На правах рукописи

Соколова Юлия Сергеевна

МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА ДАННЫХ НА ОСНОВЕ ИНСТРУМЕНТАРИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ИНФОРМАЦИИ И БИОИНСПИРИРОВАННОГО МОДЕЛИРОВАНИЯ

Специальность: 05.13.01 – Системный анализ, управление и обработка информации (технические системы)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Рязань – 2018

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Рязанский государственный радиотехнический университет» (ФГБОУ ВО «РГРТУ») на кафедре «Вычислительная и прикладная математика».

Научный руководитель: **Демидова Лилия Анатольевна**, доктор технических наук, профессор, профессор кафедры вычислительной и прикладной математики ФГБОУ ВО «РГРТУ», г. Рязань.

Официальные оппоненты: **Дивеев Асхат Ибрагимович**, доктор технических наук, профессор, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук;

Ковшов Евгений Евгеньевич, доктор технических наук, профессор, профессор ФГБОУ ВО «Московский государственный технологический университет «СТАНКИН».

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

Защита диссертации состоится «*26*» *сентября* 20*18* г. в *12* часов на заседании диссертационного совета по защите докторских и кандидатских диссертаций Д 212.211.01 в ФГБОУ ВО «РГРТУ» по адресу: **390005, г. Рязань, ул. Гагарина, д. 59/1**.

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «РГРТУ» и на сайте http://rsreu.ru.

Автореферат разослан «___» _____ 20*18* г.

Ученый секретарь
диссертационного совета Д 212.211.01,
доктор технических наук,
доцент

Александр Викторович Пруцков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В настоящее время наиболее перспективные тенденции развития информационных технологий в сфере поддержки принятия решений связаны с разработкой программных средств интеллектуального анализа данных (ИАД, Data Mining), ориентированных на обработку многомерных сложноорганизованных структур данных больших объёмов, являющихся неотъемлемой частью экономики, банковской деятельности, производства, маркетинга, телекоммуникаций, веб-аналитики, медицины и пр. Стремительное увеличение количества информации, подлежащей обработке и анализу в различных прикладных областях, служит убедительным доводом в пользу выполнения разработок, направленных на создание математических, программных и аппаратных средств, предназначенных для быстрого и качественного решения задач обработки и анализа данных.

В ИАД особое место занимает проблема классификации, поскольку необходимость в проведении декомпозиции объектов встречается при решении широкого круга прикладных задач: при анализе кредитного риска, в медицинской диагностике, при распознавании рукописных символов (почерка), при категоризации текстов и т.п. Не менее актуально проведение качественной классификации данных в технических системах, например, при обработке изображений, полученных при дистанционном зондировании Земли, в т.ч. при обработке гиперспектральных изображений, при идентификации различных объектов (пешеходов, лиц и т.п.) на снимках, при осуществлении мероприятий неразрушающего контроля, при прогнозировании брака в зависимости от настроек оборудования и т.п. Следовательно, можно говорить об острой востребованности программно-математических средств принятия обоснованных и адекватных решений, позволяющих определить класс принадлежности объекта с максимальной точностью при приемлемых временных затратах.

На практике часто приходится проводить декомпозицию объектов в случае, когда об их внутренних связях ничего неизвестно и заранее неизвестна группировка объектов, на основе которой можно определить принципы для их разделения. В таком случае в качестве первой задачи анализа, требующей решения, можно рассматривать задачу кластеризации с целью выявления внутренней структуры в данных, на основании которой в дальнейшем можно будет формулировать более детальные задачи о поиске внутренних связей, влияющих на группировку объектов в исходном наборе.

Следует отметить, что не существует универсальных алгоритмов и методов классификации и кластерного анализа. Более того, применение различных инструментов моделирования к одному и тому же набору объектов может привести к различным результатам. Это связано с тем, что в основу этих инструментов заложены различные принципы моделирования, различаются и используемые в них метрики, функции близости, критерии оптимальности, алгоритмы оптимизации, способы выбора начальных приближений, способы работы с разнотипными характеристиками и т.п. В связи с этим возникает необходимость в получении результирующего классификационного решения (КР), объединяющего результаты разбиений, полученные при реализации нескольких декомпозиционных алгоритмов (ДКА), и совершающего меньшее число ошибок, чем каждый из этих алгоритмов.

При решении задач классификации и кластеризации характеристики объектов задаются, как правило, точными числовыми значениями. Однако иногда для значений характеристик объектов возможно определить лишь интервалы принадлежности. Следовательно, существует необходимость в разработке математического аппарата, позволяющего принимать адекватные и обоснованные КР с использованием данных, представленных, в том числе, в виде интервальных значений характеристик объектов.

В настоящее время разработаны десятки методов и алгоритмов решения задач классификации, успешно применяемых в разнообразных прикладных областях. Среди таких методов и алгоритмов наиболее известными являются: 1) байесовский классификатор; 2) искусственные нейронные сети; 3) алгоритмы ближайшего соседа и *k*-ближайших соседей (*k* Nearest Neighbor, *k*NN-алгоритм); 4) деревья решений; 5) алгоритм опорных векторов (Support Vector Machine, SVM-алгоритм) и др. Анализ алгоритмов классификации показывает, что зачастую они не обеспечивают получение приемлемых решений ввиду недостаточно обоснованного выбора значений параметров этих алгоритмов, а поиск эффективных решений приводит к значительным временным затратам из-за необходимости выполнения многократных реализаций используемых ДКА с целью выбора оптимальных значений их параметров. В связи с этим целесообразно исследовать перспективы использования биоинспирированных алгоритмов (БИА) стохастической оптимизации, в частности алгоритма роя частиц (Particle Swarm Optimization, PSO-алгоритма), в задаче разработки классификатора с целью выбора значений параметров, обеспечивающих высокое качество классификации данных. Процедура поиска наилучшего решения с помощью БИА имитирует некоторый природный процесс либо коллективное поведение определенных видов животных и растений с учетом их видовых особенностей. БИА постоянно демонстрируют свою эффективность и активно используются, поскольку позволяют находить решения таких задач оптимизации, для которых поиск решений традиционными численными методами оказывается неэффективным или вообще невозможным. БИА не гарантируют сходимости к глобальному решению задачи оптимизации, однако позволяют получить хорошее субоптимальное решение за приемлемое время.

В ряде случаев по результатам решения задачи декомпозиции объектов приходится выбирать из образовавшихся классов не все объекты, а только некоторые из них. Следовательно, возникает необходимость в их упорядочении (сортировке).

При наличии у объекта по каждой из характеристик нескольких значений, полученных, например, на основании показаний от нескольких приборов, сведений от нескольких источников информации или при групповом экспертном оценивании, целесообразно использовать подход к решению задач классификации и упорядочения объектов, основанный на применении инструментария теории мультимножеств, позволяющий учесть все (в т.ч. и противоречивые) значения (оценки).

С учетом вышесказанного можно сделать следующий вывод: от того, насколько качественно, адекватно и обоснованно выполнен анализ изучаемых объектов, в частности их классификация, упорядочение и последующий отбор, зависит эффективность принимаемых решений. Использование комплексного подхода к решению задач анализа данных с применением технологий интеллектуальной обработки информации и инструментария теории мультимножеств позволит создать качественно новые программные средства, обеспечивающие для задач классификации и упорядочения объектов повышение обоснованности и объективности принятия решений при приемлемых временных затратах, что является актуальной научно-технической задачей.

Степень разработанности темы исследования. Основополагающие научные разработки в области теории распознавания и классификации представлены трудами таких авторов, как: Дж. Нейман, Э. Пирсон, Р. Фишер, Ф. Розенблат, М.А. Айзерман, В.Н. Вапник, А.Я. Червоненкис, В.Д. Мазуров, А.Г. Ивахненко, Н.Г. Загоруйко, Ю.И. Журавлёв, Г.С. Лбов, К.В. Рудакова, В.В. Рязанов, О.В. Сенько и др.

Основные принципы современной теории кластеризации базируются на работах Дж. Рубина, С. МакНотона, Д. Дюрана, П. Оделла, Д.А. Вятченина, Т. Кохонена, Г. Болла, Д. Холла, Г. Ланса, У. Уильямса, Р. Дженсена, Г. Миллигана, Х. Фридмана, Н.Г. Загоруйко, Дж.К. Беждека, Дж.К. Данна, Р.Н. Дейва, Дж.М. Келлера, Я. Охаши и др.

Решению задач оптимизации, которые трудноразрешимы классическими методами, с использованием поисковых алгоритмов, реализующих эволюционные вычисления и имитирующие процессы, протекающие в биологических организмах, посвящены труды Дж.Г. Холланда, Н.А. Барричелили, А. Фрейзеры, Д. Рутковской, Л.А Гладкова, Л.Дж. Фогеля, А.П. Ротштейна, Е.С. Семёнкина, А.П. Карпенко и др.

В работах А.Б. Петровского рассматриваются подходы к решению проблемы принятия адекватных и обоснованных решений в задачах классификации, упорядочения и отбора объектов с использованием инструментария теории мультимножеств.

Несмотря на наличие значительного количества работ, посвященных вопросам классификации, кластеризации, упорядочивания и оптимизации, существует необходимость в разработке новых подходов, которые позволили бы существенно повысить обоснованность и адекватность принимаемых решений в сфере анализа данных.

Объект исследования – технологии обработки информации и ИАД.

Предмет исследования – методы и алгоритмы анализа данных на основе инструментария интеллектуальной обработки информации и биоинспирированного моделирования.

Цель работы – повышение обоснованности и адекватности принимаемых решений, в том числе и в условиях неопределённости, посредством разработки эффективных методов и алгоритмов классификации и упорядочения объектов, основанных на инструментарии интеллектуальной обработки информации и биоинспирированном моделировании, позволяющих устранить недостатки существующих аналогов.

Для реализации поставленной цели необходимо решить следующие **задачи**.

1. Произвести обзор основных подходов к проблемам классификации, кластеризации и упорядочивания объектов.
2. Выполнить анализ применимости SVM-алгоритма и алгоритмов кластеризации в условиях неопределённости к задаче классификации объектов, представленных числовыми значениями по характеристикам оценивания.
3. Выполнить анализ применимости SVM-алгоритма к задаче упорядочения объектов, представленных числовыми значениями по характеристикам оценивания.
4. Выполнить анализ применимости мультимножественного подхода к задаче классификации и упорядочения объектов, представленных интервальными значениями по характеристикам оценивания.
5. Разработать модификацию PSO-алгоритма, обеспечивающую одновременный подбор типа функции ядра, значений её параметров и значения параметра регуляризации при приемлемых временных затратах на разработку SVM-классификатора, характеризующегося высоким качеством классификации объектов, представленных числовыми значениями по характеристикам оценивания.
6. Повысить обоснованность принимаемых КР посредством интеграции частных SVM-классификаторов в ансамбль.
7. Разработать новый метод классификации объектов в случае отсутствия информации об их возможной классовой принадлежности, реализующий совместное использование SVM-алгоритма и алгоритмов кластеризации, в рамках которого предложить и исследовать варианты формирования набора эталонных объектов, которые могут быть использованы для получения обучающей и тестовой выборок данных при разработке SVM-классификатора.
8. Выполнить классификацию и упорядочение объектов, представленных интервальными значениями по характеристикам оценивания, на основе инструментария теории мультимножеств с применением лингвистической шкалы.
9. Разработать программное обеспечение (ПО), реализующее ИАД с применением предложенных методов и алгоритмов.

Методы исследования. В работе использовались методы математической статистики, эволюционных вычислений, оптимизации, системного анализа, теории нечётких множеств, теории мультимножеств, интеллектуальной обработки информации, модульного и объектно-ориентированного программирования.

Соответствие паспорту специальности. Содержание диссертации соответствует п.4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации» и п.5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации» паспорта научной специальности 05.13.01 «Системный анализ, управление и обработка информации (технические системы)».

Научная новизна заключается в следующих научно обоснованных результатах, полученных в ходе диссертационного исследования.

1. Разработан модифицированный PSO-алгоритм, который, в отличие от классического PSO-алгоритма, осуществляет одновременный поиск субоптимальных значений типа функции ядра, значений параметров функции ядра и значения параметра регуляризации. Использование такого подхода к поиску субоптимальных значений параметров SVM-классификатора позволяет в 2–3 раза сократить временные затраты на его разработку и обеспечить при этом высокое качество классификации (с точностью до 98%) и упорядочения объектов.

2. Предложена модификация метода формирования ансамбля SVM-классификаторов, базирующегося на применении алгоритма максимальной декорреляции (АМД), обеспечивающая разнообразие частных классификаторов в ансамбле и, как результат, более качественное КР, точность которого на 4–5% превосходит точность лучшего SVM-классификатора, используемого в ансамбле, в соответствии с выбранной стратегией принятия решений. В отличие от традиционных методов формирования ансамблей в предлагаемом методе реализован новый подход как к интеграции результатов частных классификаций, так и к способу определения результирующего КР.

3. Разработан алгоритм формирования двухуровневого классификатора сложноорганизованных многомерных данных больших объёмов на основе совместного использования SVM- и PSO-алгоритмов, позволяющий сократить объём данных за счёт включения в набор, используемый для создания обучающей и тестовой выборок, наиболее значимых для построения SVM-классификатора объектов, представленных опорными векторами, что позволяет в 2–3 раза сократить временные затраты на разработку SVM-классификатора и обеспечить при этом высокое качество классификации объектов.

4. Разработан двухэтапный метод классификации объектов, основанный на совместном использовании SVM- и *k*NN-алгоритмов и обеспечивающий повышение в среднем на 3% точность классификации данных посредством уменьшения числа ошибочно классифицированных объектов внутри полосы, разделяющей классы.

5. Разработан метод классификации объектов в случае отсутствия информации об их возможной классовой принадлежности, основанный на совместном использовании SVM-алгоритма и алгоритмов кластеризации в условиях неопределённости. Метод может использоваться для уточнения результатов кластеризации, полученных с применением алгоритмов кластеризации в условиях неопределённости, а также с целью дальнейшей классификации новых данных. В рамках метода разработаны и исследованы на эффективность варианты формирования набора эталонных объектов, которые могут быть использованы для получения обучающей и тестовой выборок данных при построении SVM-классификатора.

6. Предложен подход к принятию решений по классификации и упорядочению объектов, реализующий представление неточных знаний на основе лингвистических переменных и позволяющий рассмотреть различные стратегии формирования обобщающих решающих правил классификации (ОРПК) и упорядочения с применением инструментария теории мультимножеств.

Теоретическая и практическая значимость работы. Основные положения диссертации вносят вклад в развитие средств анализа данных: предложены и исследованы новые методы и алгоритмы анализа данных на основе инструментария интеллектуальной обработки информации и биоинспирированного моделирования, в том числе и в условиях неопределённости. Выводы и результаты исследования ориентированы на практическое применение предложенных методов и алгоритмов в системах ИАД. Самостоятельное практическое значение имеет разработанное ПО, реализующее предложенные методы и алгоритмы и позволяющее использовать их при решении практических задач ИАД.

Основные положения, выносимые на защиту.

1. Модифицированный PSO-алгоритм, используемый при определении субоптимальных значений параметров SVM-классификатора, позволяющий существенно сократить временные затраты на разработку SVM-классификатора и обеспечивающий высокую точность классификации и упорядочения объектов.

2. Модифицированный метод формирования ансамбля SVM-классификаторов, базирующийся на применении АМД, реализующий интеграцию частных SVM-классификаторов в ансамбль и обеспечивающий более качественное КР в соответствии с выбранной «лучшей» стратегией принятия решений.

3. Алгоритм формирования двухуровневого классификатора сложноорганизованных многомерных данных больших объёмов, позволяющий сократить объём данных, используемых для построения классификатора, и обеспечивающий высокую точность классификации объектов и упорядочения при приемлемых временных затратах.

4. Двухэтапный метод классификации, основанный на совместном использовании SVM- и *k*NN-алгоритмов, обеспечивающий повышение точности КР за счёт уменьшения числа ошибок внутри полосы, разделяющей классы.

5. Метод классификации объектов, основанный на совместном использовании SVM-алгоритма и алгоритмов кластеризации в условиях неопределённости, позволяющий проводить классификацию и упорядочение в случае отсутствия информации о возможной классовой принадлежности объектов, а также варианты формирования набора эталонных объектов, которые могут быть использованы для получения обучающей и тестовой выборок данных при построении SVM-классификатора.

6. Подход к принятию решений по классификации и упорядочению объектов, характеристики которых представлены интервальными значениями оценок, позволяющий рассмотреть различные стратегии формирования ОРПК и упорядочения с применением инструментария теории мультимножеств.

7. Программное обеспечение, реализующее бинарную классификацию и упорядочение объектов с применением предложенных методов и алгоритмов ИАД и подтверждающее их адекватность и работоспособность.

Степень достоверности и апробация результатов. Достоверность научных результатов подтверждена непротиворечивостью получаемых экспериментальных результатов, их соответствием эталонным данным, результатами сравнительного анализа с реализациями подобных методов и алгоритмов в известных программах ИАД, положительным рецензированием предлагаемых методов и алгоритмов квалифицированными специалистами, свидетельствами о регистрации программ для ЭВМ, наличием актов внедрения исследований в организациях.

Основные положения и результаты диссертационной работы докладывались и обсуждались на 6 всероссийских и 12 международных конференциях: XIX, XX и XXI Всероссийских научно-технических конференциях студентов, молодых учёных и специалистов «Новые информационные технологии в научных исследованиях (НИТ-2014, НИТ-2015, НИТ-2016)» (г. Рязань, 2014, 2015, 2016); XI Международной научно-технической конференции «Искусственный интеллект в XXI веке» (г. Пенза, 2014); 16th International Symposium on Advanced Intelligent Systems (ISIS2015) (Mokpo, South Korea, 2015); XXXIV Международной научно-технической конференции «Математические методы и информационные технологии в экономике, социологии и образовании» (г. Пенза, 2015); III Международной конференции «Устойчивость и процессы управления» (г. Санкт-Петербург, 2015); V Всероссийской конференции студентов и молодых учёных «Молодёжная наука в развитии регионов» (г. Пермь, 2015); XX-th International Open Science Conference «Modern Informatization Problems in Economics and Safety» (Yelm, WA, USA, 2015); 18-й Международной научно-технической конференции «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций» (г. Рязань, 2015); Всероссийской научно-технической конференции «Интеллектуальные и информационные системы» (г. Тула, 2015); XXI-th International Open Science Conference «Modern Informatization Problems in the Technological and Telecommunication Systems Analysis and Synthesis» (Yelm, WA, USA, 2016); Международной научно-технической и научно-методической конференции «Современные технологии в науке и образовании (СТНО-2016, СТНО-2017)» (г. Рязань, 2016, 2017); Международной научно-практической конференции «Математика: фундаментальные и прикладные исследования и вопросы образования» (г. Рязань, 2016); XII Международном симпозиуме «Интеллектуальные системы» (г. Москва, 2016).

Результаты диссертационной работы внедрены и используются в: 1) АО «РКЦ «Прогресс» (филиал ОКБ «Спектр») при обработке гиперспектральной информации, получаемой от функционирующих космических аппаратов серии «Ресурс-П», с целью проведения идентификации объектов земной поверхности; 2) Эксперт-Центре АО «НИКИМТ-Атомстрой» при реализации программного модуля классификации при создании автоматизированных систем неразрушающего контроля изделий; 3) ООО «Независимый центр оценки и экспертиз» при оценке инвестиционной привлекательности объектов коммерческой недвижимости; 4) Рязанской торгово-промышленной палате при выборе наиболее эффективных инвестиционных проектов среди предприятий Рязанской области; 5) ПАО «Прио-Внешторгбанк» для оценки кредитоспособности заемщиков; 6) учебном процессе ФГБОУ ВО «РГРТУ».

Публикации. По теме диссертации опубликовано 44 работы, в том числе: 8 статей в ведущих рецензируемых научных журналах из Перечня ВАК РФ, 6 работ в базе данных Scopus, 4 работы в базе данных Web of Science, 8 статей в межвузовских сборниках научных трудов, 18 тезисов докладов на международных и всероссийских научно-технических конференциях; получено 5 свидетельств о государственной регистрации программ для ЭВМ.

Личный вклад соискателя в получение результатов, изложенных в диссертации. Автором сформулированы основные идеи защищаемых методов и алгоритмов. Им же было разработано и зарегистрировано ПО, реализующие предложенные методы и алгоритмы. Работы, выполненные в соавторстве, подчинены общей постановке научной проблемы и предложенной автором концепции её решения.

Структура и объём работы. Диссертационная работа состоит из введения, четырех глав с выводами, заключения, списка использованных источников и приложений. Общий объём работы 294 страницы, включая 26 таблиц, 71 рисунок и библиографический список из 202 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении аргументирована актуальность исследования, сформулированы цель и задачи диссертационной работы, показана научная новизна полученных результатов и область их практического применения, приведены основные положения, выносимые на защиту.

В первой главе выполнена постановка задач классификации, кластеризации и упорядочения объектов, произведен обзор основных ДКА; установлено, что SVM-алгоритм является одним из наиболее эффективных инструментов проведения классификации, а кластеризация объектов с использованием алгоритмов кластеризации в условиях неопределённости является более «естественной», чем чёткая кластеризация, поскольку позволяет одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью принадлежности или/и типичности кластерам. Показано, что для корректной декомпозиции необходима предварительная нормировка значений характеристик объекта, рассмотрены различные способы нормировки, из которых был выбран способ, преобразующий числовое значение характеристики в отрезок [0; 1].

Исследованы подходы к определению качества принимаемого КР с целью выявления их недостатков и достоинств. Рассмотрены показатели качества КР, одним из которых является показатель точности, равный отношению числа верно классифицированных объектов к общему числу объектов. Показано, что не существует универсального показателя качества классификации, поскольку выбор лучшего классификатора производится исходя из требований к задаче: для каждого случая индивидуально выбирается соотношение между уровнем числа ошибок первого или второго рода.

Выполнено исследование SVM-алгоритма, в ходе которого определена его высокая способность к обобщению (принятию КР). При разработке SVM-классификатора исходные данные (учебный набор) разбивают случайным образом на обучающую и тестовую (контрольную) выборки. Обучающая выборка используется для обучения классификатора, тестовая выборка, составляющая от 1/10 до 1/3 от числа объектов учебного набора, не используется при обучении классификатора на основе SVM-алгоритма, а применяется для проверки точности классификатора.

Цель SVM-алгоритма, относящегося к инструментам машинного обучения (МО), – найти гиперплоскость, разделяющую классы, максимально удалённую от объектов обучающей выборки. Поиск разделяющей гиперплоскости сводится к задаче квадратичной оптимизации, которая эквивалентна двойственной задаче поиска седловой точки функции Лагранжа. В случае разработки SVM-классификатора на основе SVM-алгоритма определяется классифицирующая функция

$$f(z) = \sum_{i=1}^{S} \lambda_i y_i \kappa(z_i, z) + b,$$

на основе которой принимается КР, сопоставляющее объект z классу с меткой «−1» или «+1», в соответствии с правилом

$$A(z) = sign(f(z)) = sign\Big(\sum_{i=1}^{S} \lambda_i y_i \kappa(z_i, z) + b\Big),$$

где $\kappa(z_i, z)$ − функция ядра (ФЯ) классификатора; b − значение параметра, задающего смещение гиперплоскости, разделяющей классы, относительно начала координат; λ_i − множитель Лагранжа, $\lambda_i \geq 0$; y_i − известное КР («−1» или «+1»), S − число объектов в обучающей выборке. Условие $f(z) = 0$ определяет разделяющую гиперперплоскость.

Показано, что модуль значения классифицирующей функции $f(z)$ − $|f(z)|$ − позволяет оценить удалённость объекта z от гиперплоскости, разделяющей классы, и может быть использован для упорядочения объектов внутри классов с метками «−1» и «+1».

Установлено, что применение SVM-классификатора ограничено в связи с трудностями определения оптимальных значений его параметров, к которым относятся параметр регуляризации, тип ФЯ и параметры функции ядра. Показано, что для получения адекватных результатов классификации с использованием SVM-классификатора необходимо его многократная разработка при различных составах обучающей и тестовой выборок для принятия окончательного решения о результатах классификации.

Исследованы показатели качества чёткой кластеризации и кластеризации в условиях неопределённости с целью выявления их недостатков и достоинств. Указано, что при кластеризации в условиях неопределённости в качестве показателя качества кластеризации целесообразно использовать индекс Се – Бени и его модификации, обеспечивающие получение адекватных результатов кластеризации и характеризующиеся невысокой вычислительной сложностью, а при декомпозиции, проводимой с использованием чётких алгоритмов кластеризации, – индекс оценки силуэта.

Определена целесообразность использования ансамблей алгоритмов для получения более точных КР и решений по упорядочению объектов.

Показано, что, в случае наличия у объектов нескольких оценок по каждой из его характеристик, в качестве математической модели их представления для определения принципов классификации и упорядочения следует использовать мультимножества, позволяющее учесть все, в том числе несовпадающие и противоречивые, оценки объектов по характеристикам оценивания.

Во второй главе представлена разработка методов и алгоритмов, направленных на повышение качества анализа данных с использованием SVM-алгоритма.

Модифицированный PSO-алгоритм используется при определении субоптимальных значений параметров SVM-классификатора, при которых достигается максимально высокое качество классификации, определяемое максимальной точностью классификации на обучающей и тестовой выборках при минимальном числе опорных векторов. В этом алгоритме каждой i-й частице роя ($i = \overline{1,m}$) соответствует набор значений параметров, описывающих её позицию в пространстве поиска: (T_i, x_i^1, x_i^2, C_i), где T_i – номер типа ФЯ: $T_i = 1$ – для полиномиальной однородной, $T_i = 2$ – полиномиальной неоднородной, $T_i = 3$ – радиальной базисной, $T_i = 4$ – радиальной базисной функции Гаусса, $T_i = 5$ – сигмоидной ФЯ; x_i^1, x_i^2 – значения параметров ФЯ i-й частицы; C_i – значение параметра регуляризации. В отличие от традиционного PSO-алгоритма, где позиция частицы в пространстве поиска определяется набором значений (x_i^1, x_i^2, C_i), в модифицированном алгоритме дополнительно вводится параметр T_i, определяющий тип функции ядра.

В модифицированном PSO-алгоритме возможно «перерождение» частицы: из тех частиц, у которых значение параметра $T_i \neq \widetilde{T}$ ($i = \overline{1,m}$), где значение $\widetilde{T}$ соответствует типу ФЯ, при котором частицы показывают максимально высокое качество классификации, выбирают p% частиц, показавших самое низкое качество классификации, и изменяют значение параметра T_i этих частиц на значение $\widetilde{T}$. При этом также изменяют значения параметров x_i^1, x_i^2, C_i «перерождаемой» частицы так, чтобы они соответствовали новому типу ФЯ $\widetilde{T}$ (то есть попадали в соответствующие диапазоны).

В результате выполнения модифицированного PSO-алгоритма определяется частица с субоптимальным набором значений параметров $(\widetilde{T}, \widetilde{x}^1, \widetilde{x}^2, \widetilde{C})$, обеспечивающая высшее качество классификации на включенных в поиск типах ФЯ.

В модифицированном методе формирования ансамбля SVM-классификаторов, базирующегося на применении АМД, предлагается в качестве порогового значения

выбирать то, при котором все пять стратегий классификации (стратегия максимума, минимума, медианы, суммы или произведения) показывают стабильное улучшение качества классификации.

Алгоритм формирования двухуровневого SVM-классификатора может быть представлен следующей последовательностью шагов.

Шаг 1. Обучить на учебном наборе данных t частных SVM-классификаторов с использованием различных обучающих выборок данных TR_1, …, TR_t с целью получения t частных SVM-классификаторов, составляющих группу. При обучении частных SVM-классификаторов следует использовать различные типы ФЯ, различные значения параметров ФЯ и различные значения параметра регуляризации.

Шаг 2. Получить по результатам обучения всех частных классификаторов наборы опорных векторов SV_1, …, SV_t, которые расположены на границах разделяющей гиперплоскости и несут всю информацию о разделении классов, в результате объединения которых сформировать набор SV из ℓ объектов ($\ell \le s$, где s – число объектов учебного набора).

Шаг 3. Выделить из набора SV поднабор SV^+, состоящий из L объектов, ($L \le \ell$), которые были определены теми или иными частными классификаторами как опорные векторы и реальный класс которых совпал с классом, в который этот объект был отнесен тем или иным частным классификатором. Это необходимо для того, чтобы заведомо ложные данные не участвовали в обучении итогового SVM-классификатора. Остальные объекты набора SV составят поднабор SV^- из $\ell - L$ объектов. Поднабор SV^+ будет использоваться для обучения, а поднабор SV^- – для тестирования итогового SVM-классификатора.

Шаг 4. Определить с помощью PSO-алгоритма субоптимальные значения параметров SVM-классификатора, используя в качестве обучающей выборки поднабор SV^+, а в качестве тестовой – поднабор SV^-.

Шаг 5. Построить итоговый SVM-классификатор на основе типа ФЯ, значений его параметров и значения параметра регуляризации, найденных PSO-алгоритмом.

Шаг 6. Выполнить доклассификацию объектов учебного набора данных, не вошедших в поднаборы SV^+ и SV^-.

Шаг 7. Оценить точность разработанного двухуровневого классификатора, а также время, затраченное на его разработку.

Двухэтапный метод классификации, основанный на совместном использовании SVM- и *k*NN-классификаторов, может быть представлен следующей последовательностью действий.

Этап 1. Разработка SVM-классификатора с определением возможности разработки *k*NN-классификатора.

1.1. Разработка SVM-классификаторов с последующим выбором лучшего из них в смысле обеспечения наиболее высокой точности классификации на учебном наборе данных U осуществляется на основе сформированных случайным образом обучающей и тестовой выборках. Заданные при разработке SVM-классификатора тип функции ядра, значения параметров ядра, значение параметра регуляризации определяют гиперплоскость, разделяющую объекты на два класса с метками «–1» и «+1». Оценка качества классификации объектов осуществляется с применением различных показателей качества классификации.

При разработке SVM-классификатора используется учебный набор данных $U = \{< z_1, y_1 >,...,< z_s, y_s >\}$, в котором каждый кортеж $< z_i, y_i >$ содержит информацию об объекте $z_i \in Z$ и число $y_i \in Y = \{-1;+1\}$, определяющее метку класса, к кото-

рому принадлежит объект z_i. Набор объектов Z представляет собой объединение набора объектов Z^-, метка класса которых принимает значение «−1», и набора объектов Z^+, метка класса которых принимает значение «+1», т.е. $Z = Z^- \cup Z^+$. Объект $z_i \in Z$ представлен q-мерным вектором числовых характеристик $z_i = (z_i^1, z_i^2, \ldots, z_i^q)$ (нормированных значениями из отрезка [0; 1]), где z_i^l – числовое значение l-й характеристики i-го объекта ($i = \overline{1,s}$, $l = \overline{1,q}$).

1.2. Определение областей Ω^- и Ω^+, содержащих все ошибочно классифицированные объекты, оказавшиеся в наборах Z^- и Z^+ соответственно; d_{Ω^-} – ширины области Ω^-; d_{Ω^+} – ширины области Ω^+; N_{Ω^-} – числа объектов в области Ω^-; N_{Ω^+} – числа объектов в области Ω^+. Формирование на основе областей Ω^- и Ω^+ итоговой Ω-области, включающей в себя все ошибочно классифицированные объекты, образующие вместе с правильно классифицированными объектами, попавшими в Ω-область, и соответствующими метками классов объектов из Ω-области, набор данных $G = \{< z_1, y_1 >, \ldots, < z_{N_\Omega}, y_{N_\Omega} >\}$, в котором каждый кортеж $< z_i, y_i >$ содержит информацию об объекте z_i из Ω-области и соответствующую z_i метку класса $y_i \in Y = \{-1; +1\}$.

Возможны два варианта формирования Ω-области, в результате реализации которых будут получены: 1) *асимметричная* относительно разделяющей гиперплоскости Ω-область $\Omega = \Omega^- \cup \Omega^+$; 2) *симметричная* относительно разделяющей гиперплоскости Ω-область, содержащая объекты, находящиеся относительно разделяющей гиперплоскости на расстоянии, не превышающем $\Delta = \max\{d_{\Omega^-}, d_{\Omega^+}\}$.

1.3. Формирование набора данных $W = U \setminus G$ удалением из учебного набора данных U кортежей набора данных G. Набор W будет состоять из кортежей набора U, классовая принадлежность объектов для которых SVM-классификатором была определена правильно. Объекты этого набора будут использованы для разработки *k*NN-классификатора. Поскольку в Ω-области кроме ошибочно классифицированных объектов может находиться и некоторое число объектов, классифицированных правильно, то возможно, что число кортежей в наборе W окажется существенно меньше, чем в U, и их будет недостаточно для последующей разработки *k*NN-классификатора.

Этап 2. Разработка *k*NN-классификатора.

2.1. Разработка на основе набора данных $W = U \setminus G$ *k*NN-классификаторов, устанавливающих классовую принадлежность всех объектов Ω-области при различных значениях числа k ближайших соседей из набора W с использованием различных способов голосования и различных способов оценки близости между объектами. Выбор лучшего *k*NN-классификатора в смысле обеспечения наиболее высокой точности классификации всех объектов Ω-области и фиксация значений параметров лучшего *k*NN-классификатора: варианта Ω-области, используемого способа оценки близости между объектами, способа голосования и оптимального числа соседей.

2.2. Сравнение качества итоговой классификации объектов с применением лучшего *k*NN-классификатора с качеством классификации, полученным после разработки SVM-классификатора, с целью выявлении целесообразности применения сформированного таким образом классификатора для определения классовой принадлежности новых объектов, при этом целесообразность применения определяется улучшением показателей качества классификации объектов из набора Z.

Алгоритм решения задачи упорядочения объектов с применением результатов работы SVM-классификатора может быть представлен следующей последовательностью шагов.

Шаг 1. Для заданного набора объектов (прецедентов) построить SVM-классификатор, субоптимальные значения параметров которого определить, например, с использованием традиционного или модифицированного PSO-алгоритма.

Шаг 2. Для объектов упорядочиваемого класса (класса с меткой «–1» или «+1») рассчитать значения функции $f(z)$, возвращаемые SVM-классификатором.

Шаг 3. Отсортировать объекты упорядочиваемого класса по убыванию абсолютных значений $f(z)$.

Успешность применения SVM-алгоритма при решении классификационных задач определяется качеством используемых обучающей и тестовой выборок данных. Однако зачастую приходится решать вопросы классификации объектов в условиях отсутствия какой-либо априорной информации о возможной классовой принадлежности хотя бы части объектов, наличие которой позволило бы сформировать обучающую и тестовую выборки для разработки SVM-классификатора.

При отсутствии информации о классовой принадлежности классифицируемых объектов предлагается метод, основанный на применении алгоритмов кластеризации в условиях неопределённости (алгоритмов нечёткой, возможностной и возможностно-нечёткой кластеризации – FCM-, PCM- и PFCM-алгоритмов), использование которых позволит снять неопределённость, касающуюся классового разделения объектов, и сформировать обучающую и тестовую выборки данных для разработки SVM-классификатора. Таким образом, использование этого метода позволит, во-первых, решить проблему формирования обучающей и тестовой выборки, во-вторых, уточнить результаты кластеризации.

Для решения проблемы формирования обучающей и тестовой выборок данных при применении SVM-алгоритма для разработки SVM-классификатора предлагается использовать результаты кластеризации объектов, полученные с применением одного или нескольких алгоритмов кластеризации.

Для решения проблемы уточнения результатов кластеризации объектов, полученных с применением одного или нескольких алгоритмов кластеризации, предлагается использовать SVM-классификатор, разработанный с применением SVM-алгоритма.

При решении классификационных задач в случае отсутствия информации о возможной классовой принадлежности объектов особое внимание уделено разработке и исследованию вариантов формирования набора эталонных объектов, который в дальнейшем может быть использован для создания обучающей и тестовой выборок, применяемых при разработке SVM-классификатора. По результатам исследований был сделан вывод о том, что наиболее перспективным является вариант формирования набора эталонных объектов из объектов, результирующая принадлежность к кластерам для которых определяется с применением кластерного ансамбля на основе матрицы подобия векторов меток кластеров и алгоритма спектральной факторизации, поскольку его использование обеспечивает создание более качественного (с точки зрения повышения точности КР с применением SVM-классификатора) набора эталонных объектов.

В третьей главе представлен математический аппарат, применяемый для проведения классификации и упорядочения объектов на основе инструментария теории мультимножеств, даже в случае, когда их характеристики определены с использованием лингвистической шкалы, позволяющей реализовать принципы описания и обработки неточных данных на основе лингвистических переменных. Использование инструментария мультимножеств позволит учесть все значения оценок объекта по его характеристикам, что востребовано, например, в задачах технической диагностики, когда значения параметров системы оцениваются сразу несколькими приборами, которые показывают несовпадающие значения измеряемых параметров, а также в зада-

чах, в которых анализируются данные от нескольких источников информации или результаты группового экспертного оценивания.

В случае применения лингвистического подхода к формированию ОРПК каждому объекту по каждой характеристике будет выставляться не чёткая числовая оценка, а некоторая интервальная оценка вида $[\alpha,\beta]$. При этом левая граница интервала $[\alpha,\beta]$ будет соответствовать чисто консервативной стратегии оценивания, правая – чисто рискованной, а середина этого интервала – нейтральной стратегии. Если сопоставить некоторой стратегии оценивания объектов показатель δ ($\delta \geq 0$), то оценка, соответствующая этой стратегии, может быть вычислена как $(\beta + \delta \cdot \alpha)/(\delta + 1)$.

Использование показателя δ позволит выполнить разработку ОРПК при различных стратегиях оценивания объектов на основе мультимножеств с чёткими числовыми оценками объектов (из интервала вида $[\alpha,\beta]$) по характеристикам оценивания. Тогда каждому объекту z_i в соответствие может быть поставлено мультимножество вида

$$Z_i = \{k_{Z_i}(p_1^1)\cdot p_1^1,...,k_{Z_i}(p_1^{u_1})\cdot p_1^{u_1},...,k_{Z_i}(p_q^1)\cdot p_q^1,..,k_{Z_i}(p_q^{u_q})\cdot p_q^{u_q},k_{Z_i}(w_1)\cdot w_1,k_{Z_i}(w_2)\cdot w_2\},$$

где $k_{Z_i}(p_j^{l_j})$ и $k_{Z_i}(w_c)$ – число приборов (источников информации, экспертов), сопоставивших объекту z_i значения $p_j^{l_j}$ и w_c соответственно; символ «·» обозначает взаимосвязь между $k_{Z_i}(p_j^{l_j})$ и $p_j^{l_j}$, а также между $k_{Z_i}(w_c)$ и w_c ($i=\overline{1,s}$; $j=\overline{1,q}$; $c=\overline{1,2}$; $l_j=\overline{1,u_j}$); $p_j^{l_j}$ – l_j-е значение j-й характеристики; w_c – класс объекта, равный c.

Для мультимножеств, представляющих объекты z_i ($i=\overline{1,s}$), формируются ОРПК, относящие объекты к заданным классам наилучшим образом в смысле близости к предварительным индивидуальным сортировкам.

Инструментарий теории мультимножеств позволяет решить задачу упорядочения объектов в случае, когда их характеристики определены с использованием лингвистической шкалы. Объекты могут быть упорядочены по удалённости от «антиидеального» (наихудшего) объекта $Z_{min} = \{m\cdot p_1^1,0,..,0,m\cdot p_2^1,0,..,0,...,m\cdot p_q^1,0,...,0\}$ или по близости к $Z_{max} = \{0,..,0,m\cdot p_1^{u_1},0,..,0,m\cdot p_2^{u_2},...,0,...,0,m\cdot p_q^{u_q}\}$ – «идеальному» (наилучшему) объекту, которым все m приборов (источников информации, экспертов) дали соответственно низшие и высшие оценки по всем характеристикам оценивания.

В четвертой главе приведены результаты исследований предлагаемых методов и алгоритмов, способствующих повышению качества принимаемых КР и решений по внутриклассовому упорядочению объектов, на реальных наборах данных и наборах, традиционно используемых для тестирования методов и алгоритмов МО.

Реальные наборы данных были сформированы на основе результатов гиперспектральной съемки от космического аппарата «Ресурс-П» № 1: была создана база спектральных эталонов, содержащая 220 гиперспектральных характеристик (отображающих зависимости между длиной волны и значением коэффициента спектрального отражения) объектов природного и искусственного происхождения. Из этой базы эталонов были созданы наборы данных AqwaObj и AntroObj (таблица 1), которые использовались для идентификации объектов водного и антропогенного происхождения.

В качестве тестовых наборов были использованы данные проекта Statlog и репозитория задач МО. В частности, рассматривались наборы данных для обработки информации в системах передачи данных (Spam), классификации высокочастотных радарных сигналов, возвращаемых из ионосферы (Ionosphere) и решения задач классификации (диагностики) в различных предметных областях (таблица 1).

В результате выполнения процедуры поиска с помощью традиционного и модифицированного PSO-алгоритмов на экспериментальных наборах удалось найти значе-

ния параметров, улучшающих качество SVM-классификатора, разработанного с использованием значений параметров, установленных по умолчанию. Оба алгоритма определили в качестве оптимальных одинаковый тип ФЯ, близкие значения параметров ФЯ и параметра регуляризации, а также близкие значения точности обучения и тестирования SVM-классификатора, построенного с учетом найденных значений параметров. По времени поиска модифицированный PSO-алгоритм оказался более эффективным – затраченное им время на поиск искомого решения оказалось меньше (в 2–3 раза), чем время поиска с применением традиционного PSO-алгоритма. При этом поиск производился в одинаковых диапазонах изменения значений параметров SVM-классификатора и при одинаковых значениях параметров PSO-алгоритма.

В таблице 1 для ряда экспериментальных наборов приведены лучшие значения показателя общей точности ($Accur$), числа ошибок I (Er_I) и II рода (Er_{II}) для SVM-классификаторов, разработанных в авторском ПО с использованием: 1) значений параметров, заданных по умолчанию (SVM_{def}); 2) значений параметров, определенных в качестве оптимальных с применением традиционного или модифицированного PSO-алгоритмов (SVM+PSO); 3) ансамбля SVM-классификаторов (SVM-ансамбль); 4) двухэтапного метода классификации (SVM+kNN). Ячейки таблицы 1, соответствующие лучшему варианту классификатора, обеспечивающему максимальное значение показателя общей точности и минимальное число ошибок классификации, выделены жирным шрифтом.

Таблица 1. Результаты SVM-классификации в авторском ПО

Набор данных	$s \times q$	SVM_{def}		SVM+PSO		SVM-ансамбль		SVM+kNN	
		$Accur$,%	Er_I+Er_{II}	$Accur$,%	Er_I+Er_{II}	$Accur$,%	Er_I+Er_{II}	$Accur$,%	Er_I+Er_{II}
AqwaObj	220×127	98,64	1+2	**99,55**	**1+0**	–	–	99,09	1+1
AntroObj	220×127	90,45	5+16	97,27	6+0	**98,64**	**2+1**	93,63	7+7
Spam	4601×57	95,94	176+11	98,07	53+36	**98,59**	**64+1**	–	–
Ionosphere	351×34	97,72	7+1	**99,43**	**1+1**	–	–	**99,43**	**1+1**
Heart	270×13	95,93	5+6	97,78	3+3	–	–	**98,15**	**2+3**
WDBC	569×30	99,30	3+1	99,82	1+0	–	–	**100**	**0+0**
Firms	60×12	95,00	0+3	**100**	**0+0**	–	–	**100**	**0+0**
Australian	690×14	95,36	19+13	**96,81**	**6+16**	–	–	–	–
German	1000×24	93,40	4+62	95,10	21+28	**98,70**	**12+1**	95,90	12+29
МОТП12	400×2	91,50	27+7	**97,25**	**3+8**	–	–	–	–
Среднее		*95,32*	–	*98,11*	–	*98,64*	–	*98,03*	–

Показано, что SVM-классификаторы, разработанные со значениями параметров ФЯ и значением параметра регуляризации, установленными по умолчанию, по точности уступают SVM-классификаторам, при разработке которых использовались значения параметров, определенные процедурой поиска с помощью алгоритма роя частиц. Использование PSO-алгоритма при разработке SVM-классификатора для различных наборов данных позволяет повысить точность классификации в среднем до 98 % (при этом увеличение точности в среднем составляет 3%). Увеличение точности классификации более чем на 3% наблюдается также в случае использования ансамблей классификаторов, отобранных по принципу АМД, при этом точность этого решения на 4–5% выше точности лучшего SVM-классификатора, используемого в ансамбле. Использование предлагаемого двухэтапного метода классификации, основанного на совместном использовании SVM- и kNN-алгоритмов, позволяет повысить точность SVM-классификатора, разработанного с использованием радиальной базисной ФЯ при установленных по умолчанию значениях параметров, в среднем на 3%.

Выполнен сравнительный анализ результатов работы SVM-классификаторов, разработанных с применением авторского ПО, с результатами классификаций, представленными в литературе по ИАД и полученными в пакетах STATISTICA StatSoft и IBM SPSS Modeler, позволяющих производить классификацию данных с использованием SVM-алгоритма. Установлено, что SVM-классификаторы, разработанные с применением авторского ПО, по точности не уступают, а чаще превосходят классификаторы, разработанные в указанных пакетах (таблица 2).

Таблица 2. Сравнительный анализ показателей общей точности SVM-классификаторов

Средство построения SVM-классификатора	**Значение показателя общей точности на наборах данных, %**										
	AqwaObj	**AntroObj**	**Spam**	**Ionosphere**	**Heart**	**WDB C**	**Firms**	**Australian**	**German**	**Banknote**	**MOTII12**
SPSS Modeler	98,18	94,55	92,7	96,87	92,96	98,77	90	93,33	93,3	100	73,25
STATISTICA	95,46	63,64	87,68	93,73	84,82	97,01	75	86,09	78	98,32	71,5
Авторское ПО	99,55	98,64	98,59	99,43	98,15	100	100	96,81	98,7	100	97,25

Проведены экспериментальные исследования двухуровневого классификатора. Установлено, что существенное сокращение (в среднем в 2 раза) числа объектов за счет исключения из исходного набора объектов, не являющихся опорными векторами, по результатам построения группы SVM-классификаторов не снижает качества классификации, а позволяет в 2–3 раза сократить временные затраты на разработку итогового SVM-классификатора с использованием модифицированного PSO-алгоритма.

Исследована эффективность применения вариантов формирования набора эталонных объектов и процедуры уточнения КР, полученного с применением алгоритмов кластеризации при различных вариантах формирования обучающей выборки для SVM-алгоритма, в случае отсутствия какой-либо информации о возможной классовой принадлежности хотя бы части объектов, наличие которой позволило бы сформировать обучающую и тестовую выборки данных для разработки SVM-классификатора. Для рассматриваемых наборов производилось разбиение на два кластера с использованием различных алгоритмов кластеризации в условиях неопределённости. Результаты разбиения сравнивались с реальными значениями классификации исходных данных. На основе результатов кластеризации (с использованием предложенных четырех вариантов) формировался набор эталонных объектов для обучения SVM-классификатора с последующим его применением к остальным объектам, для уточнения результатов кластеризации. Подбор значений параметров SVM-классификатора во всех случаях производился с использованием модифицированного алгоритма роя частиц.

Таблица 3. Показатели использования вариантов формирования эталонных объектов

Результаты экспериментов	WDBC				Heart				Australian			
	1	2	3	4	1	2	3	4	1	2	3	4
Всего объектов	569	569	569	569	270	270	270	270	690	690	690	690
Число объектов, вошедших в набор эталонных	484	469	548	455	195	194	247	216	559	616	650	552
Число объектов, ошибочно отнесенных в эталонные	36	27	87	18	33	23	47	28	94	101	110	78
Число объектов, для уточнения КР	85	100	21	114	75	76	23	54	131	74	40	138
Число правильно доклассифицированных объектов	82	84	19	98	57	47	13	32	100	47	28	96
Точность классификации,%	93,15	92,44	84,36	94,02	81,11	80,74	78,89	81,48	81,88	81,45	82,32	82,60

Сравнение полученных КР с применением алгоритмов кластеризации в условиях неопределённости и SVM-классификатора с данными реальной классификации (таблица 3) свидетельствует о целесообразности применения предлагаемого метода разработки SVM-классификатора при отсутствии информации о возможной классовой принадлежности объектов. При этом наиболее перспективным является вариант формирования набора эталонных объектов из объектов, результирующая принадлежность к кластерам для которых определяется с применением кластерного ансамбля на основе матрицы подобия векторов меток кластеров и алгоритма спектральной факторизации (вариант 4 в таблице 3).

Для рассматриваемых наборов данных с использованием значения функции $f(z)$, возвращаемого SVM-классификатором для объекта z, выполнено упорядочение объектов внутри каждого класса: чем больше положительное значение $f(z)$, тем точнее объект z определяется в класс с меткой «+1», и чем меньше отрицательное значение $f(z)$, тем точнее объект z определяется в класс с меткой «–1»; значение $f(z)$, равное «–1» и «+1», показывает попадание объекта z на границу классов.

С применением инструментария теории мультимножеств и лингвистической шкалы по результатам группового экспертного оценивания выполнены классификация и упорядочение объектов, в роли которых рассмотрены конкурсные проекты (КП), представленные интервальными значениями по характеристикам оценивания. При этом каждый эксперт группы проводил оценивание и независимую индивидуальную сортировку КП. Исследована зависимость формирования ОРПК от выбора стратегии оценивания КП. Получены результаты упорядочения (ранги) КП по удалённости от «антиидеального» (наихудшего) КП при различных стратегиях оценивания КП. Показано, что выбор той или иной стратегии оценивания может оказать существенное влияние на результаты классификации и упорядочения КП.

В заключении сформулированы основные научные результаты, полученные в ходе диссертационной работы, приведены рекомендации по дальнейшим исследованиям в направлении повышения качества принятия КР.

В приложениях приводятся свидетельства о государственной регистрации программ для ЭВМ, акты внедрения результатов диссертационной работы, представлен материал, дополняющий основной текст диссертации (таблицы и рисунки).

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В диссертационной работе решена актуальная научно-техническая задача, связанная с разработкой и исследованием эффективных методов и алгоритмов классификации и упорядочения объектов с использованием SVM-алгоритма, алгоритмов кластеризации в условиях неопределённости, теории мультимножеств, оптимизационных алгоритмов.

Итогом работы стали следующие основные научные и практические результаты.

1. Разработана и исследована модификация PSO-алгоритма – эволюционного алгоритма поисковой оптимизации, позволяющая в случае разработки SVM-классификатора сократить время на поиск субоптимальных значений параметров классификатора и обеспечивающая высокую точность классификации и упорядочения объектов.

2. Разработан модифицированный метод формирования ансамбля SVM-классификаторов, базирующийся на применении АМД, реализующий интеграцию частных SVM-классификаторов в ансамбль и обеспечивающий более точное КР в соответствии с выбранной «лучшей» стратегией принятия решений.

3. Разработан алгоритм формирования двухуровневого SVM-классификатора на основе SVM-алгоритма, позволяющий принимать высокоточные решения по класси-

фикации и упорядочению сложноорганизованных многомерных данных больших объёмов при приемлемых временных затратах.

4. Разработан двухэтапный метод классификации, основанный на совместном использовании SVM- и *k*NN-алгоритмов, способствующий повышению качества SVM-классификатора, разработанного с использованием радиальной базисной функции ядра при установленных по умолчанию значениях параметров за счёт уменьшения числа ошибок внутри полосы, разделяющей классы.

5. Разработан метод классификации объектов в случае отсутствия информации о возможной классовой принадлежности объектов, основанный на совместном использовании SVM-алгоритма и алгоритмов кластеризации в условиях неопределённости. Метод может использоваться для уточнения результатов кластеризации, полученных с применением алгоритмов кластеризации в условиях неопределённости, для упорядочения объектов, а также с целью дальнейшей классификации новых объектов. В рамках данного метода предложены и исследованы на эффективность варианты формирования набора эталонных объектов, которые могут быть использованы для получения обучающей и тестовой выборок данных при построении SVM-классификатора, обеспечивающие создание более качественного набора объектов для обучения.

6. На основе инструментария теории мультимножеств предложен подход к формированию ОРПК и упорядочению объектов в случае, когда их характеристики определены с использованием лингвистической шкалы.

7. Разработано ПО, реализующее бинарную классификацию и упорядочение объектов с применением предложенных методов и алгоритмов ИАД.

Разработанные методы и алгоритмы позволяют: 1) принимать обоснованные и адекватные решения, в том числе в условиях неопределённости и неточности исходной информации; 2) минимизировать временные затраты, связанные как с необходимостью многократных реализаций классических ДКА с целью выбора соответствующих оптимальных значений параметров, обеспечивающих принятие адекватных решений, так и с необходимостью сбора и учета точных и полных исходных данных (что может быть принципиально невозможным).

Таким образом, в диссертационной работе разработаны, исследованы и апробированы методы и алгоритмы, способствующие повышению обоснованности и адекватности принимаемых решений при решении задач классификации и упорядочения, что является существенным вкладом в теорию и практику обработки информации и ИАД.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в ведущих рецензируемых научных изданиях Перечня ВАК РФ

1. Соколова, Ю.С. Лингвистический подход к задаче классификации конкурсных проектов с применением инструментария теории мультимножеств / Л.А. Демидова, Ю.С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2014. – № 4 (50-1). − С. 109−117.

2. Соколова, Ю.С. Комплексный анализ конкурсных проектов на основе инструментария теории мультимножеств с применением лингвистической шкалы / Л.А. Демидова, Ю.С. Соколова // Современные проблемы науки и образования. – 2014. – № 6. − С. 55.

3. Соколова, Ю.С. Использование SVM-алгоритма для уточнения решения задачи классификации объектов с применением алгоритмов кластеризации / Л.А. Демидова, Ю.С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2015. – № 1 (51). − С. 103−113.

4. Соколова, Ю.С. Аспекты применения алгоритма роя частиц в задаче разработки SVM-классификатора / Л.А. Демидова, Ю.С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2015. – № 3 (53). – С. 84–92.

5. Соколова, Ю.С. Использование модифицированного алгоритма роя частиц в задаче разработки SVM-классификатора / Л.А. Демидова, Ю.С. Соколова // Прикаспийский журнал: управление и высокие технологии. – 2016. – № 1 (33). − С. 26−38.

6. Соколова, Ю.С. Разработка ансамбля SVM-классификаторов с использованием декорреляционного алгоритма максимизации / Л.А. Демидова, Ю.С. Соколова // Информатика и системы управления. − 2016. – № 1 (47). − С. 95−105.

7. Соколова, Ю.С. Разработка двухуровневого классификатора сложноорганизованных многомерных данных больших объёмов / Л.А. Демидова, Ю.С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2016. – № 2 (56). − С. 71−82.

8. Соколова, Ю.С. Классификация данных на основе SVM-алгоритма и алгоритма *k*-ближайших соседей / Л.А. Демидова, Ю.С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2017. – № 4 (62). − С. 119−132.

Публикации в изданиях, индексируемых в международных базах

9. Sokolova, Yu. Use Of Fuzzy Clustering Algorithms' Ensemble For SVM Classifier Development / L. Demidova, E. Nikulchev, Yu. Sokolova // International Review on Modelling and Simulations (IREMOS). – 2015. − Vol. 8, no. 4. – P. 446–457. (**Scopus**)

10. Sokolova, Yu. Training Set Forming For SVM Algorithm With Use Of The Fuzzy Clustering Algorithms Ensemble On Base Of Cluster Tags Vectors Similarity Matrices / L. Demidova, Yu. Sokolova // 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP). − Saint-Petersburg, 2015. – P. 619–622. (**Scopus**)

11. Sokolova, Yu. Modification Of Particle Swarm Algorithm For The Problem Of The SVM Classifier Development / L. Demidova, Yu. Sokolova // 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP). − Saint-Petersburg, 2015. – P. 623–627. (**Scopus**)

12. Sokolova, Yu. Big Data Classification Using The SVM Classifiers With The Modified Particle Swarm Optimization And The SVM Ensembles / L. Demidova, E. Nikulchev, Yu. Sokolova // International Journal of Advanced Computer Science and Applications (IJACSA). – 2016. – Vol. 7, no. 5. − P. 294−312. (**Web of Science**)

13. Sokolova, Yu. The SVM Classifier Based On The Modified Particle Swarm Optimization / L. Demidova, E. Nikulchev, Yu. Sokolova // International Journal of Advanced Computer Science and Applications (IJACSA). – 2016. – Vol. 7, no. 2. − P. 16–24. (**Web of Science**)

14. Sokolova, Yu. Development Of The SVM Classifier Ensemble For The Classification Accuracy Increa se / L. Demidova, Yu. Sokolova // ITM Web of Conferences. − 2016. − Vol. 6. – P. 02003. (**Web of Science**)

15. Sokolova, Yu. SVM Classifiers: the Objects Identification on the Base of their Hyperspectral Features / L. Demidova, Yu. Sokolova, S. Trukhanov // 2017 Seminar on Systems Analysis. ITM Web of Conferences. – 2017. – Vol. 10. – P. 02003. (**Web of Science**)

16. Sokolova, Yu. A Novel SVM-*k*NN Technique for Data Classification / L. Demidova, Yu. Sokolova // 6-th Mediterranean Conference on Embedded Computing (MECO' 2017). – 2017. – P. 459–462. (**Scopus**)

17. Sokolova, Yu. Two-Level Intellectual Classifier Based on the SVM Algorithm / L. Demidova, Yu. Sokolova // 6-th Mediterranean Conference on Embedded Computing (MECO' 2017). – 2017. – P. 463–466. (**Scopus**)

Свидетельства о государственной регистрации программ для ЭВМ

18. Соколова, Ю.С. Мультимножественный классификатор конкурсных проектов на основе лингвистической шкалы / Л.А. Демидова, Ю.С. Соколова // Свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности № 2014663203 от 18.12.2014.

19. Соколова, Ю.С. Классификация данных на основе алгоритмов интеллектуального анализа / Л.А. Демидова, Ю.С. Соколова // Свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности № 2016612094 от 18.02.2016.

20. Соколова, Ю.С. Двухуровневый интеллектуальный классификатор сложноорганизованных многомерных данных / Л.А. Демидова, Ю.С. Соколова // Свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности № 2016619974 от 01.09.2016.

21. Соколова, Ю.С. Ансамбль SVM-классификаторов на основе декорреляционного алгоритма максимизации / Л.А. Демидова, Ю.С. Соколова // Свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности № 2016660479 от 15.09.2016.

22. Соколова, Ю.С. SVM-*k*NN-классификатор данных / Л.А. Демидова, Ю.С. Соколова // Свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности № 2017662535 от 10.11.2017.

Публикации в других изданиях

23.Соколова, Ю.С. Решение задачи классификации данных с использованием алгоритма нечётких *c*-средних и SVM-алгоритма / Ю.С. Соколова // Задачи системного анализа, управления и обработки информации: межвузовский сборник научных трудов. − М.: МТИ, 2015. – С. 140–147.

24.Соколова, Ю.С. Принципы выполнения SVM-классификации в аналитических приложениях / Ю.С. Соколова // Математическое и программное обеспечение вычислительных систем: межвузовский сборник научных трудов / под ред. А.Н. Пылькина. – М.: Горячая линия-Телеком, 2015. − С. 124−130.

Публикации в трудах Международных и Всероссийских конференций

25.Соколова, Ю.С. Разработка решающих правил классификации конкурсных проектов на основе мультимножеств с применением лингвистического подхода / Ю.С. Соколова // Новые информационные технологии в научных исследованиях и в образовании: материалы XIX Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. – Рязань: РГРТУ, 2014. – С. 3−5.

26.Соколова, Ю.С. Особенности реализации программного обеспечения для классификации объектов на основе SVM-алгоритма и алгоритмов кластеризации / Ю.С. Соколова // Молодежная наука в развитии регионов: Материалы V Всероссийской конференции студентов и молодых ученых. − Пермь: ПНИПУ, 2015. – С. 432−436.

27.Sokolova, J.S. Cluster ensembles' development on the base of SVM-algorithm / J.S. Sokolova // Modern informatization problems in economics and safety: Proceedings of the XX-th International Open Science Conference (Yelm, WA, USA, January 2015). Editor in Chief Dr. Sci., Prof. O.Ja. Kravets − Yelm, WA, USA: Science Book Publishing House, 2015. − P. 38−43.

28.Соколова, Ю.С. Программная реализация модифицированного алгоритма роя частиц для выбора ядра и его параметров при разработке SVM-классификатора / Ю.С. Соколова // Проблемы передачи и обработки информации в сетях и системах телекоммуникаций: Материалы 18-й Международной научно-технической конференции. − Рязань: РГРТУ, 2015. – С. 297−300.

29.Соколова, Ю.С. Особенности использования декорреляционного алгоритма максимизации в задаче разработки ансамбля SVM-классификаторов / Ю.С. Соколова // Современные технологии в науке и образовании – СТНО-2016 [текст]: сб. тр. междунар. науч.-техн. и науч.-метод. конф.: в 4 т. Т.3./ под общ. ред. О.В. Миловзорова. – Рязань: РГРТУ, 2016; – С. 68–71.

30.Соколова, Ю.С. Использование набора опорных векторов в качестве обучающей выборки при разработке двухуровневого SVM-классификатора / Ю.С. Соколова // Математика: фундаментальные и прикладные исследования и вопросы образования: материалы Международной научно-практической конференции. − Рязань: РГУ, 2016. – С. 277−282.

31.Соколова, Ю.С. Решение задачи сортировки с использованием SVM-алгоритма / Ю.С. Соколова // Новые информационные технологии в научных исследованиях: материалы XXI Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. – Рязань: РГРТУ, 2016. – С. 166−168.

Соколова Юлия Сергеевна

МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА ДАННЫХ НА ОСНОВЕ ИНСТРУМЕНТАРИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ИНФОРМАЦИИ И БИОИНСПИРИРОВАННОГО МОДЕЛИРОВАНИЯ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать 04.07.2018. Формат бумаги 60×84 1/16.
Бумага офсетная. Печать трафаретная. Усл. печ. л. 2,0.
Уч.-изд. л. 2,0. Тираж 100 экз.
Рязанский государственный радиотехнический университет.
390005, г. Рязань, ул. Гагарина, д. 59/1.