для тестирования используется 20% выборки. Результаты классификации приведены в таблице 3. В каждом опыте, представленном в таблице 3, исследователь задает только порог информативности. В первом опыте значения порога информативности равны нулю для каждого класса. В следующих опытах – значениям средней информативности, полученным в предыдущем опыте. После появления непокрытых наблюдений варьируются значения порога информативности только для одного класса.

Таблица 3 – Результаты классификации для задачи выявления спама при изменении порога информативности, H_0

Номер опыта	Множество правил	Количество правил	Средняя информативность, Hcp	Порог информативности, H_0	Покрытие отр. наблюдений	Покрытие пол. наблюдений	Количество непокрытых наблюдений	Точность классификации, %
1	отр.	234	7,84	0	120	10	0	96
	пол.	134	4,49	0	10	57		89
2	отр.	132	8,51	7,84	134	10	0	93
	пол.	79	5,49	4,49	10	70		85
3	отр.	68	8,85	8,51	141	10	1	87
	пол.	39	6,05	5,49	10	77		79
4	отр.	68	8,85	8,51	141	10	0	98
	пол.	79	5,49	4,49	10	70		87
5	отр.	34	9,03	8,85	146	10	0	96
	пол.	79	5,49	4,49	10	70		89

Согласно полученным результатам (таблица 3) можно отметить, что модификация метода, связанная с данной процедурой, позволяет упростить классификатор, поскольку количество правил, которые его составляют, сокращается в 4 раза относительно полного набора правил для данной задачи. При этом точность классификации либо не уменьшается, либо уменьшается незначительно.

В таблице 4 приведено сравнение результатов классификации по точности для 6 алгоритмов, полученных в системе анализа данных WEKA, с результатами разработанного метода логического анализа данных (LAD). Выборки для каждой задачи разделены случайным образом на обучающую (80%) и экзаменующую (20%). Проведено по 20 экспериментов для каждого метода, результаты экспериментов усреднены.

Таблица 4 – Сравнение алгоритмов классификации

Задача	Алгоритм / Показатель	1-R	RIPPER	CART	C4.5	Random Forest	Adaboost	LAD
Выявление спама	Количество верно классифицированных наблюдений, %	82,6	91,3	90,2	90,2	89,1	91,3	92,4
Радарное сканирование ионосферы	Количество верно классифицированных наблюдений, %	78,6	82,8	82,8	81,4	84,2	88,5	90

Согласно данным, приведенным в таблице 4, метод логического анализа данных по точности решения задач превосходит сравниваемые с ним алгоритмы классификации. Кроме того, преимуществом метода является возможность соблюдать баланс между различными критериями сравнения алгоритмов классификации путем целенаправленной настройки параметров метода.

В **заключении** диссертации приведены основные результаты и выводы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В ходе выполнения диссертационной работы получены следующие результаты:

1. Проведен анализ существующих логических алгоритмов классификации, алгоритмов поиска информативных закономерностей для них, и основных программных систем, решающих практические задачи классификации. Отмечено, что для классификации наиболее приемлем алгоритм, основанный на голосовании правил.

2. Разработана алгоритмическая процедура выбора базовых наблюдений для формирования закономерностей, отличающаяся от известных целенаправленным выбором базовых наблюдений, получаемых путем применения алгоритма «k-средних» к множеству наблюдений обучающей выборки.

3. Разработана алгоритмическая процедура наращивания закономерностей, полученных на базе оптимизационной модели с максимальным покрытием наблюдений обучающей выборки.

4. Создана модель оптимизации для формирования закономерностей, отличающаяся от известных наличием в целевой функции весового коэффициента покрываемого наблюдения, а также возможностью захвата наблюдений другого класса.

5. Разработана алгоритмическая процедура построения классификатора как композиции информативных закономерностей, отличающаяся от известных совместным использованием критерия бустинга для оценки информативности закономерностей и новой итеративной процедуры выбора порога информативности.

6. Модифицирован метод логического анализа данных на основе разработанных алгоритмических процедур, при использовании которых повышается интерпретируемость классификатора и улучшаются его обобщающие способности.

7. В результате решения практических задач эмпирически проверена пригодность оптимизационных моделей для формирования информативных закономерностей и эффективность разработанных алгоритмических процедур для метода логического анализа данных.

8. Проведено сравнение по точности метода логического анализа данных с другими алгоритмами классификации на практических задачах. В результате метод показал лучшие результаты по точности решения предложенных задач.

Таким образом, в диссертационной работе разработаны, исследованы и проверены на практических задачах модификации для метода логического анализа данных, основанные на создании оптимизационных моделей для формирования информативных закономерностей и алгоритмических процедур сокращения количества правил в классификаторе при сохранении высокой точности, что является вкладом в теорию и практику интеллектуального анализа данных.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в ведущих рецензируемых научных журналах и изданиях

1. Кузьмич, Р.И. Модификация целевой функции при построении паттернов для увеличения различности правил в модели классификации / Кузьмич Р.И., Масич И.С. // Системы управления и информационные технологии. – 2014. – N2 (56). – С. 14-18.

2. Кузьмич, Р.И. Применение процедуры кластеризации для генерирования объектов с целью сокращения числа паттернов в модели классификации / Кузьмич Р.И., Виноградова А.И. // Вестник КрасГАУ. – 2013. – № 9(84). – С. 51-55.

3. Кузьмич, Р.И. Построение модели классификации как композиции информативных паттернов / Кузьмич Р.И., Масич И.С. //

Системы управления и информационные технологии. – 2012. – N2 (48). – С. 18-22.

4. Кузьмич, Р.И. Сравнительный анализ методов классификации данных на практических задачах прогнозирования и диагностики / Масич И.С., Краева Е.М., Кузьмич Р.И., Гулакова Т.К. // Системы управления и информационные технологии. – 2011. – N1(43). – С. 20-25.

5. Кузьмич, Р.И. Модель логического анализа для решения задачи прогнозирования инфаркта миокарда / С.Е. Головенкин, Т.К. Гулакова, Р.И. Кузьмич, И.С. Масич, В.А. Шульман // Вестник СибГАУ. – 2010. – Вып. 4 (30). – С. 68-73.

Публикации в сборниках трудов конференций

6. Кузьмич, Р.И. Исследование зависимости точности классификации от степени формируемых паттернов в методе логического анализа данных / Р.И. Кузьмич // Материалы XVIII Междунар. науч. конф., посвящ. 90-летию со дня рождения генер. конструктора ракет.-космич. систем акад. М. Ф. Решетнева (11–14 нояб. 2014, г. Красноярск) : в 3 ч. / под общ. ред. Ю. Ю. Логинова ; Сиб. гос. аэрокосмич. ун-т. – Красноярск, 2014. – Ч. 2. - С. 80-82.

7. Кузьмич, Р.И. Применение информативных паттернов для построения модели классификации при решении задач медицинской диагностики / Кузьмич Р.И., Ступина А.А., Масич И.С. // Материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР–2013». – Томск: В-Спектр, 2013: В 5 частях. – Ч. 3. – С. 183-186.

8. Кузьмич, Р.И. Программная реализация логических алгоритмов классификации для прогнозирования осложнений инфаркта миокарда / Ступина А.А., Кузьмич Р.И. // Материалы Всероссийской молодежной научной конференции с международным участием «Современные проблемы фундаментальных и прикладных наук» / ФГБОУ ВПО «Кемеровский технологический институт пищевой промышленности» Кемерово, Кузбассвузиздат; 2011. – С. 84-87.

9. Кузьмич, Р.И. Поисковые алгоритмы псевдобулевой оптимизации в задаче классификации данных / И.С. Масич, Р.И. Кузьмич // Материалы XV международной научной конференции «Решетневские чтения», Сиб. гос. аэрокосмич. ун-т. - Красноярск, 2011. – Ч. 2. - С. 472-473.

10. Кузьмич, Р.И. Генерирование объектов для построения паттернов с целью сокращения модели классификации / Р.И. Кузьмич, И.С. Масич // Материалы XV международной научной конференции «Решетневские чтения», Сиб. гос. аэрокосмич. ун-т. - Красноярск, 2011. – Ч. 2. - С. 462-463.

11. Кузьмич, Р.И. Разработка программной системы на основе логических алгоритмов классификации для решения задач медицинской диагностики и прогнозирования / Ступина А.А., Масич И.С., Кузьмич Р.И., Ступин О.Г. // Сборник научных трудов по материалам XIV

Международной научно-технической конференции «Фундаментальные и прикладные проблемы приборостроения и информатики» – М.: МГУПИ, 2011. – С. 112-117.

12.Кузьмич, Р.И. Создание программной системы для решения задачи прогнозирования осложнений инфаркта миокарда / Р.И. Кузьмич // Материалы XLIX Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии / Новосиб. гос. ун-т. Новосибирск, 2011. – С. 115.

13. Кузьмич, Р.И. Сравнение методов классификации данных на практических задачах прогнозирования и диагностики / И.С. Масич, Р.И. Кузьмич // Материалы III Международной молодежной научно-технической конференции «Молодежь, техника, космос» - Санкт-Петербург: БГТУ,2011. – С. 215-217.

14. Кузьмич, Р.И. Определение важности признаков при формировании паттернов в задаче классификации / Р.И. Кузьмич // Материалы XIV международной научной конференции «Решетневские чтения», Сиб. гос. аэрокосмич. ун-т. - Красноярск, 2010. – Ч. 2. - С. 394-395.

Зарегистрированные программные системы

15. Кузьмич, Р.И. Логический анализ данных в задачах классификации / И.С. Масич, Р.И. Кузьмич, Е.М. Краева. – М: Роспатент, 2011. № гос. рег. 2011612265.

Кузьмич Роман Иванович
Модифицированный метод логического анализа
данных для задач классификации

Автореферат

Подписано к печати
Формат 60х84/16. Бумага писчая. Печ. л. 1.0
Тираж 100 экз. Заказ № _____
Отпечатано в отделе копировальной и множительной техники СибГАУ
660014 г. Красноярск, пр. им. газеты «Красноярский рабочий», 31