На правах рукописи

Масич Игорь Сергеевич

МЕТОД ОПТИМАЛЬНЫХ ЛОГИЧЕСКИХ РЕШАЮЩИХ ПРАВИЛ ДЛЯ КЛАССИФИКАЦИИ ОБЪЕКТОВ

05.13.01 – системный анализ, управление и обработка информации (космические и информационные технологии)

Автореферат диссертации на соискание ученой степени доктора технических наук

Красноярск – 2019

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнева», г. Красноярск

Научные консультанты: доктор технических наук, профессор
Антамошкин Александр Николаевич
доктор технических наук, доцент
Казаковцев Лев Александрович

Официальные оппоненты:
Дмитриев Михаил Геннадьевич, доктор физико-математических наук, профессор, Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, главный научный сотрудник
Дулесов Александр Сергеевич, доктор технических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Хакасский государственный университет им. Н. Ф. Катанова», кафедра информационных технологий и систем, профессор
Крутиков Владимир Николаевич, доктор технических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Кемеровский государственный университет», кафедра прикладной математики, профессор

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный технический университет»

Защита состоится 25 октября 2019 г. в 12:00 на заседании диссертационного совета Д 212.249.05, созданного на базе Сибирского государственного университета науки и технологий им. академика М.Ф. Решетнева, по адресу:
660037, г. Красноярск, проспект им. газеты Красноярский рабочий, 31.
С диссертацией можно ознакомиться в библиотеке Сибирского государственного университета науки и технологий имени академика М. Ф. Решетнева и на сайте http://www.sibsau.ru

Автореферат разослан ___._____.2019 г.

Учёный секретарь
диссертационного совета Панфилов Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. К настоящему времени в области классификации, кластеризации и распознавания сформировался ряд довольно эффективных методов, в том смысле, что они могут решать задачи распознавания с высокой точностью. Но при решении реальных задач распознавания и прогнозирования часто возникают вопросы, связанные с интерпретируемостью получаемых результатов и обоснованностью предлагаемых решений. Это, прежде всего, задачи, в которых могут быть велики негативные последствия принятия неверного решения (прогноза). И системы поддержки принятия решений, используемые для таких задач, должны обеспечивать возможность обосновывать решения и интерпретировать результат.

Такими задачами являются, в частности, задачи медицинской диагностики и прогнозирования. К примеру, при прогнозировании осложнений инфаркта миокарда есть необходимость в логических решающих правилах, имеющих подтверждение на опытных данных и способных обосновать прогнозируемое решение для новых пациентов.

Логические закономерности могут иметь широкое применение в задачах анализа и моделирования, в частности, при использовании лингвистических шаблонов для извлечения информации о событиях из открытых источников (Дмитриев М.Г.).

Требования доказательности и интерпретируемости предъявляются также к средствам, предназначенным для поддержки принятия решений при проведении испытаний, классификации и отборе электрорадиоизделий (ЭРИ) для комплектации радиоэлектронной аппаратуры космического применения. Одной из ключевых задач, возникающих при отборе изделий микроэлектроники, является задача классификации ЭРИ по однородным партиям (Орлов В.И., Федосов В.В. и др.).

Задачу выявления однородных партий можно рассматривать как задачу автоматической группировки объектов или задачу размещения (Казаковцев Л.А.). Решение такого типа задач исследовано в работах Береснева В.Л., Гимади Э.Х., Колоколова А.А., Кельманова А.В., Кочетова Ю.А., Забудского Г.Г. и др. Применение алгоритмов автоматической группировки и размещения на практических задачах рассмотрено в работах Дулесова А.С., Прутовых М.А. и др.

Для включения этапа классификации ЭРИ в программу испытаний требуется способ проверки соответствия изделия классу с помощью логических решающих правил, опирающихся на четкие значения допустимых интервалов признаков. В условиях космического производства от метода классификации требуется не только точность решения задачи, но одновременно с этим доказательность применяемого подхода и интерпретируемость формулируемого им решения, что в данном случае означает возможность разработки ужесточенных норм параметров изделий.

Интерпретируемость означает возможность записать правило классификации в виде инструкции, понятной человеку. Под доказательностью понимается наличие веских объективных аргументов, подтверждающих предлагаемое системой решение. Более формально, E. Boros, Y. Crama, P.L. Hammer и др. называют

классификатор доказательным по отношению к некоторой обучающей выборке, если каждое правило подтверждено (не менее чем одним) наблюдением и не противоречит ни одному наблюдению, и, кроме того, никакое правило не может быть заменено другим обоснованным правилом, которое имеет большую поддержку в данной выборке.

Направленность на доказательность смещает акцент в развитии подходов к обучению: основная цель уже не в том, чтобы получить высокое отношение правильно распознанных наблюдений, а в том, чтобы предоставить убедительные обоснования для каждой отдельной классификации. Другими словами, нас интересует априорное обоснование правил, а не их апостериорная эффективность. Один из подходов к априорному обоснованию состоит в применении алгоритмов определения информационной энтропии (Дулесов А.С.).

Наиболее интересными с этой точки зрения являются алгоритмы построения решающих правил с использованием аппарата алгебры логики. Известными подходами к распознаванию такого типа являются алгоритмы КОРА (Бонгард М.М., Вайнцвайг М.Н.), ТЭМП (Лбов Г.С.), предназначенные для определения логических закономерностей в виде конъюнкций значений признаков, метод «тупиковых тестов» (Чегис И.А., Яблонский С.В.), а также алгоритмы вычисления оценок, совмещающие метрические и логические принципы классификации (Журавлёв Ю.И.). Решающее правило в этих случаях задаётся в виде алгоритмической процедуры – для распознавания наблюдений используется голосование по конъюнкциям или по «тупиковым тестам». Для выявления закономерностей алгоритм КОРА использует перебор всевозможных конъюнкций с числом литералов (степенью) не более некоторого заданного числа. Эти подходы получили развитие в работах Загоруйко Н.Г., Журавлева Ю.И., Дюковой Е.В., Рудакова К.В., Воронцова К.В.

Рабочей альтернативой являются алгоритмы стратегии «отделяй и властвуй» («separate-and-conquer» или по-другому «стратегии покрытия»), которая берёт начало в семействе алгоритмов AQ (Algorithm for synthesis of quasi-minimal covers, Michalski) и заключается в последовательном исключении покрытых найденной закономерностью обучающих наблюдений и поиске следующей закономерности на оставшихся наблюдениях. Основной упор в этих алгоритмах делается на быстрое построение классификатора (с помощью, как правило, жадной эвристической процедуры), но не на выявление множества наиболее информативных закономерностей, так как все закономерности, за исключением первой, строятся лишь на оставшейся части обучающих наблюдений.

Одним из наиболее основательных подходов к выявлению и использованию логических закономерностей представляет логический анализ данных (Hammer P.L.). Первоначально в данном подходе для выявления закономерностей использовались две техники перебора: снизу вверх (поиск допустимых закономерностей путём добавления литералов) и сверху вниз (исключение литералов для повышения покрытия), которые дополняли друг друга, что является очень трудоёмкой процедурой. В дальнейшем задача выявления закономерностей была сформулирована как задача условной псевдобулевой оптимизации: поиск

закономерности с наибольшим покрытием при условии недопустимости (или ограниченной допустимости) покрытия обучающих наблюдений другого класса. Для решения этой задачи использовался жадный алгоритм (для нахождения приближенного решения) либо линейная аппроксимация целевой псевдобулевой функции со сведением к задаче целочисленного линейного программирования, решение которой является приближенным решением исходной задачи.

В данном исследовании показано, что такой подход не обеспечивает нахождения оптимальных закономерностей, а именно закономерностей, удовлетворяющих критерию доказательности («сильных» закономерностей), которые являются наиболее привлекательными для повышения точности и интерпретируемости результата распознавания. Но, кроме этого, к получаемым закономерностям могут предъявляться дополнительные требования – удовлетворение условиям простоты или избирательности – в результате получаемые закономерности будут обладать своими особенностями. Этим обусловлена необходимость усовершенствования методов решения задачи выявления и использования логических решающих правил для распознавания.

Идея настоящей диссертации состоит в разработке метода поиска оптимальных закономерностей различных типов – сильных первичных и сильных охватывающих закономерностей – и их совместном использовании для построения логических решающих правил.

Степень проработанности проблемы. Методология анализа данных, состоящая в комбинации идей из оптимизации, комбинаторики и булевых функций, была впервые предложена Питером Хаммером (P. Hammer) в 1986 г. и названа логическим анализом данных. Затем этот подход был развит в работах Crama Y., Ibaraki T., Bores E., Kogan A., Alexe G., Alexe S., Bonates T., Anthony M. и др. Обзор результатов приведен в работе Чикалова И.В. и др. (2013). Подход получил широкое практическое применение, описанное в работах Caserta, Reiners (2016), Lejeune, Lozin, Lozina и др. (2019), Bruni, Bianchi, Dolente, Leporelli (2019), Bain, Avila-Herrera, Subasi (2018), Kim, Choi (2015), Yan, Ryoo (2017), Shaban, Meshreki, Yacout и др. (2017), Ragab, Koujok, Ghezzaz и др. (2019). Тем не менее, оставались открытыми ключевые вопросы, касающиеся оптимального назначения порогов при бинаризации количественных признаков, поиска оптимальных закономерностей различных типов, выбора типа закономерностей для решения практических задач, анализа большого количества данных, повышения компактности классификатора.

Объектом диссертационного исследования являются задачи классификации, результат решения которых требует обоснования и доказательности, **предмет исследования** – методы и алгоритмы их решения.

Цель диссертационного исследования состоит в повышении точности решения задач классификации с требованием обоснования результатов распознавания и интерпретации в виде логических правил.

Поставленная цель достигается путем решения следующих **задач**:

- сравнительный анализ известных алгоритмов выявления закономерностей в данных и их использования для решения задач классификации;

- разработка новых способов выявления закономерностей на основе моделей оптимизации, построение моделей для нахождения закономерностей, удовлетворяющих различным критериям (простоты, доказательности, избирательности);
- сравнительный анализ применения разных типов закономерностей и разработка способов совместного использования закономерностей различных типов для повышения качества распознавания;
- построение единой модели оптимизации на основе метода логического анализа данных для поиска пары закономерностей (первичной и охватывающей);
- сведение задачи выявления закономерностей, оптимальных по критериям простоты, доказательности и избирательности, к задачам комбинаторной оптимизации и разработка алгоритмов их решения;
- разработка алгоритма поиска пары закономерностей (сильной первичной и сильной охватывающей) с использованием алгоритмов оптимизации;
- анализ существующих способов бинаризации вещественных признаков для применения метода оптимальных логических решающих правил и разработка способа оптимального назначения порогов вещественных признаков для их дискретизации, при котором наблюдения разных классов оказываются наиболее различимы, а число самих порогов ограничено;
- разработка процедуры ускорения поиска закономерностей, позволяющей применять метод оптимальных логических решающих правил для случаев большого объема данных;
- разработка способа отбора закономерностей в методе оптимальных логических решающих правил для повышения интерпретируемости классификатора без потерь в точности распознавания;
- апробация метода оптимальных логических решающих правил на решении практических задач из разных областей.

Методы исследования. Основные теоретические и прикладные результаты получены с применением методов системного анализа, исследования операций, теории оптимизации, теории множеств.

Новые научные результаты и положения, выносимые на защиту:

1. Впервые предложен метод нахождения сильных первичных и сильных охватывающих закономерностей на основе решения задачи комбинаторной оптимизации – метод оптимальных логических решающих правил. Метод состоит в использовании новых моделей оптимизации и алгоритмов поиска оптимальных решений, основанных на использовании свойств классов задач псевдобулевой оптимизации и принципах схемы ветвей и границ. Разработанный метод позволяет строить классификаторы, обеспечивающие высокую точность (в сравнении с известными классификаторами, основанными на правилах) и при этом способные интерпретировать и обосновывать результаты распознавания.
2. Разработан новый алгоритм условной оптимизации монотонных псевдобулевых функций на основе схемы метода ветвей и границ и поиска среди граничных точек допустимой области. Показано, что применение предлагаемого алгоритма к решению задачи поиска оптимальных закономерностей обеспечивает

нахождение сильных первичных закономерностей за время, допускающее решение задач в интерактивном режиме (в то время как известные алгоритмы находят лишь первичные закономерности), тем самым повышая покрытия отдельных закономерностей и точность классификатора в целом.

3. Построена новая оптимизационная модель для поиска закономерностей, максимальных по отношениям доказательности и избирательности. Показано, что для выявления сильной охватывающей закономерности на основе предлагаемой модели достаточно нахождение любой крайней точки допустимой области.

4. Впервые предложен подход к решению задачи распознавания, заключающийся в использовании одновременно различных видов закономерностей – сильных первичных и сильных охватывающих. Предлагаемый подход приводит к снижению числа нераспознанных наблюдений (по сравнению с использованием одних охватывающих закономерностей), к повышению точности распознавания (по сравнению с использованием одних первичных закономерностей) и даёт возможность более детально оценить уровень надёжности результата распознавания за счёт уточнённой схемы принятия решения, использующей информацию о числе и виде закономерностей, покрывающих наблюдение.

5. Впервые разработан алгоритм поиска пары закономерностей (сильной первичной и сильной охватывающей) с использованием новой модели оптимизации и нового алгоритма псевдобулевой оптимизации. Предлагаемый алгоритм обеспечивает нахождение пар сильных закономерностей с одинаковым покрытием обучающей выборки, имеющих наименьшее (первичная закономерность) и наибольшее (охватывающая) число условий.

6. Разработана новая модель оптимизации для назначения порогов при бинаризации вещественных признаков. Предлагаемая модель обеспечивает нахождение оптимального размещения порогов вещественных признаков, при котором наблюдения разных классов являются наиболее различимыми при заданном максимальном числе порогов, что приводит к повышению покрытий отдельных закономерностей и повышению точности классификатора.

7. Впервые предложена комплексная процедура ускорения поиска закономерностей, состоящая в выборе базовых наблюдений для формирования закономерностей, отборе признаков путём решения задачи псевдобулевой оптимизации и применении приближенного варианта нового алгоритма псевдобулевой оптимизации для выявления закономерностей. Предлагаемая процедура делает возможным применение метода оптимальных логических решающих правил для случаев большого объёма данных без существенного снижения точности классификатора.

8. Предложен новый способ повышения интерпретируемости классификатора, основанного на правилах, состоящий в применении нового алгоритма псевдобулевой оптимизации, новой схемы использования одновременно двух видов закономерностей (сильной первичной и сильной охватывающей) и в отборе достаточного числа закономерностей с ограниченным числом условий на основе решения задачи условной псевдобулевой оптимизации. Предлагаемый способ

позволяет повысить обобщающую способность отдельных закономерностей и уменьшить общее число закономерностей, необходимых для распознавания, что позволяет получать более компактные решающие правила.

9. Предложен новый метод классификации электрорадиоизделий космического применения по однородным партиям, состоящий в формировании логических решающих правил, использующих результаты тестовых воздействий. Предложенный метод позволяет эффективно решать задачу формирования однородных партий при проведении отбраковочных испытаний ЭРИ.

Значение для теории: Семейство алгоритмов классификации, основанных на правилах, дополнено новым подходом построения логических алгоритмов классификации, имеющим более высокую точность. Повышение точности обеспечивается новыми моделями и алгоритмами оптимизации для поиска логических решающих правил, максимальных по отношениям простоты, доказательности и избирательности, новым способом назначения порогов при бинаризации вещественных признаков, новым способом классификации объектов на основе совместного использования сильных первичных и сильных охватывающих закономерностей. Результаты диссертации уже получили дальнейшее развитие в работах последователей. На основе метода оптимальных логических решающих правил разрабатываются новые алгоритмы и модели, например: Кузьмич и др. (2018).

Практическая ценность: Результаты исследования могут быть востребованы в различных областях при решении задач распознавания и прогнозирования, особенно там, где решение, принимаемое по результатам работы системы распознавания, должно быть обосновано, а цена ошибки может быть велика.

Предлагаемый метод оптимальных логических решающих правил создаёт основу для синтеза новых систем поддержки принятия решений при распознавании и прогнозировании. Наиболее важным преимуществом таких систем является способность интерпретировать получаемые решения и обосновывать даваемые рекомендации. Зачастую наличие таких возможностей является первостепенным для работы пользователя при решении практических задач распознавания и прогнозирования.

Применение предлагаемого метода для решения задачи прогнозирования осложнений инфаркта миокарда дает значимые преимущества (по сравнению, например, с ранее использованным способом решения задачи с помощью искусственных нейронных сетей), заключающиеся в возможности объяснения и обоснования результатов прогнозирования, с точностью распознавания, превосходящей известные логические алгоритмы классификации.

Предлагаемый метод был применен для решения задачи классификации электрорадиоизделий (ЭРИ) по однородным партиям, что позволило получить явные правила классификации, и выполнять классификацию на основе части признаков (результатов тестовых воздействий), которые являются значимыми (значимыми комплексно или комбинаторно, а не только индивидуально).

Практическая реализация результатов. Результаты исследования использованы для решения задачи классификации ЭРИ по производственным

партиям для комплектации радиоэлектронной аппаратуры космических аппаратов. Построенные модели оптимизации для нахождения закономерностей в данных и разработанные алгоритмы псевдобулевой оптимизации легли в основу программы для ЭВМ, позволяющей выявлять закономерности в данных тестовых испытаний ЭРИ и использовать их для классификации ЭРИ. Программа классификации ЭРИ по производственным партиям легла в основу СППР в АО «ИТЦ – НПО ПМ» (г. Железногорск).

Апробация. Основные положения и результаты работы докладывались и прошли всестороннюю апробацию на международных и всероссийских научных и научно-практических конференциях и семинарах. В их числе: VII Международная конференция «Проблемы оптимизации и их приложения» (Optimization Problems and Their Applications, OPTA-2018, г. Омск), Седьмая Международная конференция «Системный анализ и информационные технологии» (САИТ-2017, г. Светлогорск Калининградской области), 30th International Business Information Management Association Conference (IBIMA 2017, Мадрид, Испания), Международная научно-практическая конференция «Решетневские чтения» (2017, 2018 гг., Красноярск-Железногорск), VI Международная конференция «Проблемы оптимизации и экономические приложения» (Омск, 2015), Всероссийская научно-практическая конференция «Информационно-телекоммуникационные системы и технологии» (ИТСиТ-2014, Кемерово) и др.

Основные научные результаты были получены в рамках государственного задания Министерства науки и высшего образования РФ (проект 2.5527.2017/8.9), отдельные практические результаты – в рамках хозяйственного договора между СибГУ им. М.Ф. Решетнёва и АО «ИТЦ-НПО ПМ» (2014-2019 гг.).

Основные результаты исследований были отмечены Правительством и Законодательным собранием Красноярского края Государственной премией Красноярского края в области профессионального образования в 2016 году.

Публикации. По материалам диссертации опубликовано 84 работы, в том числе 25 статей в ведущих российских рецензируемых периодических изданиях, рекомендуемых ВАК РФ для опубликования основных научных результатов диссертационных исследований, 12 статей в изданиях, включенных в международные базы цитирования Web of Science, Scopus и Mathematical Reviews, 5 монографий. Зарегистрировано 6 программ для ЭВМ.

Структура и объем работы. Диссертация изложена на 261 странице, состоит из введения, шести глав, заключения и списка литературы из 250 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** дана общая характеристика проблемы, обоснована актуальность выбранной темы исследования, определены цели и задачи, сформулированы основные положения, выносимые на защиту, научная новизна и практическая значимость полученных результатов.

В **первой главе** проведен анализ современного состояния методов классификации, основанных на правилах.

Наиболее популярными алгоритмами построения логических правил для распознавания являются алгоритмы «отделяй и властвуй» или покрывающие

алгоритмы. Все они имеют один и тот же цикл верхнего уровня: в начале алгоритм «отделяй и властвуй» ищет правило, которое объясняет часть его обучающих наблюдений, «отделяет» эти наблюдения и рекурсивно «властвует» над оставшимися наблюдениями, формируя правила до тех пор, пока непокрытых наблюдений не останется. Это гарантирует, что каждый экземпляр исходного набора обучающей выборки покрывается, по крайней мере, одним правилом (Furnkranz).

Стратегия «отделяй и властвуй» берет свое начало в семействе алгоритмов AQ (Michalski, 1969) под названием «стратегия покрытия». Термин «отделяй и властвуй» был придуман Пагальо и Хаусслером (Pagallo, Haussler, 1990) для описания стратегии обучения: сформировать правило, которое покрывает часть заданного обучающего множества, и рекурсивно формировать другие правила, которые покрывают часть из оставшихся примеров до тех пор, пока непокрытых примеров не останется.

Алгоритмы «отделяй и властвуй» были разработаны для различных задач обучения. Например (по типам формируемых правил):

Множества правил в виде ДНФ: AQ (Michalski, 1969), PRISM (Cendrowska, 1987), PFOIL (Mooney, 1995), GROW (Cohen, 1993), RIPPER (Cohen, 1995), COR-AL (Лбов Г.С., Котюков В.И., Манохин А.Н., 1973), GARIPPER (Yang, Tiyyagura, Chen, Honavar, 1999), RipMC (Asadi, Shahrabi, 2016), DiRUC (Govada, Thomas, Samal, Sahay, 2016).

Множества правил в виде КНФ: PFOIL-CNF (Mooney, 1995), ICL (De Raedt, Van Laer, 1995).

Списки решений: CN2 (Clark, Niblett, 1989; Clark, Boswell, 1991), GREEDY3 (Pagallo, Haussler, 1990), FOIDL (Mooney, Califf, 1995), ДРЭТ (Лбов Г.С., 1981), ACORI (Asadi, Shahrabi, 2016).

Логические программы: INDUCE (Michalski, 1980), FOIL (Quinlan, 1990; Quinlan, Cameron-Jones, 1995), I-REP (Furnkranz, Widmer, 1994), PROGOL (Muggleton, 1995), ProbFOIL (Raedt, Thon, 2011).

Правила регрессии: IBL-SMART (Widmer, 1993), RULE (Weiss, Indurkhya, 1993, 1995), FORS (Karalic, Bratko, 1995), PCR (Zenko, 2005), GuideR (Sikoraa, Wróbel, Gudyś, 2018).

Задача обучения с помощью правил заключается в следующем (Furnkranz). Даны положительные и отрицательные примеры целевого концепта, описанные фиксированным числом признаков. Цель алгоритма - найти описание целевого концепта в форме явных правил, сформулированных в терминах тестов для определенных значений признаков. Результирующий набор правил должен быть способен правильно распознавать экземпляры целевого концепта и отличать их от объектов, которые не относятся к целевому концепту.

Существуют различные подходы к решению этой задачи. Наиболее часто используемой альтернативой является обучение дерева решений через стратегию «разделяй и властвуй» («divide-and-conquer», Quinlan, 1986). Высокая популярность деревьев решений связана с их эффективностью в обучении и классификации (Bostrom, 1995). Кроме того, деревья решений можно легко

превратить в правила, создав одно правило для каждого пути от корня до листа.

Тем не менее, есть несколько аспектов, которые делают формирование правил по стратегии «отделяй и властвуй» порой более привлекательными:

1. Деревья решений часто оказываются довольно большими и трудны для понимания. Quinlan отметил, что даже усеченные деревья решений могут быть слишком громоздкими и сложными для понимания рассматриваемой ситуации, и разработал процедуры упрощения деревьев решений в наборы правил. Дополнительные подтверждения этого исходят от Rivest, который показал, что списки решений (упорядоченные наборы правил) с не более чем k условиями на правило являются более понятными, чем деревья решений с глубиной k. Аналогичные выводы были сделаны у Bostrom.

2. Обучающие алгоритмы в деревьях решений создают неперекрывающиеся правила, что накладывает серьезные ограничения на формируемый набор правил. Одной из проблем, возникающих из этого ограничения, является проблема репликации поддерева (Pagallo, Haussler): наличие одинаковых поддеревьев в разных местах дерева решений, что вызвано ограничением на неперекрываемость правил в стратегии разделения. Алгоритмы типа «отделяй и властвуй» не создают такого ограничения и, следовательно, не восприимчивы к этой проблеме.

Алгоритмы типа «отделяй и властвуй» быстро работают, и в результате получается небольшой упорядоченный список правил, который, к тому же, легко интерпретируется. Но такие алгоритмы имеют значимые изъяны. Во-первых, эти алгоритмы являются жадными алгоритмами локального поиска и поэтому не обеспечивают оптимального решения. Во-вторых, может быть недостаточно информации для принятия решения о классификации, так как решение принимается только на основе выполнения лишь одного правила (или невыполнения всех), а не на основе «голосования» правил. В-третьих, если список правил длинный, то решение трудно интерпретировать, так как нужно учитывать предшествующие правила, которые относятся к различным классам, и объяснение получается запутанным.

Использование в логических процедурах распознавания ряда улучшений, таких как бустинг, голосование правил, выделение признаков и пограничных объектов (Дюкова, Журавлев, Прокофьев, 2017), позволяет существенно повысить эффективность классификатора.

Логический анализ данных (Logical Analysis of Data – LAD) обладает рядом преимуществ перед алгоритмами типа «отделяй и властвуй». Во-первых, все получаемые правила могут быть оптимальными по используемому критерию или отношению (простота, избирательность или доказательность, а также их возможные совмещения). Во-вторых, классификатор не просто разделяет области классов, а строит аппроксимацию областей набором правил. Для оценки силы разделения этих областей может быть использовано понятие «зазора» (Bonates, 2010). В-третьих, голосование правил является средством оценки достоверности принадлежности к каждому классу.

Несмотря на относительно высокую вычислительную сложность логического анализа данных, этот подход представляется перспективным, а разработка и

использование алгоритмов псевдобулевой оптимизации, реализующих свойства решаемого класса задач по формированию решающих правил, делают возможным применение этого подхода за время, допускающее решение задач в интерактивном режиме, при этом позволяя извлечь из данных максимум полезной информации.

Во **второй главе** исследуется подход к выявлению и использованию логических решающих правил, называемый логическим анализом данных.

Предлагаемый метод к поддержке принятия решений при распознавании основан на подходе к выявлению и использованию закономерностей, называемом логическим анализом данных (P. Hammer). Логический анализ данных - это методология анализа данных, которая объединяет идеи и концепции из оптимизации, комбинаторики и булевых функций.

Исследуемый подход к распознаванию, основанный на выявлении в данных правил, изначально разработан для распознавания объектов, признаки которых могут принимать лишь два допустимых значения, то есть являются бинарными. Для задач распознавания, в которых объекты описываются вещественными или какими-либо другими признаками, необходимо применить процедуру бинаризации, которая должна быть неотъемлемым этапом при построении классификатора.

Наибольшую сложность представляет кодирование количественных признаков. Так как число различных значений некоторого количественного признака всех объектов выборки может быть велико, наиболее популярным и обоснованным подходом является дискретизация количественного признака, при которой весь диапазон значений разбивается на интервалы, и каждое значение может относиться только к одному интервалу. Граничные значения этих интервалов называются порогами. Очевидно, что от выбора числа интервалов и расположений порогов зависит работа всего метода распознавания и качество распознавания.

Исходным критерием при дискретизации является качество распознавания объектов, а именно точность и интерпретируемость. Но оценить точность распознавания можно только после прохождения всех этапов построения классификатора, поэтому применить этот критерий при дискретизации не представляется возможным.

Наиболее подходящими критериями, которые можно вычислить непосредственно в процессе бинаризации (а не в результате распознавания контрольных наблюдений), являются следующие.

1. Различимость объектов разных классов. Если два объекта разных классов в пространстве исходных признаков различались друг от друга значениями количественных признаков, то после дискретизации этих признаков может оказаться, что эти объекты совпадут, то есть будут иметь одни и те же значения бинаризованных признаков, что может негативно отразиться на качестве распознавания.

2. Минимум интервалов дискретизации. Чем меньше будет использовано порогов при дискретизации, тем меньше будет число бинарных признаков, что

снизит размерность задачи и, возможно, позволит её решить более успешно. Кроме того, меньшее число интервалов положительно сказывается на интерпретируемости результатов распознавания.

3. Робастность дискретизации. В случае наличия ошибок в измерениях значений признаков может представлять проблему то, что эти значения находятся близко к порогам. Поэтому более робастными являются пороги, расстояния от которых до ближайших значений признака больше.

Известный способ выбора порогов при бинаризации действительных признаков состоит в решении задачи оптимизации (Boros, Hammer, Ibaraki, Kogan, 1997). Модель оптимизации направлена на снижение числа порогов, достаточных для разделения наблюдений различных классов.

$$\sum_{j=1}^{d}\sum_{i=1}^{k_i} y_{ij} \to \min_{Y},$$

$$\sum_{j=1}^{d}\sum_{i=1}^{k_i} a_{ij}^{zw} y_{ij} \geq 1, z \in K^{+}, w \in K^{-},$$

где $y_{ij} \in \{0,1\}$ - управляющая переменная, определяющая использование соответствующего порога, $a_{ij}^{zw} = \left| x_{ij}^{z} - x_{ij}^{w} \right|$ для $z \in K^{+}$ и $w \in K^{-}$.

При практическом использовании приведенной выше модели оптимизации выяснилось, что получаемые с помощью нее решения имеют недостатки. Во-первых, число порогов в найденном решении может быть слишком мало, что уменьшает разделяющую способность классификатора, который строится на основании этих порогов, и негативно влияет на точность распознавания. Для преодоления этого недостатка используется модификация – в правой части ограничений вместо единицы используется целое число $h \geq 1$. Но оказывается, что при назначении в правой части ограничений большего числа (то есть любые два объекта различных классов должны различаться более чем одним порогом) число порогов может быть не уменьшено вообще по отношению к исходному числу потенциальных порогов.

Во-вторых, в этой модели пороги всех признаков смешаны. В полученном решении для некоторого признака может оставаться большое число порогов, в то время как для многих других признаков пороги могут отсутствовать.

Поэтому часто оказывается, что приближенное решение этой задачи может быть предпочтительнее, чем оптимальное, так как это позволяет добиться лучшего компромисса в балансе "вычислительная сложность / разделяющая способность".

В связи с этим возникла необходимость решать новую задачу: выбрать такое расположение приемлемого числа порогов, при котором объекты различных классов разделены как можно сильнее.

Пусть имеются два класса наблюдений K^{+} и K^{-}: $K^{+} \cup K^{-} = K, K^{+} \cap K^{-} = \emptyset$. Наблюдения описываются действительными признаками. Расположим все возможные значения каждого признака j=1,…,d для наблюдений данной выборки в порядке возрастания: $b_1^{(j)} < b_2^{(j)} < b_3^{(j)} < \ldots$.

Потенциальные пороги можно взять как середины между двумя ближайшими значениями признака: $\beta_i^j = \left(b_i^{(j)} + b_{i+1}^{(j)}\right)/2$.

Считается, что в качестве потенциальных порогов имеет смысл брать только те (из выше указанных), для которых ближайшие значения признака с разных сторон от порога соответствуют наблюдениям разных классов, т.е. существуют $z \in K^+$ и $w \in K^-$ (или наоборот) такие, что $z_j = b_i^{(j)}$ и $w_j = b_{i+1}^{(j)}$.

Таким образом, для каждого действительного признака, описывающего объект распознавания, имеем упорядоченное по возрастанию множество потенциальных порогов $\beta_1^j, \beta_2^j, ..., \beta_{k_j}^j$.

Бинарные переменные указывают, превысило ли значение признака объекта соответствующий порог

$$x_{ij} = \begin{cases} 1, & b_j \geq \beta_i^j, \\ 0, & b_j < \beta_i^j. \end{cases}$$

Рассмотрим наблюдения разных классов $z \in K^+$ и $w \in K^-$, для которых определим величины $a_{ij}^{zw} = \left| x_{ij}^z - x_{ij}^w \right|$. Если $a_{ij}^{zw} = 1$, то наблюдения z и w различны по бинарной переменной x_{ij}, то есть значения численного признака b_j у этих наблюдений лежат по разные стороны порога β_i^j.

Введем бинарную переменную y_{ij}, указывающую, будет ли порог β_i^j использоваться при дискретизации

$$y_{ij} = \begin{cases} 1, & \text{если порог } \beta_i^j \text{ используется,} \\ 0, & \text{в противном случае.} \end{cases}$$

Предлагается оптимизационная модель для выбора таких порогов, при которых объекты разных классов оказываются наиболее различимы, а число самих порогов ограничено. В качестве критерия при выборе порогов примем суммарное число признаков, по которым объекты всевозможных пар различаются между собой. При этом любые два объекта должны быть различны хотя бы по одному признаку. Кроме того, следует ограничить число применяемых для каждого численного признака порогов.

Таким образом, получаем следующую задачу оптимизации:

$$\sum_{(z,w)\in(K^+)\times(K^-)} \sum_{j=1}^{d} \left(1 - \prod_{i=1}^{k_j} \left(1 - a_{ij}^{zw} y_{ij}\right) \right) \to \max_Y,$$

$$\sum_{j=1}^{d} \sum_{i=1}^{k_j} a_{ij}^{zw} y_{ij} \geq 1, z \in K^+, w \in K^-,$$

$$\sum_{i=1}^{k_j} y_{ij} \leq h, j = 1,...,d,$$

$$y_{ij} \in \{0,1\}, i = 1,...,k_j, j = 1,...,d,$$

где d – число вещественных признаков,
k_j – число потенциальных порогов для j-го признака,

h – наибольшее заданное число порогов для каждого численного признака.

В **третьей главе** исследуются модели оптимизации для выявления логических закономерностей в данных.

Рассмотрим задачу распознавания объектов, описываемых бинарными признаками и разделенных на два класса $K = K^+ \cup K^- \subset B_2^n$, где $B_2^n = \{0,1\}^n$, $B_2^n = B_2 \times B_2 \times \cdots \times B_2$. При этом классы не пересекаются: $K^+ \cap K^- = \emptyset$.

Наблюдение $X \in K$ описывается бинарным вектором $X = (x_1, x_2, ..., x_n)$ и может быть представлено как точка в гиперкубе пространства бинарных признаков B_2^n. Наблюдения класса K^+ будем называть положительными точками выборки K, а наблюдения класса K^- - отрицательными точками выборки.

Рассмотрим подмножество точек из B_2^n, у которых некоторые переменные фиксированы и одинаковы, а остальные принимают произвольные значения:

$$T = \{x \in B_2^n \,|\, x_i = 1 \text{ для } \forall i \in A \text{ и } x_j = 0 \text{ для } \forall j \in B\},$$

для некоторых подмножеств $A, B \subseteq \{1,2,...,n\}$, $A \cap B = \emptyset$. Это множество может быть также определено в виде булевой функции, принимающей истинное значение для элементов множества: $t(x) = (\underset{i \in A}{\wedge} x_i) \wedge (\underset{j \in B}{\wedge} \bar{x}_j)$.

Множество точек x, для которых $t(x) = 1$, обозначим $S(t)$. Множество $S(t)$ является подкубом в булевом гиперкубе B_2^n, число точек подкуба равно $2^{(n-|A|-|B|)}$.

Бинарная переменная x_i или её отрицание $\bar{x}_i$ в терме называется литералом. Запись x_i^α обозначает x_i, если $\alpha = 1$, и $\bar{x}_i$, если $\alpha = 0$. Таким образом, терм представляет собой конъюнкцию различных литералов, которая не содержит одновременно некоторую переменную и ее отрицание. Множество литералов в терме t обозначим $Lit(t)$.

Будем считать, что терм t покрывает точку $a \in B_2^n$, если $t(a) = 1$, то есть эта точка принадлежит соответствующему подкубу.

Под *закономерностью P* (или *правилом*) в данном случае понимается терм, который покрывает хотя бы одно наблюдение некоторого класса и не покрывает ни одного наблюдения другого класса. То есть закономерность соответствует подкубу, имеющему непустое пересечение с одним из множеств (K^+ или K^-) и пустое пересечение с другим множеством (K^- или K^+ соответственно). Закономерность P, которая не пересекается с K^-, будем называть положительной, а закономерность P', которая не пересекается с K^+ - отрицательной. В дальнейшем для определенности будем рассматривать только положительные закономерности. Множество наблюдений, которые покрываются закономерностью P, обозначим $Cov(P)$. Закономерности являются элементарными блоками для построения логических решающих функций.

Не существует единственного однозначного критерия для сравнения логических закономерностей между собой. При анализе различных данных к качеству и особенностям формируемых закономерностей могут предъявляться разные требо-

вания. В соответствие с работой Хаммера, Когана и др. (2004) для оценки качества чистых (однородных, не покрывающих наблюдений других классов) закономерностей используем три отношения частичного предпорядка — простоты, избирательности и доказательности, а также их возможные совмещения.

Отношение простоты (или компактности) часто используется для сравнения закономерностей между собой, в том числе получаемых разными алгоритмами обучения. Закономерность P_1 предпочтительнее P_2 по отношению *простоты* (обозначим $P_1 \succeq_\sigma P_2$), если $Lit(P_1) \subseteq Lit(P_2)$.

Закономерность P является *первичной*, если после удаления любого литерала из $Lit(P)$ образуется терм, который не является (чистой) закономерностью (то есть покрывает наблюдения другого класса). Очевидно, что оптимальность закономерности по отношению простоты тождественна утверждению, что эта закономерность является первичной.

Поиск более простых закономерностей имеет вполне обоснованные предпосылки. Во-первых, такие закономерности являются лучше интерпретируемыми и понятными для человека при их использовании для принятия решения. Во-вторых, часто считается, что более простые закономерности имеют лучшую обобщающую способность, и их использование приводит к лучшей точности распознавания. Однако, это утверждение спорно, более того, было показано, что уменьшение простоты может приводить к большей точности.

Использование простых, а значит, коротких закономерностей приводит к тому, что уменьшается число неверно распознанных положительных наблюдений (ошибочных отрицательных), но в то же время это может приводить к увеличению числа неверно распознанных отрицательных наблюдений (ошибочных положительных). Естественный путь к уменьшению числа ошибочных положительных наблюдений — это формирование более избирательных наблюдений, что достигается уменьшением размера подкуба, определяющего закономерность.

Закономерность P_1 предпочтительнее P_2 по отношению *избирательности* (обозначим $P_1 \succeq_\Sigma P_2$), если $S(P_1) \subseteq S(P_2)$.

Следует отметить, что два рассмотренных выше отношения противоположны друг другу, то есть $Lit(P_1) \subseteq Lit(P_2) \Leftrightarrow S(P_1) \supseteq S(P_2)$.

Закономерность, максимальная по отношению избирательности, является *минтермом*, то есть закономерностью, покрывающей единственное положительное наблюдение. Использование этого отношения самого по себе, естественно, не эффективно, так как получаемые закономерности (минтермы) не обладают какой-либо обобщающей способностью. Но отношение избирательности является чрезвычайно полезным при использовании совместно с другими отношениями, что будет рассмотрено далее.

Еще одним полезным отношением является отношение, основанное на покрытии, то есть числе положительных наблюдений обучающей выборки, удовлетворяющих условиям закономерности. Несомненно, что закономерности с большим покрытием обладают большей обобщающей способностью. А наблюдения обучающей выборки, покрываемые закономерностью, являются своего рода доказательством применимости этой закономерности для принятия решения.

Закономерность P_1 предпочтительнее P_2 по отношению *доказательности* (обозначим $P_1 \succeq_\varepsilon P_2$), если $Cov(P_1) \supseteq Cov(P_2)$. Закономерности, максимальные по отношению доказательности, называются *сильными*. То есть закономерность P является сильной в том случае, если не существует закономерности P' такой, что $Cov(P') \supset Cov(P)$.

Важно отметить, что рассматриваемые отношения не являются полностью независимыми. Так, отношения простоты и избирательности противоположны друг другу. Более того, можно отметить следующие зависимости:

$$P_1 \succeq_\sigma P_2 \Rightarrow P_1 \succeq_\varepsilon P_2;\ P_1 \succeq_\Sigma P_2 \Rightarrow P_2 \succeq_\varepsilon P_1.$$

Для того чтобы комбинировать рассмотренные отношения между собой, используются два способа их совмещения.

Для заданных критериев π и ρ закономерность P_1 предпочтительнее P_2 по пересечению $\pi \wedge \rho$ (обозначим $P_1 \succeq_{\pi\wedge\rho} P_2$), если и только если $P_1 \succeq_\pi P_2$ и $P_1 \succeq_\rho P_2$.

Для заданных критериев π и ρ закономерность P_1 предпочтительнее P_2 по лексикографическому уточнению $\pi \mid \rho$ (обозначим $P_1 \succeq_{\pi|\rho} P_2$), если и только если $P_1 \succ_\pi P_2$ или ($P_1 \approx_\pi P_2$ и $P_1 \succeq_\rho P_2$).

Из всех возможных комбинаций отношений, внимание заслуживают лишь три из них. Закономерности, максимальные по $\Sigma \wedge \varepsilon$, называются *охватывающими*. Закономерности, максимальные по $\varepsilon \mid \sigma$, называются *сильными первичными*. И закономерности, максимальные по $\varepsilon \mid \Sigma$, называются *сильными охватывающими*.

Из всех типов закономерностей, полученных в соответствие с рассмотренными выше отношениями и их комбинациями, наиболее полезными для выявления информативных закономерностей и их использования для поддержки принятия решений при распознавании представляются закономерности сильные первичные и сильные охватывающие.

Один из путей в составлении набора закономерностей для алгоритма распознавания – это поиск закономерностей, опирающихся на значения признаков конкретных объектов.

Выделим некоторое наблюдение $a \in K^+$. Обозначим P^a закономерность, покрывающую наблюдение a. Те переменные, которые зафиксированы в P^a, равны соответствующим значениям признаков объекта a.

Рассмотрим задачу поиска максимальной закономерности P^a, то есть такого терма, который помимо наблюдения a покрывает как можно больше положительных наблюдений и ни одного отрицательного.

Для задания закономерности P^a введем бинарные переменные $Y = (y_1, y_2, ..., y_n)$

$$y_j = \begin{cases} 1, & i\text{-ый признак фиксирован в } P^a, \\ 0, & \text{в противном случае.} \end{cases}$$

Некоторая точка $b \in K^+$ будет покрываться закономерностью P^a только в том случае, если $y_i = 0$ для всех i, для которых $b_i \neq a_i$. С другой стороны, некоторая

точка $c \in K^-$ не будет покрываться закономерностью P^a в том случае, если $y_i = 1$ хотя бы для одной переменной i, для которой $c_i \neq a_i$.

Таким образом, задачу нахождения максимальной закономерности можно записать в виде задачи поиска таких значений $Y = (y_1, y_2, ..., y_n)$, при которых получаемая закономерность P^a покрывает как можно больше точек $b \in K^+$ и не покрывает ни одной точки $c \in K^-$ (Bonates, Hammer, Kogan, 2007):

$$\sum_{b \in K^+} \prod_{\substack{i=1 \\ b_i \neq a_i}}^{n} (1 - y_i) \to \max, \tag{1}$$

$$\sum_{\substack{i=1 \\ c_i \neq a_i}}^{n} y_i \geq 1 \text{ для всех } c \in K^-. \tag{2}$$

Эта задача является задачей условной псевдобулевой оптимизации, то есть задачей, в которой целевая функция и левые части ограничений являются псевдобулевыми функциями – вещественными функциями булевых переменных. Целевая функция и функции ограничений в этой задаче унимодальны и монотонны.

Для поиска максимальных отрицательных закономерностей задача формулируется подобным образом.

Следует заметить, что любая точка $Y = (y_1, y_2, ..., y_n)$ соответствует подкубу в пространстве булевых признаков $X = (x_1, x_2, ..., x_n)$, включающему в себя базовое наблюдение a. При $Y \in O_k(Y^1)$ (т.е. Y отличается от Y^1 значением k координат), где $Y^1 = (1,1,...,1)$, число точек этого подкуба равно 2^k.

Целевая функция (1) является нелинейной. Bonates, Hammer, Kogan (2007) рассматривают сведение задачи (1)-(2) к задаче целочисленного линейного программирования (ЦЛП). Но в результате сильно возрастает размерность задачи, и они отказываются от практического применения такого подхода и прибегают к эвристическим алгоритмам: жадному алгоритму и линейной аппроксимации целевой функции, позволяющей свести задачу к задаче ЦЛП, решение которой является приближенным решением исходной задачи.

Рассмотрим свойства задачи оптимизации (1)-(2). Приведем основные понятия (Антамошкин, 1989). Точки $X^1, X^2 \in B_2^n$ назовем *k-соседними*, если они отличаются значением k координат, $k = \overline{1,n}$. Множество $O_k(X)$, $k = \overline{0,n}$, всех точек B_2^n, k-соседних к точке X, назовем *k-м уровнем* точки X.

Унимодальную функцию f назовем *монотонной* на B_2^n, если $\forall X^k \in O_k(X^*), k = \overline{1,n}$ выполняется $f(X^{k-1}) \leq f(X^k), \forall X^{k-1} \in O_{k-1}(X^*) \cap O_1(X^k)$, и *строго монотонной*, если это условие выполняется со знаком строгого неравенства.

Рассмотрим основные свойства множества допустимых решений задачи условной псевдобулевой оптимизации. Имеется задача вида

$$C(X) \to \max_{X \in S \subset B_2^n}, \tag{3}$$

где $C(X)$ - монотонно возрастающая от X^0 псевдобулевая функция, $S \subset B_2^n$ - некоторая подобласть пространства булевых переменных, определяемая заданной системой ограничений, например: $A_j(X) \le H_j$, $j = \overline{1,m}$.

Введем ряд понятий для подмножества точек пространства булевых переменных. Точка $Y \in S$ является *граничной точкой* множества S, если существует $X \in O_1(Y)$, для которой $X \notin S$. Точку $Y \in O_i(X^0) \cap S$ будем называть *крайней точкой* множества S с базовой точкой $X^0 \in S$, если $\forall X \in O_1(Y) \bigcap O_{i+1}(X^0)$ выполняется $X \notin S$. Ограничение, определяющее подобласть пространства булевых переменных, будем называть *активным*, если оптимальное решение задачи условной оптимизации (3) не совпадает с оптимальным решением соответствующей задачи оптимизации без учета ограничения.

При этом доказано следующее свойство множества допустимых решений:

Свойство 1. Если целевая функция является строго монотонной унимодальной функцией, а ограничение активно, то оптимальным решением задачи (3) будет точка, принадлежащая подмножеству крайних точек множества допустимых решений S с базовой точкой X^0, в которой целевая функция принимает наименьшее значение:

$$C(X^0) = \min_{X \in B_2^n} C(X).$$

Рассмотрим отдельное ограничение в задаче оптимизации (1)-(2): $A_j(Y) \ge 1$,

где $A_j(Y) = \sum_{\substack{i=1 \\ c_i^j \ne a_i}}^{n} y_i$ для всех $c^j \in K^-$, $j = \{1,2,...,|K^-|\}$.

Введем обозначение $$\delta_i^j = \begin{cases} 1, \text{ если } c_i^j \ne a_i; \\ 0, \text{ если } c_i^j = a_i. \end{cases}$$

Тогда $$A_j(Y) = \sum_{i=1}^{n} \delta_i^j y_i$$

Функция $A_j(Y)$ монотонно убывает от точки $Y^1 = (1,1,...,1)$.

Крайними точками допустимой области являются точки $Y_k \in O_{n-1}(Y^1)$ (либо, что то же самое, $Y_k \in O_1(Y^0)$), причем такие, что Y_k отличаются от $Y^0 = (0,0,...,0)$ значением k-ой компоненты, при которой $\delta_k^j = 1$.

Множеством допустимых решений является объединение подкубов, образуемых крайними точками допустимой области и точкой Y^1:

$$\bigcup_{k:\delta_k^j = 1} K(Y_k, Y^1).$$

Теперь перейдем ко всей системе ограничений: $A_j(Y) \ge 1$, $j = 1,2,...,|K^-|$.

Этой системе будут удовлетворять точки, принадлежащие множеству

$$\bigcap_{j=1}^{|K^-|} \bigcup_{k:\delta_k^j = 1} K(Y_k^j, Y^1),$$

которое, в конечном счете, является объединением конечного числа подкубов. Крайние точки допустимой области могут находиться на совершенно разных уровнях точки Y^1. А их количество, в худшем случае, может достигать значения $C_n^{[n/2]}$, то есть мощности среднего уровня.

Далее рассмотрим целевую функцию $C(Y) = \sum_{b \in K^+} \prod_{\substack{i=1 \\ b_i \neq a_i}}^{n} (1 - y_i)$ для некоторого «базового» наблюдения $a \in K^+$. Либо можно записать

$$C(Y) = \sum_{j=1}^{|K^+|} \prod_{i=1}^{n} (1 - \Delta_i^j y_i), \text{ где } \Delta_i^j = \begin{cases} 1, \text{ если } b_i^j \neq a_i; \\ 0, \text{ если } b_i^j = a_i. \end{cases}$$

Функция $C(X)$ монотонно возрастает от точки $Y^1 = (1,1,...,1)$, принимая в ней значение 1, что соответствует покрытию «базового» наблюдения a. Наибольшее значение, равное $|K^+|$, функция $C(X)$ принимает в точке $Y^0 = (0,0,...,0)$.

Положим, что ближайшее к a наблюдение $b \in K^+$ отличается значением s компонент. Во всех точках $Y \in O_k(Y^1)$, $k = 0,1,...,s-1$, значение целевой функции будет одинаковым и равным 1. Наличие такого множества постоянства затрудняет работу алгоритмов оптимизации, начинающих поиск из допустимой точки Y^1 и ведущих его по соседним точкам, так как вычисление целевой функции в системе окрестностей, состоящей из соседних точек, не дает информации о наилучшем направлении поиска. При решении практических задач больших размерностей это множество постоянства может быть таким, что ему принадлежит большая часть точек допустимой области.

Другим следствием наличия множеств постоянства целевой функции является то, что оптимальным решением является не только точка, принадлежащая подмножеству крайних точек допустимой области, а это может быть целое множество точек, представляющее собой множество постоянства целевой функции. При этом справедливы следующие утверждения.

Утверждение 1. Крайние точки допустимой области задачи (1)-(2) соответствуют первичным закономерностям.

Утверждение 2. Оптимальное решение задачи (1)-(2) соответствует сильной первичной закономерности.

Таким образом, применяя приближенные алгоритмы оптимизации, можно утверждать, что найденная закономерность будем являться первичной, но она не обязательно будет являться сильной. Если же использовать точный алгоритм оптимизации, то найденная закономерность будет являться сильной первичной закономерностью.

Модель оптимизации определяется, прежде всего, тем, каким образом введены переменные, определяющие альтернативы задачи оптимизации. В рассмотренной выше модели оптимизации альтернативы определяются включением или не включением в закономерность условий, которые выполняются в некотором базовом наблюдении. Рассмотрим альтернативный способ задания закономерности.

Введем в рассмотрение бинарную переменную

$$z_b = \begin{cases} 1, \text{если наблюдение } b \in K^+ \text{ покрывается } P^a, \\ 0, \text{в противном случае.} \end{cases}$$

Для того чтобы получаемая закономерность была охватывающей, введем в ее состав все литералы, имеющиеся у всех положительных наблюдений, которые эта закономерность покрывает, то есть те литералы, для которых выполняется условие: $\prod_{b\in K^+}\left(1-\left|b_j - a_j\right|\cdot z_b\right)=1$.

Теперь сформулируем задачу поиска максимальной закономерности как максимизацию покрытия положительных наблюдений при условии недопустимости покрытия отрицательных наблюдений:

$$\sum_{b\in K^+} z_b \to \max_Z, \tag{4}$$

$$\sum_{\substack{j=1 \\ c_j \neq a_j}}^{n} \prod_{b\in K^+}\left(1-\left|b_j - a_j\right|\cdot z_b\right)\geq 1 \text{ для любого } c\in K^-. \tag{5}$$

Решив эту задачу оптимизации, мы определим множество положительных наблюдений, покрываемых искомой закономерностью. Саму закономерность определим, используя характеристические переменные $Y=(y_1, y_2,..., y_n)$:

$$y_j = \prod_{b\in K^+}\left(1-\left|b_j - a_j\right|\cdot z_b\right),\ j=1,...,n.$$

В такой постановке задачи оптимизации целевая функция является строго монотонной, а множества постоянства целевой функции отсутствуют.

Утверждение 3. В задаче (4)-(5) крайние точки допустимой области, и только они, соответствуют сильным охватывающим закономерностям.

В **четвертой главе** описывается применение закономерностей для принятия решений при распознавании, и исследуются схемы использования семейств закономерностей различных типов.

Предположим, что найдено некоторое число положительных и отрицательных закономерностей. Согласно логическому анализу данных, для принятия решения относительно принадлежности некоторого распознаваемого наблюдения одному из классов, используется следующее правило (рис.1):

1) Если наблюдение покрывается только положительными закономерностями, то оно считается положительным.

2) Если наблюдение покрывается только отрицательными закономерностями, то оно считается отрицательным.

3) Если наблюдение подчиняется условиям *t* закономерностей одного класса и *f* другого, то класс наблюдения определяется в результате голосования, например, как результат разности $t/T - f/F$, где T и F – число закономерностей этих классов.

4) Если наблюдение не покрывается ни одной закономерностью, то оно считается нераспознанным.

Здесь имеются однозначные области, которые покрываются закономерностями только одного класса; конфликтная область, точки которой покрываются закономерностями разных классов (в этом случае принадлежность к классу определяется

голосованием закономерностей); а также область, не покрываемая ни одной закономерностью (наблюдения этой области не могут быть распознаны).

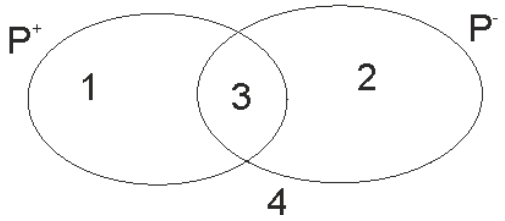


Рис. 1. Поясняющая схема для правила принятия решения

Использование закономерностей различных видов имеет свои существенные особенности:

$$Cov(P^{CO}) = Cov(P^{СП}),\ S(P^{CO}) \subseteq S(P^{СП}),\ Lit(P^{СП}) \subseteq Lit(P^{CO}),$$

где $P^{со}$ – сильная охватывающая закономерность, $P^{сп}$ – сильная первичная закономерность.

Первичные закономерности более простые, состоят из меньшего числа условий. Использование первичных закономерностей уменьшает число нераспознанных наблюдений. Использование сильных охватывающих закономерностей позволяет получать классификаторы с лучшей обобщающей способностью.

Предлагается подход, состоящий в совместном использовании двух видов закономерностей. А именно, построение и использование закономерностей попарно – сильной охватывающей и сильной первичной. Это позволяет совместить преимущества этих двух видов закономерностей.

Охватывающие закономерности являются более надежными. Первичные – более простые и затрагивают больше потенциальных наблюдений. Это позволяет повысить интерпретируемость распознавания и сделать принятие решения более обоснованным (рис.2).

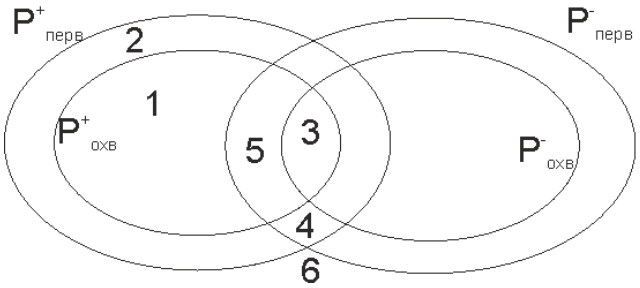


Рис. 2. Схема, поясняющая принятие решения при использовании пар закономерностей.
Расшифровка областей:

- 1- покрытие охватывающей закономерностью одного класса,
- 2- покрытие только первичными закономерностями одного класса,
- 3- покрытие охватывающими закономерностями разных классов,
- 4- покрытие только первичными закономерностями разных классов,
- 5- покрытие охватывающей закономерностью одного класса и первичными закономерностями другого класса,
- 6- нет покрытия закономерностями.

Для нахождения пары закономерностей предлагается следующий алгоритм.

Алгоритм поиска пары закономерностей (первый вариант)

1. Найти сильную первичную закономерность $P_{перв}$, решив задачу

$$\sum_{b\in K^+}\prod_{\substack{i=1\\ b_i\neq a_i}}^{n}(1-y_i)\to\max_{Y},\ \sum_{\substack{i=1\\ c_i\neq a_i}}^{n}y_i\geq 1 \text{ для всех } c\in K^-.$$

2. Определить множество наблюдений S, которые покрываются $P_{перв}$: S=Cov($P_{перв}$).

3. Найти соответствующую $P_{перв}$ охватывающую закономерность $P_{охв}=\prod_{i\in I}x_i^{\alpha_i}$, где I – множество всех индексов i, для которых i-ые компоненты всех векторов XϵS имеют одинаковой значение.

В этом способе проблема заключается в том, что для нахождения сильной первичной закономерности требуется точное решение задачи оптимизации – нахождение наилучшей крайней точки. Приближенное решение даст лишь первичную закономерность с возможно далеко не лучшим покрытием.

Решение проблемы заключается в использовании более совершенного алгоритма оптимизации, способного найти точное решение, который описывается в следующей главе.

Альтернативная схема поиска закономерностей попарно состоит в использовании второй модели оптимизации для поиска сильной охватывающей закономерности, а затем нахождении сильной первичной закономерности путем исключения части литералов. Следует отметить, что одной сильной охватывающей закономерности соответствует множество сильных первичных закономерностей с разным числом условий. Поэтому необходимо решать дополнительную задачу оптимизации, чтобы найти первичную закономерность с наименьшим числом условий:

Алгоритм поиска пары закономерностей (второй вариант)

1. Найти сильную охватывающую закономерность $P_{охв}$, решив задачу $\sum_{b\in K^+}z_b\to\max_{Z},\ \sum_{\substack{j=1\\ c_j\neq a_j}}^{n}\prod_{b\in K^+}\left(1-\left|b_j-a_j\right|\cdot z_b\right)\geq 1$ для всех $c\in K^-$.

2. Определить $y_j=\prod_{b\in K^+}\left(1-\left|b_j-a_j\right|\cdot z_b\right),\ j=1,...,n$.

3. Найти соответствующую $P_{охв}$ первичную закономерность $P_{перв}$, решив задачу:

$$\sum_{i=1}^{n}y_i'\to\min_{Y},\ y_i'\leq y_i \text{ для всех } i=1,...,n,\ \sum_{\substack{i=1\\ c_i\neq a_i}}^{n}y_i'\geq 1 \text{ для всех } c\in K^-.$$

Предлагается единая модель оптимизации для одновременного выявления пары закономерностей. За основу взята исходная модель оптимизации для поиска первичных закономерностей. Имеющиеся неоднозначности исключены путем добавления дополнительных ограничений.

Переменные $Y=(y_1,y_2,...,y_n)$ определяют сильную охватывающую закономерность. Переменные $Y'=(y_1',y_2',...,y_n')$ определяют сильную первичную закономерность. Добавлены ограничения для выявления первичной закономерности:

$$\exists c\in K^-:\sum_{\substack{i=1\\ c_i\neq a_i}}^{n}y_i'=1 \text{ или иначе } \min_{c\in K^-}\sum_{\substack{i=1\\ c_i\neq a_i}}^{n}y_i'=1.$$

В свою очередь, для охватывающей закономерности должно выполняться усло-

вие: $y_i \geq 1 - \sum_{b \in K^+} |a_i - b_i| \cdot z_b$ для всех $i = 1,...,n$,

$$\text{где } z_b = \prod_{\substack{i=1 \\ b_i \neq a_i}}^{n} (1 - y_i) = \prod_{i=1}^{n} (1 - |a_i - b_i| \cdot y_i).$$

В **пятой главе** описываются алгоритмы оптимизации для решения задач, рассмотренных в предыдущих главах, и предлагается алгоритм условной псевдобулевой оптимизации, основанный на поиске среди граничных точек и методе ветвей и границ. Использование свойств задачи и схемы метода ветвей и границ позволяет быстро исключать из рассмотрения области, в которых нет оптимального решения. Метод ветвей и границ используется разными исследователями для широкого круга задач, например, для задач размещения (Береснев В.Л., 2014).

Рассмотрим класс задач вида

$$\begin{cases} C(X) \to \max\limits_{X \in B_2^n} \\ A_j(X) \leq H_j,\ j = \overline{1,m} \end{cases}$$

где $B_2^n = \{0,1\}^n$ – пространство бинарных переменных, $C(X)$ и $A_j(X)$ – псевдобулевые функции (действительные функции бинарных переменных), в общем случае заданные алгоритмически, H_j – некоторые действительные числа. Рассмотрим класс задач, в котором целевая функция $C(X)$ и функция $A(X)$, определяющая систему ограничений, монотонно возрастают от точки $X^0 = (x_1^0,...,x_n^0)$. К этому классу задач относятся, в частности, задачи поиска оптимальных закономерностей, рассмотренные в третьей главе.

Обозначим $X^1 = (x_1^1,...,x_n^1) = (\overline{x_1^0},...,\overline{x_n^0})$. Всё множество точек пространства B_2^n можно представить в виде подкуба $K(X^0, X^1)$ - это исходный подкуб, в котором выполняется поиск оптимального решения.

Допустим, что найдена некоторая крайняя точка $X' \in B_2^n$. Тогда подкубы $K(X', X^0)$ и $K(X', X^1)$ можно исключить из дальнейшего рассмотрения.

Введём вспомогательную переменную

$$z_i = \begin{cases} x_i, & \text{if } x_i^0 = 0, \\ \bar{x}_i, & \text{if } x_i^0 = 1. \end{cases}$$

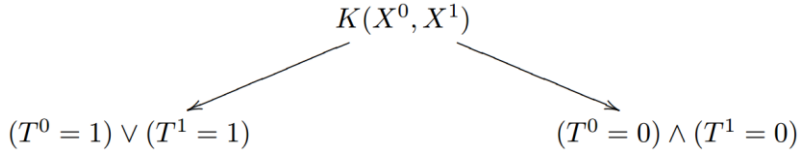
Тогда подкуб $K(X', X^0)$ можно представить как множество точек, для которых выполняется

$$T^0 = \bigwedge_{i \in B(X')} \bar{z}_i$$

А подкуб $K(X', X^1)$ можно описать выражением

$$T^1 = \bigwedge_{i \in A(X')} z_i$$

Разделим подкуб $K(X^0,X^1)$ на две части.



Левая часть, как было сказано выше, исключается из дальнейшего рассмотрения. Правая часть (T^0=0)∧(T^1=0) может быть представлена как множество подкубов.

Рассмотрим условие (T^1=0). Оно выполняется в том случае, когда $z_i = 0$ хотя бы для одного $i \in A(X')$. При $|A(X')|>1$ множество точек, выполняющих условие $T^1 = 0$, не может быть представлено в виде одного подкуба, а только как набор подкубов. Наиболее очевидный способ разбить это множество точек на k подкубов - зафиксировать поочередно значение переменной $z_i = 0$ для $i \in A(X')$. В этом случае получаем k подкубов размерностью $n-1$. Недостаток этого способа состоит в том, что полученные подкубы существенно пересекаются между собой.

Для того, чтобы этого избежать, используем следующий подход. Первый подкуб K_1^1 получим, зафиксировав одну переменную: $z_{i_1} = 0$. Для второго K_2^1 зафиксируем две переменных: $z_{i_1} = 1$ и $z_{i_2} = 0$. Для третьего K_3^1 три перменных: $z_{i_1} = 1$, $z_{i_2} = 1$ и $z_{i3} = 0$. И так далее. В результате получаем k подкубов различной размерности. Такой подход гарантирует, что получаемые подкубы не пересекаются.

То же самое проделаем для условия (T^0=0). Соответствующее множество точек следует разбить на ($n-k$) подкубов путем фиксации переменных $j \in B(X')$. Для первого подкуба K_1^0 зафиксируем одну переменную: $z_{j_1} = 1$. Для второго K_2^0 зафиксируем две переменных: $z_{j_1} = 0$ и $z_{j_2} = 1$. Для третьего K_3^0 три перменных: $z_{j_1} = 0$, $z_{j_2} = 0$, $z_{j_3} = 1$. Для ($n-k$)-го подкуба K_{n-k}^0. $z_{j_s} = 0$, $s = 1,...,n-k-1$, $z_{j_{n-k}} = 1$.

В результате получаем два набора подкубов. Множество точек, выполняющих условие (T^0=0)∧(T^1=0), соответствует объединению всевозможных пересечений пар подкубов, взятых из этих двух наборов.

Итак, найдя в подкубе некоторую крайнюю точку $X' \in O_k(X^0)$, этот подкуб разбивается на две части, одна из которых отбрасывается, а из другой образуются $k \cdot (n-k)$ новых ветвей. Каждая из этих ветвей является подкубом, с которым может быть произведена такая же процедура ветвления, что описана выше.

Обозначим $\overline{X}$ и $\underline{X}$ соответственно верхнюю и нижнюю точки некоторого подкуба.

Подкуб $K(\underline{X}, \overline{X})$ может содержать оптимальное решение, только если выполняются условия:

1. В подкубе имеются допустимые решения.
2. Верхняя граница соответствующей ветви выше найденного наилучшего решения.

Так как функция ограничения $A(X)$ монотонно возрастает от точки X^0, то в пределах подкуба $K(\underline{X}, \overline{X})$ функция $A(X)$ монотонно возрастает от точки $\underline{X}$, принимая в ней минимальное значение. Поэтому если точка $\underline{X}$ является недопустимой, то недопустимы все точки этого подкуба.

Целевая функция $C(X)$ в пределах подкуба также монотонно возрастает от точки $\underline{X}$, принимая наибольшее значение в точке $\overline{X}$. Сама эта точка может быть и недопустима, но значение $C(\overline{X})$ может быть использовано как верхняя граница ветви, соответствующей этому подкубу.

Также, если точка $\overline{X}$ является допустимой, то все остальные точки этого подкуба заведомо не лучше, кроме того, в подкубе при этом отсутствуют крайние точки, за исключением разве что $\overline{X}$.

Итак, подкуб $K(\underline{X}, \overline{X})$ исключается из дальнейшего поиска при выполнении хотя бы одного из условий:

- Точка $\underline{X}$ является недопустимой.
- Точка $\overline{X}$ является допустимой.
- Значение верхней границы $C(\overline{X})$ не превышает уже найденного наилучшего допустимого значения целевой функции.

Такая проверка, включая вычисление верхней границы, требует просмотра всего двух точек подкуба.

Для того чтобы произвести ветвление описанным выше способом, необходимо найти некоторую крайнюю точку, принадлежащую рассматриваемому подкубу. Это не обязательно должно быть наилучшее решение в данном подкубе. Тем не менее, хорошее решение может повысить рекорд (наилучшее найденное допустимое решение) и улучшает отсев новых ветвей. Для нахождения крайней точки используется жадный алгоритм, также могут использоваться алгоритмы случайного поиска (Растригин Л.А., Антамошкин А.Н., Крутиков В.Н. и др.).

На рис. 3 и 4 приведены результаты экспериментального исследования работы алгоритма на задачах условной псевдобулевой оптимизации, сгенерированных случайным образом.

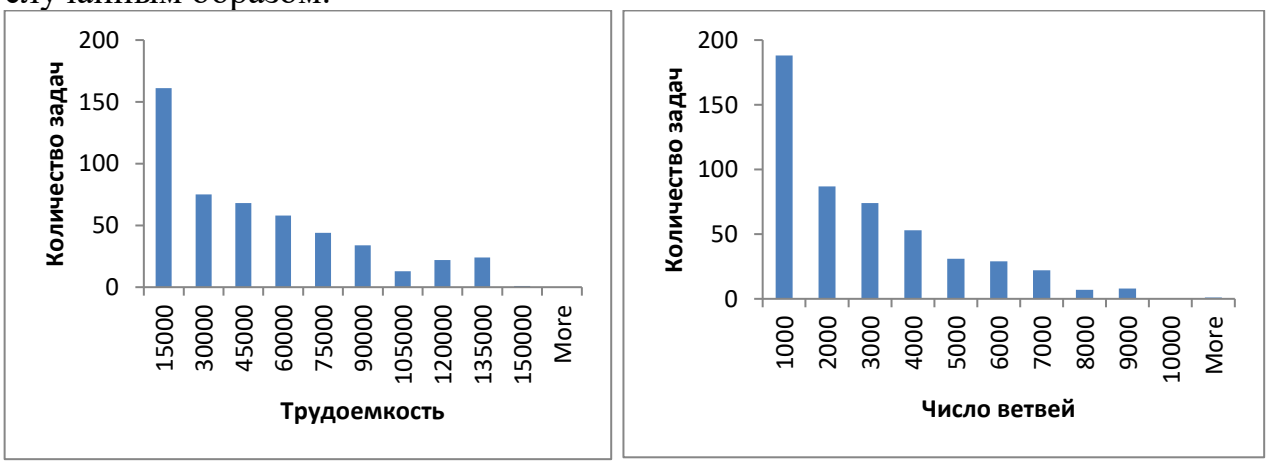


Рис. 3. Трудоемкость и число ветвей для $n = 20$ (точное решение найдено)

На гистограммах показаны распределения трудоемкости и числа ветвей, полученных в результате точного решения 500 задач размерностью 20. Гистограммы рис. 3 относятся к полному решению задач, то есть не остается открытых ветвей, и таким образом, точность решения доказана. Гистограммы рис. 4 показывают трудоемкость и число ветвей для момента, когда наилучшее решение уже найдено, но точность еще не доказана, то есть остались открытые ветви.

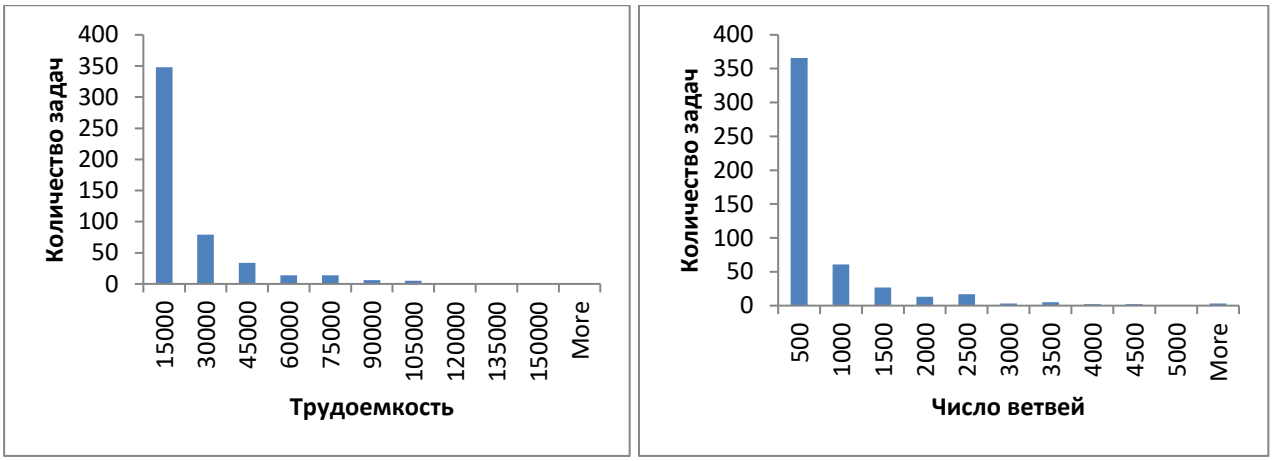


Рис. 4. Трудоемкость и число ветвей для $n = 20$ (точность не доказана)

На рис. 5 графики показывают изменение числа открытых ветвей и значения лучшего найденного решения в процессе поиска. Представлены результаты для отдельно взятых задач размерностью 20 и 100, но это является типичной картиной. Как видно, лучшее найденное значение возрастает на первых ветвлениях, достигая, возможно, оптимального значения, а затем не меняется. Остальная часть поиска нужна лишь для подтверждения оптимальности найденного решения.

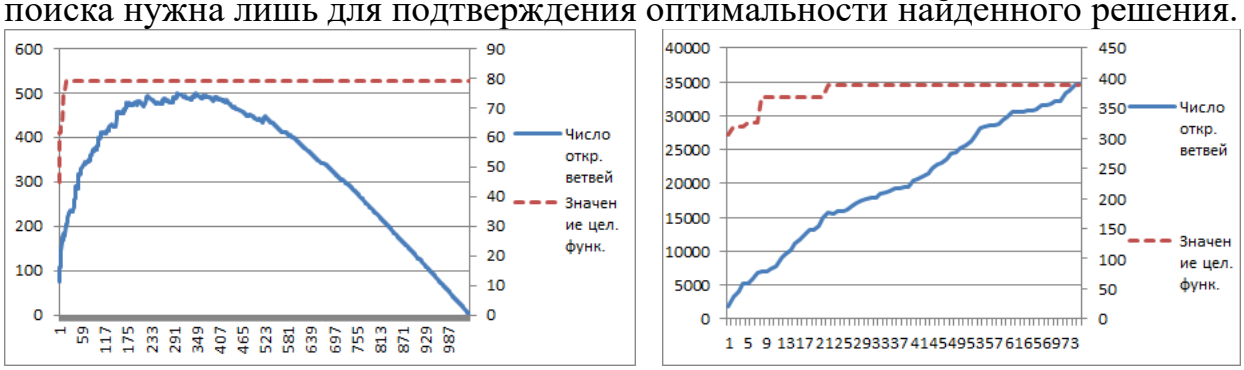


Рис. 5. Число открытых ветвей и значение рекорда для $n = 20$ и n=100.

Таким образом, из практических соображений, достаточно произвести лишь несколько итераций алгоритма, чтобы значительно улучшить результат жадного алгоритма. Идея улучшения приближенного алгоритма, после чего он способен выдавать решение равное или очень близкое оптимальному, ранее уже применялась, например, в алгоритме локального поиска по обобщенной окрестности для оптимизации псевдобулевых функций (Береснев В.Л. и др., 2011).

В **шестой главе** рассматривается применение метода оптимальных логических решающих правил на практических задачах.

Экспериментальные исследования по сравнению точности распознавания на популярных задачах (репозиторий задач машинного обучения) подтверждают конкурентоспособность предлагаемого метода (для сравнения с другими методами использовалось программное приложение Weka) (табл. 1).

Одно из практических внедрений метода – решение задачи классификации электрорадиоизделий (ЭРИ) по производственным партиям для комплектации радиоэлектронной аппаратуры космических аппаратов.

Требования к качеству электронной компонентной базы, используемой в бортовой аппаратуре, не выполняются для изделий, выпускаемых российскими предприятиями. Кроме того, поставляемые партии ЭРИ могут быть неоднородными.

Таблица 1. Сравнение точности на тестовых задачах

Набор данных	Логист. регр.	C4.5	Случ. лес	Нейр. сеть	SVM	ОЛРП
boston	0,87	0,84	0,87	**0,89**	**0,89**	0,87
bupa	0,66	0,63	0,73	0,64	0,70	**0,74**
breast c.w.	0,96	0,94	**0,97**	0,95	0,96	**0,97**
chess (krkp)	0,97	**0,99**	**0,99**	**0,99**	**0,99**	**0,99**
credit ap.	0,87	0,85	**0,88**	0,83	0,86	0,87
heart	0,83	0,80	0,83	0,81	**0,84**	0,83
hepatitis	0,76	0,65	0,72	0,73	0,77	**0,78**
pima	0,73	0,72	0,73	0,73	0,73	**0,75**
sick	0,81	**0,92**	0,83	0,85	0,82	0,83
voting	**0,96**	**0,96**	**0,96**	0,94	**0,96**	**0,96**

Тестированием ЭРИ занимается специализированное предприятие – ИТЦ НПО ПМ в Железногорске. Дефекты обнаруживаются проведением сотен неразрушающих тестов и ряда разрушающих. Распространять результаты разрушающих испытаний на всю партию можно только при условии, что эта партия однородна по технологии и сделана из одной партии сырья.

В работах Патраева В.Е., Орлова В.И., Федосова В.В. доказано, что однородная партия дает схожие результаты неразрушающих тестов. Изделия с различающимися эксплуатационными характеристиками (фактически различные партии ЭРИ) имеют различия в полученных результатах испытаний. Это дает возможность применения методов анализа данных для осуществления необходимой группировки ЭРИ.

Ранее разработана система, позволяющая производить выявление однородных производственных партий в сборной партии электрорадиоизделий космического применения (Казаковцев и Масич, 2016). Система основана на использовании алгоритма с жадной эвристикой.

Исходными данными для анализа при решении задачи являются результаты тестовых воздействий на ЭРИ по контролю вольт-амперных характеристик входных и выходных цепей микросхем. Данные представляют собой таблицу, в строках которой приведены последствия различных электрических воздействий на элементы набора однотипных ЭРИ.

Исследовалась задача повышения эффективности классификации посредством формирования информативных закономерностей и разработки процедур, позволяющих улучшить интерпретируемость классификатора. Построение подобных классификаторов может быть основано на различных методах, среди которых наиболее перспективными для данной задачи являются методы логической классификации, отличающиеся высокой интерпретируемостью результатов классификации. Интерпретируемость результатов логической классификации в условиях космического производства означает возможность разработки ужесточенных норм параметров ЭРИ.

Проведенные исследования по использованию методов классификации, основанных на правилах, показывают, что для распознавания группы электрорадиоиз-

делий достаточно использование небольшого числа логических правил (от 1 до 4). Эти правила представляют собой сравнение значения некоторого признака с определенным в процессе построения правил порогом. Примеры получаемых закономерностей при классификации микросхемы 1526ЛЕ5:

Класс a (196 изделий):

(TEST_34 < 5) ∧ (TEST_127 < 4,76) => класс a (100%)

(TEST_35 < 5) ∧ (TEST_37 < 5) ∧ (TEST_127 < 4,76) => класс a (100%)

Класс b (218 изделий):

(TEST_66 ≥ 0,682) ∧ (TEST_122 ≥ 4,98) => класс b (97,2%)

(TEST_66 ≥ 0,682) => класс b (97,7%)

Класс c (205 изделий):

(TEST_34 ≥ 5) ∧ (TEST_58 < 0,93) => класс c (99%)

Использование предлагаемого подхода дает возможность принимать решение о принадлежности изделия партии по данным небольшого числа тестов с использованием простых правил сравнения. Оказывается достаточным использование весьма небольшого числа признаков (выбранных определенным образом в процессе построения правил из всего набора признаков – результатов тестов) для успешной классификации ЭРИ. Таким образом, в работе решена задача выявления закономерностей для классификации ЭРИ по результатам дополнительных отбраковочных испытаний с целью дальнейшего прогнозирования показателей безотказности электронной компонентной базы.

Ещё одной практической задачей, решаемой в рамках данной работы, является задача прогнозирования осложнений инфаркта миокарда (ИМ). Результат болезни пациентов, у которых диагностирован ИМ, может быть различен. Необходимым этапом при назначении лечения является прогнозирование осложнений ИМ, возникновение которых может привести к серьезному ухудшению и даже летальному исходу. На таком прогнозе основывается индивидуальная терапия, которая не может быть одинаковой для всех пациентов. Выявление возможных осложнений является трудной задачей, верно и своевременно решить которую не всегда удается даже опытным специалистам.

В качестве исходных данных используется информация о течении заболевания у 1700 больных ИМ, полученная из историй болезни в Кардиологическом центре городской больницы № 20 г. Красноярска. Таблица содержит сведения о данных анамнеза каждого больного, клинике ИМ, электрокардиографических, лабораторных показателях, лекарственной терапии и особенностях течения заболевания в первые дни ИМ. Признаки принадлежат различным типам, большинство из них бинарные, но имеются также номинальные и количественные. Всего 1700 наблюдений, описанных 124 признаками. В исходных данных много пропущенных значений.

Задача заключается в прогнозировании ряда наиболее значимых осложнений: фибрилляции предсердий (ФП), фибрилляции желудочков (ФЖ), отека легких (ОЛ), разрыва сердца (РС), летального исхода (ЛИ). При этом специалисты имеют потребность в классификаторе, представляющем собой решающее правило в виде логических высказываний.

Результаты экспериментов показывают, что точность задачи предлагаемым ме-

тодом в большинстве случаев превосходит точность решений другими методами, основанными на выявлении и использовании правил (табл. 2).

Таблица 2. Точность решения задачи

Задача	RIPPER	CART	C4.5	Random Forest	Adaboost	ОЛРП
ФП	0,66	0,62	0,7	0,7	0,74	0,76
ФЖ	0,867	0,633	0,683	0,833	0,9	0,865
РС	0,786	0,857	0,714	0,857	0,893	0,965
ОЛ	0,692	0,718	0,667	0,769	0,697	0,835
ЛИ	0,74	0,74	0,66	0,76	0,74	0,86

Примеры полученных правил приведены на рис. 6.

```
========================== Отрицательные правила ======================
Правило №1   (63 <  AGE) (INF_IM  < 2) (L_BLOOD  < 12) (TEMPER_3_N  < 2)
Правило №2   (1 <  DLIT_AG) (3 <  TIME_B_S) (NA_R_1_N  < 2) (NOT_NA_3_N  < 2)
Правило №3   (SEX = 0) (STENOK_AN  < 5) (N_R_ECG_P__2 = 0) (NOT_NA_3_N  < 2)
Правило №4   (STENOK_AN  < 3) (IBS_POST  < 2) (SIM_GIPERT = 0) (ZSN_A  < 1) (INF_IM  < 2) (IM_PG_P = 0) (N_P_ECG_P__8 = 0)
Правило №5   (1 <  IBS_POST) (NR_11 = 0) (TEMPER_3_N  < 2) (GEPAR_S_N = 0)
Правило №6   (1 <  FK_STENOK) (NR_11 = 0) (RITM_ECG_P_2 = 0) (N_R_ECG_P__2 = 0) (5 < 18 ROE) (ROE 5 < 18) (TEMPER_3_N  < 3)
Правило №7   (RITM_ECG_P_5 = 0) (NITR_S = 0) (NA_R_1_N  < 1) (NOT_NA_1_N  < 3) (ANT_CA_S_N = 1)
Правило №8   (SEX = 1) (STENOK_AN  < 6) (ZAB_LEG_02 = 0) (130 <  NA_BLOOD) (TIME_B_S  < 9) (NA_R_1_N  < 1)
Правило №9   (2 <  GB) (NR_04 = 0) (ENDOCR_02 = 0) (ZAB_LEG_02 = 0) (INF_IM  < 4) (N_R_ECG_P_ = 0) (NA_R_1_N  < 2)
Правило №10   (STENOK_AN  < 4) (ANT_IM  < 1) (RITM_ECG_P = 1)
========================== Положительные правила ======================
Правило №1   (STENOK_AN  < 4) (ENDOCR_02 = 0) (FIB_G_POST = 0) (1 <  ANT_IM) (LID_S_N = 1) (GEPAR_S_N = 1)
Правило №2   (58 <  AGE) (70 <  D_AD_ORIT) (0,23 <  ALT_BLOOD)
Правило №3   (S_AD_ORIT  < 140) (R_AB_2_N  < 2)
Правило №4   (ENDOCR_02 = 0) (O_L_POST = 0) (1 <  NA_R_1_N) (LID_S_N = 1)
Правило №5   (D_AD_ORIT  < 80)
Правило №6   (D_AD_ORIT  < 90)
Правило №7   (MP_TP_POST = 0) (2 <  ANT_IM) (RITM_ECG_P_2 = 0) (K_BLOOD  < 4,1) (NA_BLOOD  < 140)
Правило №8   (FK_STENOK  < 3) (S_AD_ORIT  < 140)
Правило №9   (D_AD_ORIT  < 90)
Правило №10   (ZSN_A  < 1) (ANT_IM  < 1) (FIBR_TER_0_1 = 0) (8 <  L_BLOOD) (TEMPER_1_N  < 1)
```

Рис. 6. Примеры закономерностей

Точность решения задачи прогнозирования осложнений методом оптимальных логических решающих правил сопоставима с точностью решения методом нейронных сетей (Горбань А.Н.). При этом предлагаемый метод в явном виде предоставляет правила, по которым принимается решение, что является преимуществом этого метода для специалистов (табл. 3).

Таблица 3. Сравнение результатов

Прогнозируемое осложнение	Класс	Точность классификации	
		Нейронные сети	ОЛРП
ФП	чувствительность	0,90	0,78
	специфичность	0,70	0,74
ФЖ	чувствительность	0,76	0,83
	специфичность	0,70	0,90
РС	чувствительность	0,80	0,93
	специфичность	0,70	1,00
ОЛ	чувствительность	0,85	0,89
	специфичность	0,80	0,78
ЛИ	чувствительность	0,86	0,85
	специфичность	0,80	0,87

ЗАКЛЮЧЕНИЕ

В диссертации предложен новый метод анализа данных для поддержки принятия решений при распознавании, состоящий в выявлении в данных оптимальных логических закономерностей с помощью алгоритмов псевдобулевой оптимизации, позволяющий успешно решать задачи классификации, предоставляя обоснование и объяснение решений в виде логических правил,

подтверждаемых прецедентами.

Цель диссертации достигнута путём решения поставленных задач, а именно:

1. Проведён анализ существующих способов выявления закономерностей в данных и исследованы свойства моделей оптимизации для нахождения закономерностей. Показано, что существующие алгоритмы не гарантируют получения сильных закономерностей с максимальным покрытием. Разработана новая модель оптимизации для нахождения сильных охватывающих закономерностей в данных.

2. Изучение особенностей применения различных типов закономерностей позволило установить, что использование первичных закономерностей уменьшает число нераспознанных наблюдений, а использование сильных охватывающих закономерностей позволяет снизить ошибки в распознавании. Разработан новый подход к поддержке принятия решений при распознавании, заключающийся в совместном использовании закономерностей двух типов – сильных первичных и сильных охватывающих закономерностей.

3. Разработана единая модель оптимизации в рамках метода оптимальных логических решающих правил для поиска пары (первичной и охватывающей) закономерностей.

4. Построен новый алгоритм условной псевдобулевой оптимизации на основе схемы ветвей и границ и поиска среди граничных точек допустимой области, позволяющий находить лучшее решение, чем жадный алгоритм. Алгоритм не требует алгебраического задания функций (целевой и ограничений), функции могут быть заданы алгоритмически (оптимизация черного ящика), что позволяет его применять для задач с нелинейными функциями, в том числе для задачи поиска логических закономерностей в данных. Экспериментально показана эффективность приближенного варианта алгоритма с использованием ранней остановки: достаточно произвести лишь несколько итераций алгоритма, чтобы получить решение, лучшее, чем получаемое жадным алгоритмом.

5. Разработан алгоритм поиска пары закономерностей (сильной первичной и сильной охватывающей) с использованием нового алгоритма условной псевдобулевой оптимизации.

6. Разработана новая модель оптимизации для нахождения оптимального назначения порогов количественных признаков, которая не просто определяет наименьшее число порогов, достаточных для разделения наблюдений разных классов, а позволяет выбрать такие пороги, которые наилучшим образом разделяют наблюдения разных классов в пространстве булевых признаков.

7. Впервые предложена комплексная процедура ускорения поиска закономерностей, делающая возможным применение метода для случаев большого объема данных. Предлагаемое улучшение состоит в применении нового способа выбора базовых наблюдений для формирования закономерностей, отборе признаков путём решения задачи псевдобулевой оптимизации и применении приближенного варианта нового алгоритма псевдобулевой оптимизации.

8. Разработана процедура повышения интерпретируемости классификатора, основанного на закономерностях, заключающаяся в нахождении закономерностей

с лучшей обобщающей способностью, новой схемы использования одновременно двух видов закономерностей (сильной первичной и сильной охватывающей) и отборе достаточного числа закономерностей с ограниченным числом условий на основе решения задачи условной псевдобулевой оптимизации.

9. Решена задача выявления закономерностей для классификации электрорадиоизделий космического применения по результатам дополнительных отбраковочных испытаний с целью дальнейшего прогнозирования показателей безотказности электронной компонентной базы. Использование предлагаемого подхода дает возможность принимать решение о принадлежности изделия партии по данным небольшого числа тестов с использованием простых правил сравнения.

Совокупность новых моделей и алгоритмов являются содержанием нового метода решения задач классификации, результаты распознавания в которых должны быть обоснованы и интерпретированы в виде логических правил. Новый метод – метод оптимальных логических решающих правил – основан на нахождении оптимальных решений при выявлении логических закономерностей в соответствие с моделями оптимизации, относящимися к классу задач оптимизации монотонных псевдобулевых функций с монотонными ограничениями, при этом функции заданы алгоритмически. Для решения этих задач использован новый алгоритм условной псевдобулевой оптимизации, реализующий свойства данного класса задач и основанного на поиске среди граничных точек допустимой области и схеме ветвей и границ. Применение оптимальных логических решающих правил обеспечивает повышение точности решения задач классификации с выполнением условий доказательности и интерпретируемости.

Основные результаты работы опубликованы в следующих статьях автора.

А) **Статьи в российских периодических изданиях**, рекомендованных ВАК:

1. Масич И.С. Метод оптимальных логических решающих правил для задач распознавания и прогнозирования. // Системы управления и информационные технологии. 2019. Т. 75. № 1. С. 31-37.

2. Кузьмич Р.И., Масич И.С., Ступина А.А. Модели формирования закономерностей в методе логического анализа данных. // Системы управления и информационные технологии. 2017. Т. 67. № 1. С. 33-37.

3. Федосов В.В., Казаковцев Л.А., Масич И.С. Метод нормировки исходных данных испытаний электрорадиоизделий космического применения для алгоритма автоматической группировки. // Системы управления и информационные технологии. 2016. Т. 65. № 3. С. 92-96.

4. Орлов В.И., Казаковцев Л.А., Масич И.С. Применение критерия силуэта в алгоритме автоматической группировки электрорадиоизделий космического применения. // Вестник СибГАУ. 2016. Т. 17. № 4. С. 883-890.

5. Антамошкин А.Н., Масич И.С. Поисковые алгоритмы условной псевдобулевой оптимизации. // Системы управления, связи и безопасности. 2016. № 1. С. 103-145.

6. Казаковцев Л.А., Масич И.С., Орлов В.И., Федосов В.В. Быстрый детерминированный алгоритм для классификации электронной компонентной базы

по критерию равнонадежности. // Системы управления и информационные технологии. 2015. Т. 62. № 4. С. 39-44.

7. Антамошкин А.Н., Масич И.С. Обнаружение закономерностей в данных для распознавания объектов как задача условной псевдобулевой оптимизации. // Вестник СибГАУ. 2015. Т. 16. № 1. С. 16-21.

8. Казаковцев Л.А., Масич И.С., Орлов В.И., Федосов В.В. Детерминированный алгоритм для задач классификации электрорадиоизделий. // Информационные технологии моделирования и управления. 2015. Т. 96. № 6. С. 519-525.

9. Кузьмич Р.И., Масич И.С. Модификация целевой функции при построении паттернов для увеличения различности правил в модели классификации. Системы управления и информационные технологии. 2014. Т. 56. № 2.

10. Казаковцев Л.А., Орлов В.И., Ступина А.А., Масич И.С. Задача классификации электронной компонентной базы. // Вестник СибГАУ. 2014. № 4 (56). С. 55-61.

11. Антамошкин А.Н., Масич И.С. Выбор логических закономерностей для построения решающего правила распознавания. // Вестник СибГАУ. 2014. № 5 (57). С. 20-25.

12. Масич И.С., Краева Е.М. Отбор закономерностей для построения решающего правила в логических алгоритмах распознавания. // Системы управления и информационные технологии. 2013. Т. 51. № 1.1. С. 170-173.

13. Кузьмич Р.И., Масич И.С. Построение модели классификации как композиции информативных паттернов. // Системы управления и информационные технологии. 2012. Т. 48. № 2. С. 18-22.

14. Антамошкин А.Н., Масич И.С. Исследование свойств задач оптимизации при поиске логических закономерностей в данных. // Системы управления и информационные технологии. 2011. N4.1(46). С. 111-115.

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

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

17. Antamoshkin A.N., Masich I.S. Combinatorial optimization and rule search in logical algorithms of machine learning. // Engineering & automation problems (Проблемы машиностроения и автоматизации). V.7. N.1. 2010. P. 52-57.

18. Масич И.С. Модель логического анализа для прогнозирования осложнений инфаркта миокарда. // Информатика и системы управления. №3 (25). 2010. С. 48-56.

19. Masich I.S. Combinatorial optimization in foundry production planning. // Вестник СибГАУ, выпуск 2(23). 2009. С. 40-44.

20. Масич И.С. Комбинаторная оптимизация в задаче классификации. // Системы управления и информационные технологии. 2009. №1.2(35). С. 283-288.

21. Antamoshkin A.N., Masich I.S. Unimprovable algorithm for monotone pseudo-Boolean function conditional optimization. // Engineering & automation problems (Проблемы машиностроения и автоматизации). V. 6. N. 1. 2008. P. 71-75.

22. Antamoshkin A.N., Masich I.S. Heuristic search algorithms for monotone pseudo-boolean function conditional optimization. // Engineering & automation problems (Проблемы машиностроения и автоматизации). № 3. 2007. P. 41-45.

23. Antamoshkin A.N., Masich I.S. Identification of pseudo-Boolean function properties. // Engineering & automation problems (Проблемы машиностроения и автоматизации). 2/2007. P. 66-69.

24. Масич И.С., Шарыпова К.В. Оптимизация загрузки производственных мощностей литейного производства. // Системы управления и информационные технологии. №3(29). 2007. С. 76-80.

25. Масич И.С. Приближенные алгоритмы поиска граничных точек для задачи условной псевдобулевой оптимизации. // Вестник СибГАУ, 8470, 1(8). 2006. С. 39-43.

Б) **статьи в зарубежных изданиях**, включенных в международные базы цитирования, рекомендованные ВАК:

26. Kazakovtsev L.A., Masich I.S. A branch-and-bound algorithm for a pseudo-boolean optimization problem with black-box functions. // Facta Universitatis, Series Mathematics and Informatics. Vol. 33. No 2 (2018). P. 337–360. (WoS)

27. Kuzmich, R., Masich, I., Stupina, A., Kazakovtsev, L. Algorithmic procedure for constructing the truncated basic set of characteristics in the method of logical analysis of data // Proceedings of the 30th International Business Information Management Association Conference. IBIMA 2017. P. 5592-5597. (SCOPUS)

28. Masich I.S., Kazakovtsev L.A., Stupina A.A. Optimization Models for Detection of Patterns in Data. // Optimization Problems and their Applications. Proceedings of the School-Seminar on Optimization Problems and their Applications (OPTA-SCL 2018). Omsk, Russia, July 8-14, 2018. P. 264-275. (SCOPUS)

29. Kazakovtsev L.A., Orlov V.I., Stashkov D.V., Antamoshkin A.N., Masich I. S. Improved model for detection of homogeneous production batches of electronic components. // IOP Conference Series: Materials Science and Engineering 255. 2017. 012004. (SCOPUS)

30. Kraeva E.M., Masich I.S. Analysis and improvement of calculation procedure of high-speed centrifugal pumps. // IOP Conference Series: Materials Science and Engineering 255. 2017. 012005. (SCOPUS)

31. Antamoshkin A.N., Masich I.S. Combinatorial optimization in foundry practice. // IOP Conference Series: Materials Science and Engineering. 2016. С. 012001. (SCOPUS)

32. Antamoshkin A.N., Masich I.S. Selection of logical patterns for constructing a decision rule of recognition. // IOP Conference Series: Materials Science and Engineering. 2016. С. 012002. (SCOPUS)

33. Kraeva E.M., Masich I.S. Calculation and optimization of parameters in low-flow pumps. // IOP Conference Series: Materials Science and Engineering. 2016. С. 012019. (SCOPUS)

34. Antamoshkin A.N., Masich I.S., Kuzmich R.I. Heuristics and criteria for constructing logical patterns in data. // IOP Conference Series: Materials Science and Engineering. 2015. С. 012003. (SCOPUS)

35. Kazakovtsev L.A., Antamoshkin A.N., Masich I.S. Fast deterministic algorithm for eee components classification. // IOP Conference Series: Materials Science and Engineering. 2015. С. 012015. (SCOPUS)

36. Kazakovtsev L.A., Stupina A.A., Orlov V.I., Karaseva M.V., Masich I.S. Clustering Methods for Classification of Electronic Devices by Production Batched and Quality Classes. // FACTA UNIVERSITATIS. Ser. Math. Inform. Vol. 30. No 5. 2015. P. 567-581. (Math. Reviews)

37. Antamoshkin A., Masich I. Pseudo-Boolean Optimization in Case of an Unconnected Feasible Set. // Models and Algorithms for Global Optimization. Springer Optimization and Its Applications. Vol. 4. 2007. P. 111-122. (SCOPUS)

Монографии, препринты и главы в книгах:

1. Кузьмич Р.И., Масич И.С. Модификации метода логического анализа данных для задач классификации : монография – Красноярск: Сиб. федер. ун-т, 2018. – 180 с.
2. Орлов В.И., Федосов В.В., Казаковцев Л.А., Масич И.С., Проценко В.В., Сташков Д.В. Алгоритмическое обеспечение поддержки принятия решений по отбору изделий микроэлектроники для космического приборостроения : монография; СибГУ им. М. Ф. Решетнева. – Красноярск, 2017. – 228 с.
3. Казаковцев Л. А., Масич И. С., Орлов В. И., Проценко В. В., Федосов В. В. Разработка алгоритмического обеспечения анализа однородности партий электрорадиоизделий для комплектации РЭА КА : монография; Сиб. гос. аэрокосмич. ун-т. – Красноярск, 2016. – 192 с.
4. Антамошкин А.Н., Масич И.С., Кузьмич Р.И. Комбинаторная оптимизация при логической классификации : монография. – Красноярск: КрасГАУ. 2015. – 130 с.
5. Масич И.С. Поисковые алгоритмы условной псевдобулевой оптимизации : монография. – Красноярск: СибГАУ. 2013. – 160 с.
6. Масич И.С., Крушенко Г.Г. Алгоритмы случайного поиска в практических задачах условной оптимизации. Препринт № 1-12. – Красноярск: ИВМ СО РАН; СибГАУ. 2012. – 38 с.
7. Antamoshkin A., Masich I. Pseudo-Boolean Optimization in Case of an Unconnected Feasible Set, in: “Models and Algorithms for Global Optimization”. Springer Optimization and Its Applications, Vol. 4, 2007, p. 111-122. (SCOPUS) (глава в книге)

Программы для ЭВМ, зарегистрированные в Роспатенте:

1. Орлов В.И., Федосов В.В., Казаковцев Л.А., Масич И.С. Система интеллектуального анализа данных результатов тестовых испытаний электрорадиоизделий космического применения. Свидетельство о государственной регистрации программы для ЭВМ №2017617555 от 6.07.2017.
2. Антамошкин А.Н., Кузьмич Р.И., Масич И.С. Модифицированный метод логического анализа данных. Свидетельство о государственной регистрации программы для ЭВМ №2016619162 от 15.08.2016.
3. Казаковцев Л.А., Масич И.С., Орлов В.И., Федосов В.В. Программа автоматической группировки электрорадиоизделий. Свидетельство о государственной регистрации программы для ЭВМ №2016616924, 22.06.2016.
4. Казаковцев Л.А., Масич И.С., Орлов В.И., Федосов В.В. Система автоматизированного формирования и контроля специальных партий электрорадиоизделий. Свидетельство о государственной регистрации программы для ЭВМ №2016611353, 1.02.2016.
5. Масич И.С., Кузьмич Р.И., Краева Е.М. Логический анализ данных в задачах классификации. Свидетельство о государственной регистрации программы для ЭВМ № 2011612265, 2011.

6. Масич И.С., Краева Е.М. Алгоритмы условной псевдобулевой оптимизации для решения задач рюкзачного типа. Свидетельство о государственной регистрации программы для ЭВМ № 2011613446, 2011.