На правах рукописи

КЛЮЕВА Ирина Алексеевна

МЕТОДЫ И АЛГОРИТМЫ АНСАМБЛИРОВАНИЯ И ПОИСКА ЗНАЧЕНИЙ ПАРАМЕТРОВ КЛАССИФИКАТОРОВ

Специальность 05.13.17 – Теоретические основы информатики (технические науки)

А в т о р е ф е р а т
диссертации на соискание ученой степени кандидата технических наук

Пенза – 2021

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

Научный руководитель – доктор технических наук, доцент
Баранчиков Алексей Иванович

Официальные оппоненты: **Жуков Дмитрий Олегович**,
доктор технических наук, профессор,
ФГБОУ ВО «МИРЭА – Российский технологический университет»,
(г. Москва), профессор кафедры «Информационное противоборство»;

Балашова Ирина Юрьевна,
кандидат технических наук, доцент,
ФГБОУ ВО «Пензенский государственный университет» (г. Пенза), доцент кафедры «Математическое обеспечение и применение ЭВМ»

Ведущая организация – ФГБОУ ВО «Воронежский государственный технический университет»

Защита состоится «___»_____2021 г., в ____часов, на заседании диссертационного совета Д 212.186.01 на базе ФГБОУ ВО «Пензенский государственный университет» по адресу: 440026, г. Пенза, ул. Красная, 40.

С диссертацией и авторефератом можно ознакомиться в библиотеке ФГБОУ ВО «Пензенский государственный университет» и на сайте университета http://dissov.pnzgu.ru/ecspertiza/Tehnicheskie_nauki/klyueva

Автореферат разослан «___»_____2021 г.

Ученый секретарь
диссертационного совета **Кольчугина Елена Анатольевна**

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В настоящее время в различных прикладных областях наблюдается тенденция увеличения спроса на анализ данных в условиях перманентного увеличения объемов обрабатываемой информации. С одной стороны, это связано с потребностью своевременного и эффективного принятия решений на основе результатов проводимой аналитики, с другой − со стремительно развивающимися возможностями информационных технологий в сфере сбора, хранения и обработки данных.

При этом достаточно широкое применение в решении различных прикладных задач находят методы и алгоритмы классификации. Они используются в медицине в задачах медицинской диагностики заболеваний, в банковском деле при определении кредитоспособности заемщиков, в сфере предоставления услуг для решения задач анализа оттока клиентов, в производственной отрасли при диагностике неисправностей оборудования, при обработке текстов на естественных языках и в задачах распознавания речи и др.

Задача классификации относится к области анализа данных на основе прецедентов и подразумевает использование данных с заранее предопределённым конечным множеством классов. Различают бинарную, многоклассовую, а также одноклассовую классификацию. В связи с востребованностью анализа больших объемов данных в практических задачах наиболее актуальной является задача многоклассовой классификации (наличие более двух классов объектов).

В связи с тем, что не существует универсальных методов и алгоритмов, гарантирующих высокое качество классификации произвольных наборов данных, требуется разработка новых методов и алгоритмов, обеспечивающих повышение точности принимаемых классификационных решений в различных предметных областях.

По причине отсутствия общих стандартных рекомендаций при разработке классификаторов одной из приоритетных задач является выбор значений параметров классификаторов, при которых достигается высокое качество классификации данных. При этом целесообразно использовать оптимизационные алгоритмы, в частности, основанные на эволюционном моделировании и обеспечивающие нахождение субоптимального решения за приемлемое время.

Существенное влияние на точность классификатора может оказывать различное соотношение количества объектов разных классов в наборе данных, используемом при его разработке. В случае дисбаланса классов применяются алгоритмы восстановления сбалансированности классов, основанные на принципах сэмплинга (*sampling*). Поиск значений параметров алгоритмов сэмплинга при простом переборе может сопровождаться значительными временными затратами. В связи с этим целесообразно выполнение разработок по созданию соответствующих оптимизационных алгоритмов поиска значений параметров алгоритмов сэмплинга, обеспечивающих разработку классификатора, характеризующегося высокой точностью классификации.

При решении многих практических задач классификации используется построение ансамблей (композиций) классификаторов. Процесс ансамблирования основывается на различных способах компенсации ошибок классификаторов, входящих в ансамбль. Востребованность ансамблей связана с их высокой способностью к обобщению. При этом в зависимости от специфики решаемой задачи классификации к недостаткам ансамблирования можно отнести сложность выбора моделей базовых классификаторов для построения ансамбля, переобучение, неэффективное использование объема обучающей выборки. Перспективным направлением является совершенствование подходов построения композиций классификаторов.

Таким образом, существуют различные актуальные области исследований, направленные на повышение точности классификации и предполагающие реализацию новых разработок, результаты применения которых в настоящее время востребованы в различных прикладных областях анализа данных.

При решении целого спектра задач классификации благодаря обобщающей способности широко востребован SVM-алгоритм (*Support Vector Machine*, SVM – алгоритм опорных векторов). Поскольку математическое описание SVM-алгоритма было первоначально разработано для решения задач бинарной классификации, существенный интерес представляет исследование возможностей его применения для решения задач многоклассовой классификации, а также разработка новых методов повышения точности бинарной SVM-классификации с помощью современных подходов эволюционного моделирования на основе PSO-алгоритма (*Particle Swarm Optimization Algorithm*, PSO – алгоритма роя частиц) и построения композиции классификаторов с использованием эффективного ансамблевого алгоритма, в частности, вспомогательного RF-алгоритма (*Random Forest*, RF – алгоритма случайного леса).

Степень разработанности темы исследования. Существенный вклад в развитие направления исследований проблем и задач классификации внесли ученые: М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр, В. Н. Вапник, А. Я. Червоненкис.

Оптимизационные методы на основе эволюционного моделирования исследуются в работах Дж. Кеннеди, Р. Эберхарта, А. П. Карпенко, В. В. Курейчика, В. М. Курейчика, В. В. Емельянова.

Проблемы несбалансированности данных рассматриваются в публикациях Х. Хе, Х. Хан, А. Чжэна, А. Казари. Исследованиям применения одноклассовой классификации на основе поиска аномалий посвящены работы А. Н. Будынкова, С. И. Масолкина.

Теоретические основы построения композиций алгоритмов приведены в работах Ю. И. Журавлёва. Использование ансамблей алгоритмов для решения задач многоклассовой классификации представлено в работах М. Эйтана, Л. Рокач, А. Сивальда.

Исследованиям применения эволюционного моделирования в решении задачи поиска значений параметров классификатора, а также подходам построения ансамблей классификаторов в случае бинарной классификации посвящены работы Ю. С. Соколовой.

Несмотря на наличие значительного количества научных трудов, посвященных проблемам повышения качества решения задач классификации, исследования существующих алгоритмов показывают отсутствие абсолютно универсальных классификаторов, являющихся однозначно лучшими для классификации различных произвольных наборов данных, поскольку такие алгоритмы характеризуются определенными ограничениями, заложенными в их математический аппарат. Таким образом, целесообразна разработка новых методов и алгоритмов, обладающих высокой обобщающей способностью и позволяющих снизить недостатки существующих аналогов при решении задач классификации.

Цель диссертации – повышение эффективности решения задач бинарной и многоклассовой классификации посредством разработки методов и алгоритмов, реализующих поиск значений параметров классификаторов, построение ансамблей классификаторов, а также снижение влияния дисбаланса классов при использовании несбалансированных наборов данных.

Для достижения поставленной цели решаются следующие **основные задачи**:

1. Определить в рамках работы базовый алгоритм классификации и исследовать существующие методы поиска оптимальных значений параметров классификаторов, предложить на основе исследованных методов более эффективные алгоритмы, позволяющие сократить временные затраты на поиск оптимальных значений параметров базового классификатора для обеспечения высокой точности классификации данных.

2. Провести анализ существующих методов многоклассовой классификации и предложить новые методы, основанные на построении ансамблей алгоритмов, обеспечивающие высокую точность многоклассовой классификации.

3. Определить влияние дисбаланса классов на значения показателей качества базового классификатора и предложить алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных.

4. Исследовать методы повышения точности бинарной классификации на основе построения композиций алгоритмов и разработать эффективный алгоритм двухэтапной классификации, обеспечивающий повышение точности решений с использованием нескольких классификаторов, один из которых является основным, другой – вспомогательным.

5. Исследовать методы поиска аномалий при решении задачи одноклассовой классификации и возможность их применения в случаях наличия дисбаланса классов при бинарной классификации, разработать алгоритм двухэтапной классификации на основе поиска аномалий, обеспечивающий повышение точности классификации несбалансированных наборов данных на основе композиции классификаторов, включающей основной и вспомогательный классификаторы.

Объектом исследования является бинарная и многоклассовая классификация многомерных произвольных наборов данных в различных предметных областях.

Предметом исследования являются методы и алгоритмы классификации данных с применением машинного обучения, включающие различные способы ансамблирования и поиска значений параметров классификаторов, а также подходы к восстановлению сбалансированности классов.

Методы исследования. Для решения поставленных задач в настоящей работе использовались методы интеллектуального анализа данных, математической статистики, стохастической оптимизации, сэмплирования, поиска аномалий, математический аппарат алгоритмов машинного обучения.

Научная новизна:

1. Предложены новые алгоритмы поиска оптимальных значений параметров базового классификатора (SVM-PSO-GS, SVM-PSO-DOE-алгоритмы), *отличающиеся* от аналогов совместным применением алгоритма роя частиц и алгоритмов поиска по сетке, *позволяющие* по сравнению с алгоритмом роя частиц сократить временные затраты на поиск оптимальных значений параметров базового классификатора в среднем в 3 раза при обеспечении высокой точности классификации данных (05.13.17 п. 5).

2. Предложены новые ансамблевые методы многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2-методы), *отличающиеся* от аналогов новыми подходами к генерации метахарактеристик при нескольких разбиениях обучающей выборки, *позволяющие* повысить точность многоклассовой классификации, реализуемой с использованием композиции бинарных классификаторов, до 5–7 % (05.13.17 п. 5).

3. Предложен новый алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных (SVM-bSMOTE-алгоритм), *отличающийся* от аналогов реализацией оценки пригодности комбинаций значений параметров алгоритма сэмплинга, *позволяющий* обеспечить разработку классификатора, точность классификации которого в среднем на 3 % превышает точность базового классификатора (05.13.17 п. 5).

4. Предложен новый алгоритм двухэтапной классификации (SVM-RF-алгоритм), *отличающийся* от аналогов последовательным применением базового и вспомогательного ансамблевого классификаторов, *позволяющий* уменьшить число ошибок классификации вблизи гиперплоскости, разделяющей классы, и повысить точность бинарной классификации в среднем на 5 % по сравнению с базовым классификатором (05.13.17 п. 5).

5. Предложен новый алгоритм двухэтапной классификации на основе поиска аномалий (1-SVM-RF-алгоритм), *отличающийся* от аналогов последовательным применением одноклассового классификатора в качестве основного и вспомогательного ансамблевого классификатора, *позволяющий* повысить точность бинарной классификации несбалансированных наборов данных в среднем на 2,5 % по сравнению с базовым классификатором (05.13.17 п. 5).

Практическая ценность работы заключается в следующем:

1. Алгоритмы поиска оптимальных значений параметров базового классификатора (SVM-PSO-GS, SVM-PSO-DOE-алгоритмы) за счет совместного

применения алгоритма роя частиц и алгоритмов поиска по сетке *позволяют* по сравнению с алгоритмом роя частиц сократить временные затраты на поиск оптимальных значений параметров базового классификатора в среднем в 3 раза при обеспечении высокой точности классификации данных.

2. Ансамблевые методы многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2-методы) за счет новых подходов к генерации метахарактеристик при нескольких разбиениях обучающей выборки *позволяют* повысить точность многоклассовой классификации, реализуемой с использованием композиции бинарных классификаторов, до 5–7 %.

3. Алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных (SVM-bSMOTE-алгоритм) за счет реализации оценки пригодности комбинаций значений параметров алгоритма сэмплинга *позволяет* обеспечить разработку классификатора, точность классификации которого в среднем на 3 % превышает точность базового классификатора.

4. Алгоритм двухэтапной классификации (SVM-RF-алгоритм) за счет последовательного применения базового и вспомогательного ансамблевого классификаторов *позволяет* уменьшить число ошибок классификации вблизи гиперплоскости, разделяющей классы, и повысить точность бинарной классификации в среднем на 5 % по сравнению с базовым классификатором.

5. Алгоритм двухэтапной классификации на основе поиска аномалий (1-SVM-RF-алгоритм) за счет последовательного применения одноклассового классификатора в качестве основного и вспомогательного ансамблевого классификатора *позволяет* повысить точность бинарной классификации несбалансированных наборов данных в среднем на 2,5 % по сравнению с базовым классификатором.

Соответствие паспорту научной специальности. Область исследования и содержание диссертации соответствуют паспорту специальности 05.13.17 «Теоретические основы информатики (технические науки)» в части п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений».

Достоверность результатов исследований подтверждается внутренней непротиворечивостью основных теоретических положений и соответствием экспериментальных результатов фактическим данным о классовой принадлежности объектов классификации; воспроизводимостью результатов исследования при использовании различных тестовых наборов данных из репозитория машинного обучения UCI и платформы соревнований по машинному обучению Kaggle; корректным использованием математического аппарата известных алгоритмов машинного обучения по прецедентам, эволюционного моделирования, стандартных метрик качества классификации, современных средств программирования и библиотек машинного обучения; сравнением результатов работы с результатами других научных исследований; апробацией на всероссийских и международных научно-технических конференциях;

публикациями в рецензируемых российских и зарубежных изданиях; актами о внедрении; полученными свидетельствами о государственной регистрации программ для ЭВМ.

Основные положения, выносимые на защиту:

1. Алгоритмы поиска оптимальных значений параметров базового классификатора (SVM-PSO-GS, SVM-PSO-DOE-алгоритмы), *позволяющие* по сравнению с алгоритмом роя частиц сократить временные затраты на поиск оптимальных значений параметров базового классификатора при обеспечении высокой точности классификационных решений за счет совместного применения алгоритма роя частиц и алгоритмов поиска по сетке.

2. Ансамблевые методы многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2-методы), *позволяющие* повысить точность многоклассовой классификации, реализуемой с использованием композиции бинарных классификаторов, за счет новых подходов к генерации метахарактеристик при нескольких разбиениях обучающей выборки.

3. Алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных (SVM-bSMOTE-алгоритм), *позволяющий* за счет реализации оценки пригодности комбинаций значений параметров алгоритма сэмплинга обеспечить разработку классификатора, точность классификации которого превышает точность базового классификатора.

4. Алгоритм двухэтапной классификации (SVM-RF-алгоритм), *позволяющий* уменьшить число ошибок классификации вблизи гиперплоскости, разделяющей классы, и повысить точность бинарной классификации по сравнению с базовым классификатором за счет последовательного применения базового и вспомогательного ансамблевого классификаторов.

5. Алгоритм двухэтапной классификации на основе поиска аномалий (1-SVM-RF-алгоритм), *позволяющий* повысить точность бинарной классификации несбалансированных наборов данных по сравнению с базовым классификатором за счет последовательного применения одноклассового классификатора в качестве основного и вспомогательного ансамблевого классификатора.

Реализация и внедрение. Методы и алгоритмы, разработанные в настоящей работе, внедрены в следующих организациях:

– банк «Вятич» (ПАО) (г. Рязань) – при оценке кредитоспособности заемщиков;

– ООО «НБЛ» (г. Рязань) – для оценки спроса на аренду недвижимости;

– в учебном процессе на кафедре вычислительной и прикладной математики факультета вычислительной техники Федерального государственного бюджетного образовательного учреждения высшего образования «Рязанский государственный радиотехнический университет» (ФГБОУ ВО «РГРТУ») (г. Рязань).

Имеются соответствующие акты внедрения результатов работы.

Апробация результатов исследования. Основные результаты диссертационной работы докладывались на 18 научных конференциях (из них 4 –

всероссийских, 14 – международных): XVIII Международной научно-технической конференции «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций» (г. Рязань, 2015), XXI, XXIII Всероссийских научно-технических конференциях студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и в образовании» (НИТ-2016, НИТ-2018) (г. Рязань, 2016, 2018); Всероссийской научно-технической конференции «Интеллектуальные и информационные системы» (Интеллект-2016) (г. Тула, 2016); I и II Международных научно-технических конференциях «Современные технологии в науке и образовании» (СТНО-2016, СТНО-2017) (г. Рязань, 2016, 2017); Всероссийской научно-технической конференции «Математические методы и информационные технологии управления в науке, образовании и правоохранительной сфере» (ММИТ-2017) (г. Рязань, 2017); XII и XIII Международных симпозиумах «Интеллектуальные системы» (INTELS'2016, INTELS'2018) (г. Москва, 2016; г. Санкт-Петербург, 2018); V, VI, VII Международных Средиземноморских научно-технических конференциях «Встраиваемые вычислительные системы» (MECO 2016, MECO 2017, MECO 2018) (Черногория, г. Бар, 2016, 2017; г. Будва, 2018); III, IV, V Международных научных конференциях «Прикладные исследования и технологии» (ART-2016, ART-2017, ART-2018) (г. Москва, 2016, 2017, 2018); Семинаре по системному анализу (ITM Web of Conferences) (г. Москва, 2017); I Международной научной конференции «Системы Управления, Математическое моделирование и Автоматизация в промышленности и энергетике» (SUMMA) (г. Липецк, 2019), Международной научно-практической конференции «Big Data and Artificial Intelligence Conference» (2020).

Публикации. По теме диссертации опубликована 31 работа, в том числе 6 статей в журналах, рекомендованных ВАК РФ, 7 работ в изданиях, индексируемых в Scopus. Получены 3 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад. Представленные на защиту результаты получены автором диссертации лично. Автором самостоятельно исследованы способы повышения качества классификации и разработаны новые методы и алгоритмы, обеспечивающие повышение точности бинарной и многоклассовой классификации. В работах [1–3, 5, 7–11, 13] диссертантом получены основные экспериментальные результаты, проведены их анализ и интерпретация. В работах [5, 6], подготовленных в соавторстве с научным руководителем, которому принадлежит общая постановка задач исследования, соискателем сформулированы и реализованы подходы к решению. Работы [4, 12] выполнены диссертантом самостоятельно. Все основные положения тезисов докладов конференций сформулированы лично автором.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, перечня сокращений, списка использованных источников, пяти приложений. Основной текст работы содержит 142 страницы, 13 таблиц, 42 рисунка.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе приведены основные принципы классификации произвольных многомерных наборов данных, а также описание основных показателей качества классификации; проведен обзор известных методов и алгоритмов машинного обучения, применяемых при решении задач классификации, выделены их достоинства и недостатки.

В рамках диссертационной работы в качестве базового алгоритма классификации используется SVM-алгоритм и осуществляется ряд исследований по разработке эффективных методов и алгоритмов, позволяющих повысить точность SVM-классификатора.

В основу SVM-алгоритма, реализующего построение бинарного классификатора, заложено нахождение оптимальной разделяющей гиперплоскости посредством решения задачи квадратичного программирования, имеющей единственное решение.

Показано, что к основным проблемам, влияющим на качество SVM-классификатора и требующим проведения дополнительных исследований, относятся: поиск оптимальных значений параметров SVM-классификатора; несбалансированность классов в наборах данных, используемых при решении задач классификации; возникновение ошибок SVM-классификации вблизи гиперплоскости, разделяющей классы; проблема повышения точности многоклассовой SVM-классификации.

Во второй главе для сокращения временных затрат на построение SVM-классификатора, характеризующегося высокой точностью классификационных решений, разработаны новые алгоритмы поиска значений параметров классификации.

Предложены алгоритмы поиска значений параметров базового SVM-классификатора на основе совместного применения PSO-алгоритма и алгоритмов поиска по сетке: GS-алгоритма (*Grid Search*, GS – классического алгоритма поиска по сетке) и DOE-алгоритма (*Design of Experiment*, DOE – алгоритма поиска по сетке на основе планирования экспериментов). Разработанные **SVM-PSO-GS- и SVM-PSO-DOE-алгоритмы** обеспечивают сокращение временных затрат на поиск оптимальных значений параметров SVM-классификатора: параметра регуляризации и коэффициента радиальной базисной функции ядра.

При реализации разработанных алгоритмов предлагается на каждой итерации PSO-алгоритма выполнять уточнение положения (координат) глобально «лучшей» частицы в рое с использованием алгоритма поиска по сетке с последующим обновлением текущей популяции частиц роя. При этом из роя удаляется «худшая» частица (т.е. частица с «худшим» значением целевой функции) и вместо нее добавляется «лучшая» частица, найденная алгоритмом поиска по сетке.

На каждой итерации PSO-алгоритма обновляются границы диапазонов поиска по сетке: $[d^j_{\min}, d^j_{\max}]$ $(j=\overline{1,n})$. При этом обновление производится с учетом диапазонов $[r^j_{\min}, r^j_{\max}]$ разброса частиц в рое на текущей итерации PSO-алгоритма. В качестве значений координат χ^j $(j=\overline{1,n})$ «главного» (центрального) узла сетки используются значения координат глобально «лучшей» частицы в рое. Такой узел рассматривается в качестве центроида, вокруг которого строится сетка (рисунок 1).

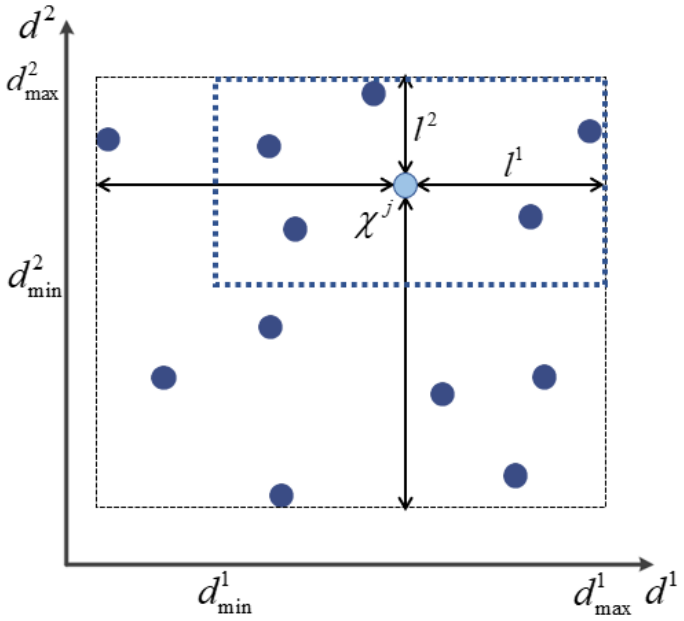

Рисунок 1 – Определение границ диапазонов поиска по сетке

Минимальное расстояние от лучшей частицы роя (центроида сетки) до границ диапазонов разброса находится как

$$l^j = \min\{\chi^j - r^j_{\min}, r^j_{\max} - \chi^j\}, \tag{1}$$

а границы диапазонов поиска по сетке определяются как

$$d^j_{\min} = \chi^j - l^j, \tag{2}$$

$$d^j_{\max} = \chi^j + l^j. \tag{3}$$

Ускорение поиска оптимального решения достигается благодаря:

– дополнительному поиску по сетке в области потенциального глобально «лучшего» положения частиц в рое;

– обновлению популяции роя частиц и удалению «худших» частиц.

Предложен *алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных* (***SVM-bSMOTE-алгоритм***), позволяющий сократить временные затраты на разработку SVM-классификатора, характеризующегося высокой точностью классификации.

Обучение классификаторов на несбалансированных наборах данных приводит к тому, что разработанный в итоге классификатор стремится классифицировать все объекты в качестве объектов мажоритарного класса (класса «большинства»), практически игнорируя менее представленный миноритарный класс (класса «меньшинства»), что обычно не соответствует фактической цели исследования.

Для решения проблемы несбалансированности набора данных применяются различные алгоритмы сэмплинга, в частности, bSMOTE-алгоритм (*Borderline Synthetic Minority Oversampling Technique*, bSMOTE – алгоритм синтезирования объектов миноритарного класса вдоль границы классов), реализующий синтезирование новых объектов миноритарного класса на основе объектов этого же класса, расположенных вблизи границы классов.

В диссертационной работе разработан алгоритм, позволяющий существенно сократить временные затраты на поиск оптимальных значений параметров bSMOTE-алгоритма по сравнению со стандартным подходом, предполагающим многократный запуск bSMOTE-алгоритма для различных значений параметров с целью выбора их комбинации, обеспечивающей разработку SVM-классификатора, характеризующегося высокой точностью классификации.

Разработанный SVM-bSMOTE-алгоритм реализует поиск оптимальных значений таких параметров, как k − число ближайших соседей, используемых для создания синтетических образцов; m − число ближайших соседей, используемых для определения, находится ли объект миноритарного класса на границе классов (рисунок 2).

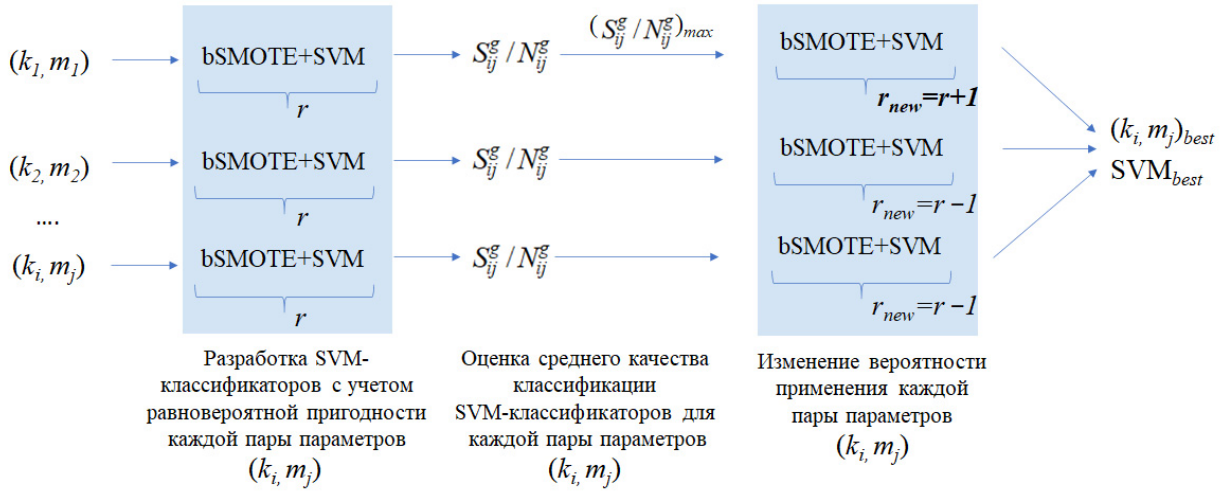

Рисунок 2 – Этапы реализации SVM-bSMOTE-алгоритма

Разработанный SVM-bSMOTE-алгоритм одновременно реализует поиск значений параметров bSMOTE-алгоритма для последующего восстановления сбалансированности классов и разработку SVM-классификатора, характеризующегося высоким качеством классификации данных. При этом на каждой итерации SVM-bSMOTE-алгоритма производится изменение вероятности применения каждой пары параметров (k_i, m_j) при реализации bSMOTE-алгоритма с последующей разработкой SVM-классификатора.

Для оценки среднего качества классификации SVM-классификаторов для каждой пары (k_i, m_j) предлагается использовать следующую последовательность шагов: 1) для каждой пары (k_i, m_j) найти общее число N_{ij}^g SVM-классификаторов, полученных за текущее число итераций SVM-bSMOTE-алгоритма; 2) для каждой пары (k_i, m_j) найти общую сумму S_{ij}^g значений показателя качества классификации (в частности, показателя сбалансированной F-меры) SVM-классификаторов, полученных за текущее число итераций SVM-bSMOTE-алгоритма; 3) для каждой пары (k_i, m_j) найти отношение S_{ij}^g / N_{ij}^g и использовать как среднее качество классификации SVM-классификаторов для пары (k_i, m_j).

В третьей главе разработаны новые методы и алгоритмы бинарной и многоклассовой классификации на основе ансамблей классификаторов.

*Предложен алгоритм двухэтапной классификации (**SVM-RF-алгоритм**), основанный на последовательном применении базового SVM-классификатора и вспомогательного RF-классификатора.*

При этом в задаче бинарной классификации реализовано уточнение решений SVM-классификатора для объектов, находящихся вблизи гиперплоскости, разделяющей классы, в экспериментально определяемой симметричной α-области, содержащей как верно, так и ошибочно классифицированные объекты.

Формирование α-области (рисунок 3) производится на основе вычисления максимального расстояния, на которое удалены от разделяющей гиперплоскости ошибочно классифицированные объекты обучающей выборки внутри разделяющей полосы:

$$d_{\max}^{\alpha} = \max\{|d^-|, |d^+|\}, \tag{4}$$

где $2 \cdot d_{\max}^{\alpha}$ – ширина α-области; d^-, d^+ – максимальное расстояние от разделяющей гиперплоскости ошибочно классифицированных объектов с метками класса «–1», «+1» соответственно.

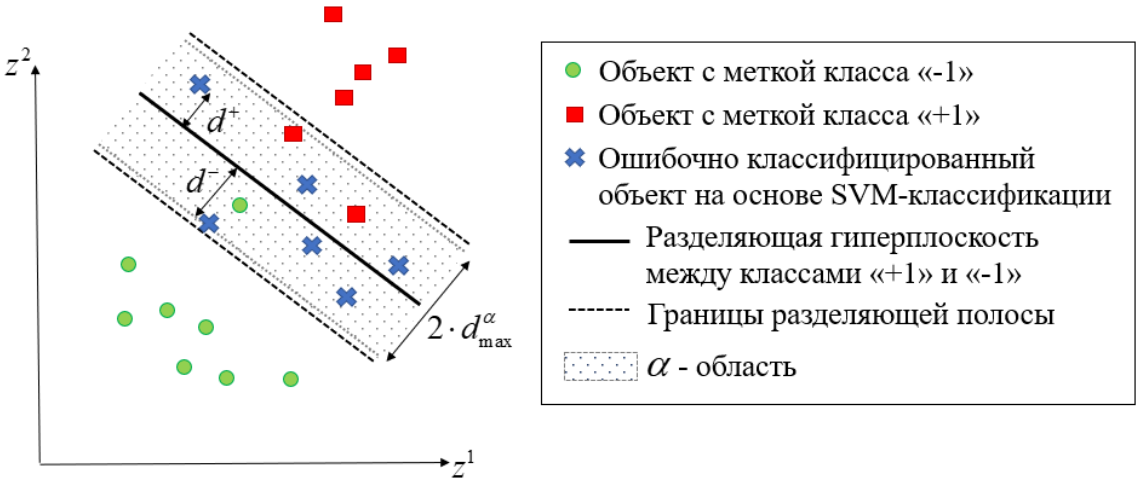

Рисунок 3 – Формирование симметричной α-области

*Предложен алгоритм двухэтапной классификации на основе поиска аномалий (**1-SVM-RF-алгоритм**).*

Ограничение применимости разработанного SVM-RF-алгоритма, реализующего уточнение классификационных решений SVM-классификатора для объектов, находящихся вблизи гиперплоскости, разделяющей классы, заключается в следующем: поскольку α-область содержит как ошибочно классифицированные объекты, так и попавшие в данную область правильно классифицированные объекты, в том случае, если ширина α-области достаточно велика или объекты внутри α-области сильно уплотнены, число объектов, используемых для обучения вспомогательного классификатора, может оказаться недостаточным для решения задачи повышения качества бинарной SVM-классификации.

В связи с этим в настоящей работе для решения задачи бинарной классификации наборов данных с несбалансированными классами предлагается использовать алгоритм двухэтапной классификации на основе поиска аномалий с применением одноклассового SVM-классификатора в качестве основного и вспомогательного RF-классификатора.

Идея 1-SVM-алгоритма (1-SVM – одноклассового SVM-алгоритма) заключается в обнаружении редких событий или аномалий. Редкие события происходят не часто, и, следовательно, у них мало примеров в классифицируемых данных, поэтому задачу поиска аномалий можно интерпретировать как задачу классификации с дисбалансом классов.

Пусть $Z_{imb} = \{(z_1, y_1),...,(z_s, y_s)\}$ – исходный набор данных с несбалансированными классами, в котором каждому объекту $z_i \in Z_{imb}$ ($i = \overline{1,s}$; s – число объектов в исходном наборе данных) соответствует метка класса $y_i \in Y = \{-1;+1\}$.

Для 1-SVM-классификации в качестве объектов нормального класса принимаются объекты мажоритарного класса из набора данных Z_{imb}, в качестве аномалий – объекты миноритарного класса из набора данных Z_{imb}.

Разработанный 1-SVM-RF-алгоритм может быть описан следующей последовательностью шагов:

Шаг 1. На основе набора данных Z_{imb} сформировать обучающую и тестовую выборки (Z_{train} и Z_{test}).

Шаг 2. Реализовать шаг 1-SVM-классификации: обучить 1-SVM-классификатор с использованием объектов только нормального класса из обучающей выборки Z_{train}^{normal} ($Z_{train}^{normal} \subset Z_{train}$) и выполнить предсказание классовой принадлежности объектов из тестовой выборки Z_{test} (определить принадлежность к нормальному классу или к аномалиям).

Шаг 3. Для объектов, которые были определены на шаге 2 как аномалии z^{anomal} ($z^{anomal} \in Z_{test}$), провести уточнение их классовой принадлежности с помощью вспомогательного RF-классификатора. Реализовать шаг RF-классификации: обучить RF-классификатор на всех объектах обучающей выборки Z_{train} и выполнить предсказание классовой принадлежности объектов-аномалий z^{anomal}.

Шаг 4. Объединить результаты 1-SVM- и RF-классификаций: для объектов тестовой выборки, принадлежность которых была отнесена к нормальному классу на шаге 2, т.е. объектов z^{normal} $(z^{normal} \in Z_{test})$, принять результат 1-SVM-классификации; для объектов тестовой выборки, определенных на шаге 2 как объекты-аномалии, т.е. объектов z^{anomal} $(z^{anomal} \in Z_{test})$, принять результат RF-классификации, полученный на шаге 3.

Шаг 5. Оценить качество проведенной двухэтапной классификации.

Предложены ансамблевые методы многоклассовой классификации с использованием новых подходов к генерации метахарактеристик при реализации стекинга (stacking). В процессе решения задачи классификации стекинг реализует построение нескольких различных базовых классификаторов для применения их классификационных решений в качестве метахарактеристик при обучении «обобщающего» классификатора (метаклассификатора).

В разработанных методах предполагается добавить новые метахарактеристики к исходному набору данных, а затем в качестве метаклассификатора использовать композицию бинарных SVM-классификаторов на основе того или иного декомпозиционного метода решения задачи многоклассовой классификации (*One-vs.-Rest*, OvR – «один против всех», или *One-vs.-One*, OvO – «один против одного»), обеспечивающего переход от исходной задачи классификации к нескольким задачам бинарной классификации. При создании метахарактеристик в качестве ответов базовых классификаторов предлагается использовать оценки вероятностей принадлежности объектов к классам.

В разработанных методах производится несколько разбиений $(t = \overline{1,T})$ обучающей выборки на K блоков для дальнейшего получения ответа базовых классификаторов для каждого k-го блока $(k = \overline{1,K})$. Таким образом, множество разбиений обучающей выборки представлено в виде $\{\{(Z_{tk}, Y_{tk}),\ k = \overline{1,K}\},\ t = \overline{1,T}\}$.

Пусть $b_{wt} \in B$ – базовый классификатор из множества $B = \{b_{11}, b_{12}, ...b_{21}, b_{22}, ...b_{wt}\}$, обеспечивающий генерацию метахарактеристик, построенный на основе w-го $(w = \overline{1,W})$ базового алгоритма со значениями параметров, соответствующими t-му $(t = \overline{1,T})$ разбиению обучающей выборки; $B_t \subset B$ – подмножество классификаторов, используемое на t-м разбиении обучающей выборки.

1. DGM-1-метод. Для каждого t-го $(t = \overline{1,T})$ разбиения обучающей выборки реализуется построение одного базового классификатора из множества B $(b_{wt} \in B)$ и использование его ответов в качестве метахарактеристик для обучения другого базового классификатора при $(t+1)$-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при реализации последнего $(t = T)$-го разбиения обучающей выборки.

2. DGM-2-метод. Для каждого t-го $(t=\overline{1,T})$ разбиения обучающей выборки реализуется построение подмножества классификаторов $B_t \subset B$ и используется усреднение их ответов в качестве метахарактеристик для обучения другого множества базовых классификаторов $B_{t+1} \subset B$ при $(t+1)$-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов на последнем $(t=T)$-м разбиении обучающей выборки.

3. **IGM-1-метод.** Для каждого t-го $(t=\overline{1,T})$ разбиения обучающей выборки с помощью кросс-валидации проводится оценка классификаторов из подмножества $B_t \subset B$ и выбор для генерации метахарактеристик классификатора с лучшим значением оценки перекрестной проверки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов при различных t-х разбиениях обучающей выборки.

4. IGM-2-метод. Для каждого t-го $(t=\overline{1,T})$ разбиения обучающей выборки с помощью кросс-валидации проводится оценка классификаторов из подмножества $B_t \subset B$. При $(t+1)$-м разбиении обучающей выборки увеличивается число построений классификаторов на основе того алгоритма, который использовался при построении классификатора с лучшей оценкой перекрестной проверки на предыдущем t-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов при различных t-х разбиениях обучающей выборки.

Таким образом, DGM-1 и DGM-2-методы реализуют обучение базовых классификаторов с использованием метахарактеристик на основе ответов базовых классификаторов с предыдущего разбиения обучающей выборки (*Dependent Generation of Metacharacters*, DGM). При реализации IGM-1 и IGM-2-методов предполагается, что создание метахарактеристик происходит независимо на разных разбиениях обучающей выборки (*Independent Generation of Metacharacters*, IGM).

Оценка перекрестной проверки, используемая в IGM-1 и IGM-2-методах, представляет собой усредненную ошибку по всем k $(k=\overline{1,K})$ блокам на текущем t-м $(t=\overline{1,T})$ разбиении обучающей выборки.

Согласно идеям стекинга, в качестве базовых классификаторов предполагается использование классификаторов, построенных на основе реализации различных по своей «природе» алгоритмов классификации. В диссертационной работе предлагается при построении базовых классификаторов применять различные ансамблевые алгоритмы на основе методов бэггинга (*bagging*) и бустинга (*boosting*).

К преимуществам разработанных методов многоклассовой классификации по сравнению с аналогами следует отнести следующее:

– использование нечеткой оценки принадлежности объектов к классам при получении значений метахарактеристик;

– оценка базовых классификаторов с помощью кросс-валидации;

– по сравнению с методом стекинга в разработанных методах при получении значений метахарактеристик для объектов тестовой выборки используется усреднение ответов базовых классификаторов, обученных на том же объеме обучающей выборки, что и при генерации метахарактеристик для объектов обучающей выборки;

– в отличие от метода блендинга разработанные методы обеспечивают эффективное использование всего объема обучающей выборки при построении метаклассификатора: при обучении метаклассификатора используется вся обучающая выборка, включающая исходный набор характеристик и метахарактеристики на основе ответов базовых классификаторов.

В четвертой главе приводится описание разработанного программного обеспечения, перечень используемых наборов данных, результаты экспериментальных исследований.

Для получения результатов использовались различные наборы данных из репозитория машинного обучения UCI и платформы соревнований по машинному обучению Kaggle. Программная реализация разработанных методов и алгоритмов выполнялась с применением среды программирования Python 3.7 и библиотеки машинного обучения scikit-learn.

В таблице 1 приведены результаты бинарной классификации. При этом в качестве результатов представлены значения (в процентах) показателя сбалансированной точности классификации (исходных наборов данных – *BA*; тестовых выборок исходных наборов данных – *BA_test*): 1) для базового SVM-алгоритма со значениями параметров, заданными по умолчанию – при значении параметра регуляризации $C = 1$ и значении параметра радиальной базисной функции ядра $\sigma = 1$ (SVM); 2) SVM-алгоритма со значениями параметров, найденными с помощью классического подхода поиска по сетке с применением перекрестной проверки (SVM-GS_CV); 3) SVM-алгоритма со значениями параметров, найденными с помощью PSO-алгоритма (SVM-PSO); 4) SVM-алгоритма со значениями параметров, найденными с помощью совместного применения PSO-алгоритма и алгоритмов поиска по сетке (SVM-PSO-GS, SVM-PSO-DOE); 5) алгоритма классификации, включающего использование SVM-классификатора и поиск значений параметров алгоритма сэмплинга (SVM-bSMOTE); 6) алгоритма двухэтапной классификации, основанного на совместном применении SVM-классификатора в качестве основного и вспомогательного RF-классификатора (SVM-RF).

В диссертационной работе проведена бинарная классификация наборов данных из таблицы 1 с использованием других известных алгоритмов машинного обучения, таких как: алгоритм логистической регрессии (*Logistic*

Regression), алгоритм решающих деревьев (*Decision Trees*), алгоритм случайного леса (*Random Forest*), алгоритм градиентного бустинга (*Gradient Boosting*), алгоритм *k* ближайших соседей (*k Nearest Neighbor*), алгоритм адаптивного бустинга (*AdaBoost*), алгоритм нейронной сети (*Neural network*), наивный байесовский классификатор (*Naive Bayes*). При этом применялись параметры алгоритмов, заданные по умолчанию в библиотеке scikit-learn версии 0.22.1. Результаты экспериментальных исследований показывают, что по сравнению с применением известных алгоритмов машинного обучения разработанные в диссертации алгоритмы бинарной классификации не уступают по качеству классификационных решений либо превосходят точность классификации различных наборов данных.

Таблица 1 – Результаты бинарной классификации

Набор данных	Показатель	*SVM*	*SVM-GS_CV*	*SVM-PSO*	*SVM-PSO-GS*	*SVM-PSO-DOE*	*SVM-bSMOTE*	SVM-RF
Heart 270x13	*BA*	93,75	86,08	94,25	95,33	95,33	94,93	99,63
	BA_test	65,91	73,98	87,08	87,08	87,08	83,12	98,15
Australian 690x14	*BA*	94,19	86,95	94,22	95,96	95,96	94,84	95,63
	BA_test	75,51	84,69	87,12	87,12	87,12	82,31	90,19
МОТП12 400x2	*BA*	90,33	93,00	96,00	96,00	96,00	95,61	–
	BA_test	86,15	82,49	95,24	95,24	95,24	86,57	
LSVT 126x309	*BA*	92,00	85,11	92,86	92,86	92,86	95,14	100
	BA_test	50,00	83,33	50,00	50,00	50,00	50,00	100
Firms 60x12	*BA*	93,33	96,67	96,67	96,67	96,67	–	100
	BA_test	75,71	83,33	83,33	83,33	83,33		100
Hepatitis 155x19	*BA*	92,18	81,59	96,06	96,06	96,06	97,82	–
	BA_test	50,00	60,65	83,15	83,15	83,15	78,79	
WDBC 699x10	*BA*	98,91	96,71	98,64	98,95	99,16	99,33	99,67
	BA_test	94,62	95,78	95,24	95,79	95,79	97,89	100
Mammographic 961x5	*BA*	84,70	84,70	87,83	88,23	88,23	88,31	84,58
	BA_test	78,98	78,98	75,61	82,44	82,44	81,05	85,50
Biodeg 1055x41	*BA*	86,87	92,30	94,70	95,00	95,00	96,75	–
	BA_test	82,31	81,21	85,77	86,00	86,00	84,04	

Результаты экспериментальных исследований, приведенные в таблице 1, демонстрируют, что разработанные алгоритмы (SVM-PSO-GS, SVM-PSO-DOE, SVM-RF, SVM-bSMOTE) обеспечивают лучшие значения показателя сбалансированной точности классификации по сравнению с базовыми SVM- и SVM-GS_CV-алгоритмами. Разработанные SVM-PSO-GS, SVM-PSO-DOE-алгоритмы могут обеспечивать более высокие значения показателя сбалансированой точности по сравнению с SVM-PSO-алгоритмом, при этом сокращаются временные затраты на поиск субоптимальных значений параметров SVM-классификатора (рисунок 4).

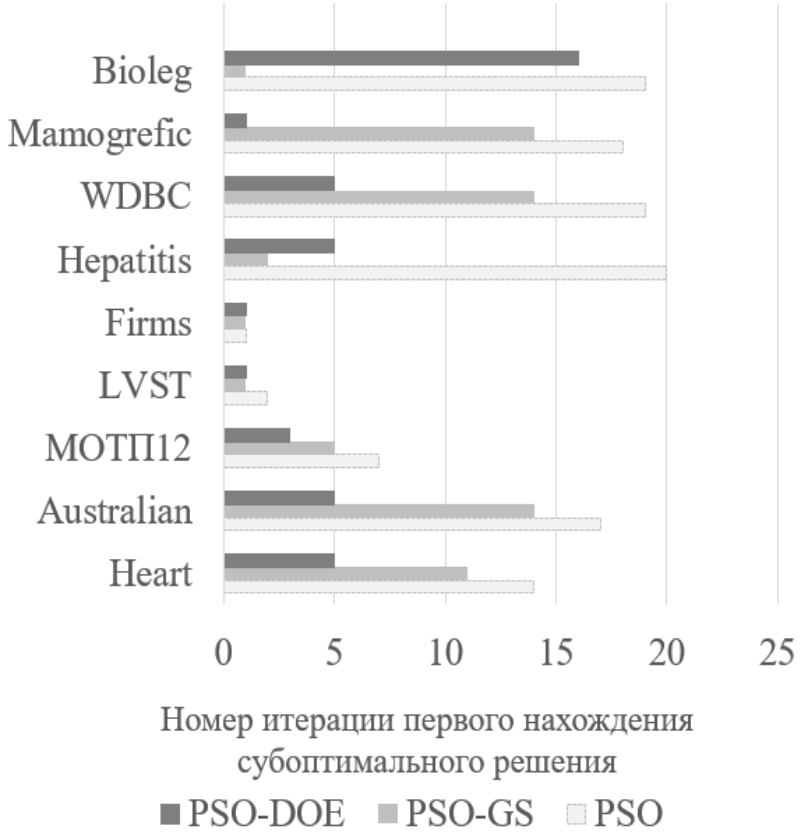

Рисунок 4 – Результаты поиска субоптимальных значений параметров SVM-классификатора

В таблицах 2–3 в качестве результатов представлены значения (в процентах) показателя сбалансированной точности классификации объектов тестовых выборок исходных наборов данных.

В таблице 2 представлены результаты исследований классификации несбалансированных наборов данных с использованием простой реализации алгоритмов (базового SVM-, 1-SVM-, RF-алгоритмов); реализации двухэтапной классификации на основе SVM- и RF-алгоритмов (SVM-RF); реализации двухэтапной классификации на основе 1-SVM- и RF-алгоритмов (1-SVM-RF).

Таблица 2 – Результаты классификации несбалансированных наборов данных

Набор данных	SVM	1-SVM	RF	SVM-RF	1-SVM-RF
SportsArticles 1000x59	84,11	63,85	85,11	85,6	86,68
Spam 4601x57	90,28	54,29	95,5	–	95,97
Biodeg 1055x41	88,08	57,38	91,73	–	92,07
WDBC 699x10	98,89	96,11	99	100	100

Результаты экспериментальных исследований SVM-RF-алгоритма в некоторых случаях не показали явных улучшений значения показателя сбалансированной точности (в таблице 2 эти случаи соответствуют наборам данных Spam, Biodeg), поскольку число объектов, используемых для обучения вспомогательного RF-классификатора, оказалось недостаточным для повышения качества SVM-классификации объектов, попавших в α-область. При этом 1-SVM-RF-алгоритм обеспечил повышение точности SVM-классификации для наборов данных Spam и Biodeg.

В таблице 3 приведены результаты многоклассовой классификации: 1) для многоклассовой SVM-классификации, основанной на применении декомпозиционных методов OvR и OvO (SVM_ovo, SVM_ovr); 2) RF-алгоритма и алгоритмов бустинга – AdaBoost-алгоритма и XGBoost-алгоритма (RF, AdaBoost, XGBoost); 3) алгоритмов на основе использования методов стекинга и блендинга (stacking, blending); 4) реализации разработанных ансамблевых методов многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2).

Таблица 3 – Результаты многоклассовой классификации

Набор данных	SVM_ovo	SVM_ovr	RF	AdaBoost	XGBoost	blending	stacking	DGM-1	DGM-2	IGM-1	IGM-2
BreastTissue 106x10	83,05	83,05	90,64	89,66	–	84,83	92,33	95,19	97,56	95,89	96,25
Segmentation 2310x19	92,92	92,92	93,26	88,59	–	95,84	95,50	97,30	96,90	96,23	97,07
Waveform 5000x21	85,13	85,13	85,53	–	85,66	84,89	85,47	86,05	85,90	86,14	85,96
Otto 6188x93	60,03	60,03	63,54	–	66,08	59,87	64,27	68,18	69,87	68,74	69,36

Результаты экспериментальных исследований, приведенные в таблице 3, показывают, что разработанные ансамблевые методы многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2) обеспечивают лучшие значения показателя сбалансированной точности по сравнению с многоклассовой SVM-классификацией на основе декомпозиционных методов OvR и OvO (SVM_ovo, SVM_ovr), многоклассовой классификацией с использованием методов бэггинга и бустинга (RF, AdaBoost, XGBoost), по сравнению с реализацией многоклассовой классификацией на основе методов стекинга и блендинга, являющимися аналогами предложенных в диссертации методов.

Приемлемое качество классификационных решений разработанных методов и алгоритмов позволяет применять их в качестве различных подходов к повышению точности базового классификатора (рисунок 5): 1) SVM-PSO-GS,

SVM-PSO-DOE-алгоритмы предлагается применять с целью нахождения оптимальных значений параметров SVM-классификатора; 2) SVM-bSMOTE-алгоритм при использовании сэмплинга для восстановления сбалансированности классов и необходимости нахождения оптимальных значений параметров bSMOTE-алгоритма в задаче SVM-классификации; 3) SVM-RF-алгоритм при необходимости сокращения числа ошибок SVM-классификации вблизи гиперплоскости, разделяющей классы; 4) 1-SVM-RF-алгоритм при необходимости повышения точности SVM-классификации несбалансированных наборов данных на основе поиска аномалий; 5) DGM-1, DGM-2, IGM-1, IGM-2-методы для повышения точности многоклассовой SVM-классификации.

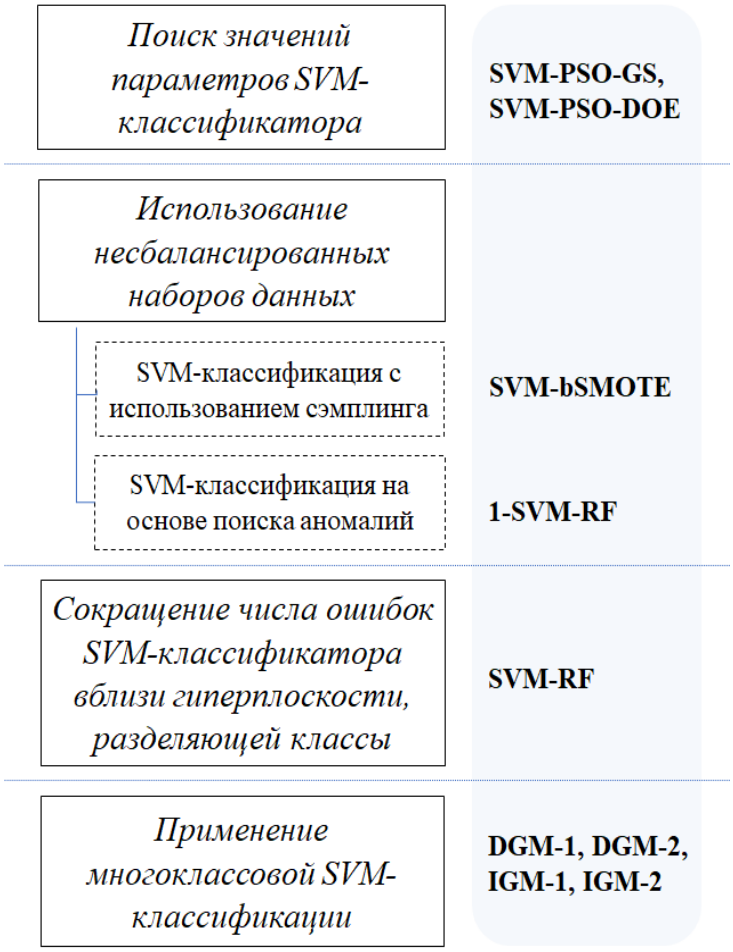

Рисунок 5 – Подходы к применению разработанных методов и алгоритмов

При этом разработанные алгоритмы, указанные в пунктах 1–4 выше, были апробированы при решении задач бинарной классификации, ансамблевые методы в пункте 5 – при решении задач многоклассовой классификации.

В заключении обобщены основные и практические результаты диссертационной работы.

Приложения включают копии свидетельств о государственной регистрации программы для ЭВМ, копии актов о внедрении результатов диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Основные результаты работы состоят в следующем:

1. Разработаны новые алгоритмы поиска оптимальных значений параметров базового классификатора (SVM-PSO-GS, SVM-PSO-DOE-алгоритмы), *отличающиеся* от аналогов совместным применением алгоритма роя частиц и алгоритмов поиска по сетке, *позволяющие* по сравнению с алгоритмом роя частиц сократить временные затраты на поиск оптимальных значений параметров базового классификатора в среднем в 3 раза при обеспечении высокой точности классификации данных.

2. Разработаны новые ансамблевые методы многоклассовой классификации (DGM-1, DGM-2, IGM-1, IGM-2-методы), *отличающиеся* от аналогов новыми подходами к генерации метахарактеристик при нескольких разбиениях обучающей выборки, *позволяющие* повысить точность многоклассовой классификации, реализуемой с использованием композиции бинарных классификаторов, до 5–7 %.

3. Разработан новый алгоритм поиска значений параметров алгоритма сэмплинга в задаче бинарной классификации с использованием несбалансированных наборов данных (SVM-bSMOTE-алгоритм), *отличающийся* от аналогов реализацией оценки пригодности комбинаций значений параметров алгоритма сэмплинга, *позволяющий* обеспечить разработку классификатора, точность классификации которого в среднем на 3 % превышает точность базового классификатора.

4. Разработан новый алгоритм двухэтапной классификации (SVM-RF-алгоритм), *отличающийся* от аналогов последовательным применением базового и вспомогательного ансамблевого классификаторов, *позволяющий* уменьшить число ошибок классификации вблизи гиперплоскости, разделяющей классы, и повысить точность бинарной классификации в среднем на 5 % по сравнению с базовым классификатором.

5. Разработан новый алгоритм двухэтапной классификации на основе поиска аномалий (1-SVM-RF-алгоритм), *отличающийся* от аналогов последовательным применением одноклассового классификатора в качестве основного и вспомогательного ансамблевого классификатора, *позволяющий* повысить точность бинарной классификации несбалансированных наборов данных в среднем на 2,5 % по сравнению с базовым классификатором.

Результаты экспериментальных исследований, характеризующиеся высокими значениями показателей качества классификации, подтверждают перспективность и целесообразность применения разработанных методов и алгоритмов при решении новых задач бинарной и многоклассовой классификации, включая случаи наличия различного соотношения классов объектов в исходных наборах данных (дисбаланс классов), необходимости поиска оптимальных значений параметров классификации за приемлемое время.

Приемлемое качество классификационных решений разработанных методов и алгоритмов позволяют найти применение в ряде востребованных

в настоящее время областях анализа данных: кредитный скоринг, обнаружение мошеннических транзакций, прогноз оттока клиентов, диагностика заболеваний в медицине, а также при классификации документов, обнаружении спама и др.

Актуальные направления дальнейших исследований могут быть связаны с разработкой новых подходов к построению ансамблей классификаторов, адаптацией разработанных алгоритмов бинарной классификации к решению задач многоклассовой классификации, применением результатов кластеризации при решении задач классификации.

Предложенные и апробированные в диссертационной работе методы и алгоритмы обеспечивают повышение точности классификации произвольных наборов данных, а также способны снижать недостатки существующих аналогов при решении задач классификации, что позволяет сделать вывод о достижении поставленной в диссертации цели исследования.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Работы, опубликованные в изданиях из перечня ВАК РФ:

1. Клюева, И. А. Разработка и исследование гибридных версий алгоритма роя частиц на основе алгоритмов поиска по сетке / Л. А. Демидова, И. А. Клюева // Вестник Рязанского государственного радиотехнического университета. – 2016. – № 57. – С. 105–116.

2. Клюева, И. А. Алгоритм подбора значений параметров bSMOTE-алгоритма в задаче SVM-классификации на основе несбалансированных наборов данных / Л. А. Демидова, И. А. Клюева // Вестник Рязанского государственного радиотехнического университета. – 2017. – № 61. – С. 67–77.

3. Клюева, И. А. Алгоритм случайного леса в задаче повышения качества SVM-классификации / Л. А. Демидова, И. А. Клюева // Вестник Рязанского государственного радиотехнического университета. – 2018. – № 65. – С. 74–83.

4. Клюева, И. А. Повышение качества многоклассовой SVM-классификации на основе feature engineering / И. А. Клюева // Cloud of science. – 2020. – Т. 7, № 1. – С. 207–218.

5. Клюева, И. А. Ансамблевые методы в задаче мнококлассовой SVM-классификации / Б. В. Костров, А. И. Баранчиков, И. А. Клюева // XXI век: итоги прошлого и проблемы настоящего плюс. – 2021. – Т. 10, № 2 (54). – С. 105–108.

6. Клюева, И. А. Использование методов очистки данных при реинжиниринге баз данных / А. И. Баранчиков, И. И. Яковлев, И. А. Клюева // Вестник Рязанского государственного радиотехнического университета. – 2021. – № 76. – С. 35–41.

Работы, опубликованные в изданиях, индексируемых в международных наукометрических базах:

7. Klyueva, I. The Study of Characteristics of the Hybrid Particle Swarm Algorithm in Solution of the Global Optimization Problem / L. Demidova, I. Klyueva,

A. Pylkin // 5th Mediterranean Conference on Embedded Computing (MECO-2016). – Bar (Montenegro), 2016. – P. 322–325. (Scopus)

8. Klyueva, I. SVM classification: optimization with the SMOTE algorithm for the class imbalance problem / L. Demidova, I. Klyueva // 6th Mediterranean Conference on Embedded Computing (MECO-2017). – Bar (Montenegro), 2017. – P. 472–475. (Scopus)

9. Klyueva, I. Intellectual Approaches to Improvement Of the Classification Decisions Quality On the Base Of the SVM Classifier / L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart // Procedia Computer Science. – 2017. – Vol. 103. – P. 222–230. (Scopus)

10. Klyueva, I. Data classification based on the hybrid versions of the particle swarm optimization algorithm / L. Demidova, I. Klyueva // 7th Mediterranean Conference on Embedded Computing (MECO-2018). – Budva (Montenegro), 2018. – P. 319–322. (Scopus)

11. Klyueva, I. Hybrid approach to improving the results of the SVM classification using the random forest algorithm / L. Demidova, I. Klyueva, A. Pylkin // Procedia Computer Science. – 2019. – Vol. 150. – P. 455–461. (Scopus)

12. Klyueva, I. Improving Quality of the Multiclass SVM Classification Based on the Feature Engineering / I. Klyueva // 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA-2019). – Lipetsk, 2019. – P. 491–494. (Scopus)

13. Klyueva, I. The two-stage classification based on 1-SVM and RF classifiers / L. Demidova, I. Klyueva // Journal of Physics: Conference Series. – 2020. – Vol. 1727. – P. 012007. (Scopus)

Работы, опубликованные в других изданиях:

14. Клюева, И. А. Подходы к модификации алгоритма роя частиц / И. А. Клюева // Информационные технологии в процессе подготовки современного специалиста : межвузовский сборник научных трудов. – Липецк : ЛГПУ, 2015. – С. 40–45.

15. Клюева, И. А. Гибридный алгоритм настройки параметров интеллектуального классификатора данных / И. А. Клюева // Математические методы в технике и технологиях. – 2015. – № 7 (77). – С. 234–237.

16. Клюева, И. А. Подбор параметров SVM-алгоритма с помощью алгоритма DOE в задаче классификации рисков ИТ-проектов / И. А. Клюева // Математическое и программное обеспечение вычислительных систем : межвузовский сборник научных трудов. – Москва : Научно-техническое издательство «Горячая линия – Телеком», 2015. – С. 48–52.

17. Клюева, И. А. Подбор параметров SVM-алгоритма с помощью алгоритма GRID SEARCH в задаче классификации рисков ИТ-проектов / И. А. Клюева // Проблемы передачи и обработки информации в сетях и системах телекоммуникаций : материалы 18-й Международной научно-технической

конференции. – Москва : Научно-техническое издательство «Горячая линия–Телеком», 2015. – С. 302–305.

18. Клюева, И. А. Исследование аспектов применимости стратегий сэмплинга для решения проблемы несбалансированности структур данных / И. А. Клюева // Новые информационные технологии в научных исследованиях : материалы XXI Всероссийской научно-технической конференции студентов, молодых ученых и специалистов (НИТ-2016). – Рязань : РГРТУ, 2016. – С. 198–200.

19. Клюева, И. А. Исследование характеристик сходимости алгоритма роя частиц и его модификации в решении задачи глобальной оптимизации / И. А. Клюева // Современные технологии в науке и образовании : сборник трудов Международной научно-технической и научно-методической конференции (СТНО-2016). – Рязань : РГРТУ, 2016. – Т. 2. – С. 46–50.

20. Клюева, И. А. Обзор многоцелевых биоинспирированных алгоритмов оптимизации / И. А. Клюева // Прикладные исследования и технологии : сборник трудов III Международной конференции молодых ученых, студентов и магистрантов (ART2016). – Москва : МТИ, 2016. – С. 129–132.

21. Клюева, И. А. Повышение качества классификационных решений SVM-классификатора на основе *k*NN-алгоритма / Л. А. Демидова, Ю. С. Соколова, И. А. Клюева // Интеллектуальные и информационные системы : материалы Всероссийской научно-технической конференции (Интеллект-2016). – Тула : ТГУ, 2016. – С. 55–58.

22. Клюева, И. А. Исследование применимости SMOTE-алгоритма при классификации несбалансированных данных / И. А. Клюева // Современные технологии в науке и образовании : сборник трудов II Международной научно-технической и научно-методической конференции (СТНО-2017). – Рязань : РГРТУ, 2017. – С. 143–146.

23. Клюева, И. А. Оптимизация поиска параметров SMOTE алгоритма в задаче SVM классификации несбалансированных наборов данных / И. А. Клюева // Математические методы и информационные технологии управления в науке, образовании и правоохранительной сфере : сборник материалов Всероссийской научно-технической конференции. – Рязань : Академия права и управления ФСИН, 2017. – С. 242–245.

24. Клюева, И. А. Разработка SVM-классификатора с применением гибридных версий алгоритма роя частиц на основе поиска по сетке / Л. А. Демидова, И. А. Клюева // Cloud of science. – 2017. – Т. 3, № 4. – С. 528–547.

25. Клюева, И. А. SVM-классификация на основе несбалансированных наборов данных в среде Python. Решение задач психодиагностики / Л. А. Демидова, И. А. Клюева // Прикладные исследования и технологии : сборник трудов IV Международной конференции молодых ученых, студентов и магистрантов (ART2017). – Москва : МТИ, 2017. – С. 129–135.

26. Klyueva, I. Development of the SVM classifier by means of the hybrid versions of the particle swarm optimization algorithm based on the GRID SEARCH /

L. Demidova, I. Klyueva // Computer modelling and new technologies. – 2017. – P. 56–63.

27. Klyueva, I. Improving the Classification Quality of the SVM Classifier for the Imbalanced Datasets on the Base of Ideas the SMOTE Algorithm / L. Demidova, I. Klyueva // ITM Web of Conferences. – Moscow, 2017. – Vol. 10. – P. 10–14.

28. Клюева, И. А. Исследование аспектов разведочного анализа данных / И. А. Клюева // Новые информационные технологии в научных исследованиях : материалы XXIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов (НИТ-2018). – Рязань : РГРТУ, 2018. – С. 153–154.

29. Клюева, И. А. Подход к классификации длительности ИТ-проектов / Л. А. Демидова, И. А. Клюева // Прикладные исследования и технологии : сборник трудов V Международной конференции молодых ученых, студентов и магистрантов (ART2018). – Москва : МТИ, 2018. – С. 39–43.

30. Клюева, И. А. Области применения интеллектуального анализа данных / И. А. Клюева // Инновационные научные исследования. – 2021. – № 7-1 (9). – С. 173–194.

31. Клюева, И. А. Современные возможности и примеры внедрения машинного обучения / И. А. Клюева // Оригинальные исследования (ОРИС). – 2021. – Т. 11, № 7. – С. 12–32.

Свидетельства о регистрации программ для ЭВМ

32. Свидетельство о государственной регистрации программы для ЭВМ 2016619800. SVM-классификатор на основе гибридных версий алгоритма роя частиц / Л. А. Демидова, И. А. Клюева; правообладатель ФГБОУ ВО «Рязанский государственный радиотехнический университет» ; заявл. 15.07.2016 ; зарег. 30.08.2016.

33. Свидетельство о государственной регистрации программы для ЭВМ 2017661030. Программный модуль подбора значений параметров алгоритма синтетического сэмплинга в задаче SVM-классификации несбалансированных наборов данных / Л. А. Демидова, И. А. Клюева ; правообладатель ФГБОУ ВО «Рязанский государственный радиотехнический университет» ; заявл. 11.08.2017 ; зарег. 02.10.2017.

34. Свидетельство о государственной регистрации программы для ЭВМ 2018666401. Программный модуль повышения качества SVM-классификации посредством гибридизации с алгоритмом случайного леса / Л. А. Демидова, И. А. Клюева ; правообладатель ФГБОУ ВО «Рязанский государственный радиотехнический университет» ; заявл. 06.12.2018 ; зарег. 17.12.2018.

Научное издание

Клюева Ирина Алексеевна

МЕТОДЫ И АЛГОРИТМЫ АНСАМБЛИРОВАНИЯ И ПОИСКА ЗНАЧЕНИЙ ПАРАМЕТРОВ КЛАССИФИКАТОРОВ

Специальность 05.13.17 – Теоретические основы информатики (технические науки)

Редактор *В. В. Чувашова*
Технический редактор *М. Б. Жучкова*
Компьютерная верстка *М. Б. Жучковой*

Распоряжение № 72/209 от 17.09.2021

Подписано в печать 20.09.2021.
Формат 60×84 1/16. Усл. печ. л. 1,4.
Тираж 100. Заказ № 453.

Издательство ПГУ.
440026, Пенза, Красная, 40.
Тел.: (8412) 66-60-49, 66-67-77; e-mail: iic@pnzgu.ru