На правах рукописи

Бардамова Марина Борисовна

АЛГОРИТМЫ ПОСТРОЕНИЯ НЕЧЕТКИХ КЛАССИФИКАТОРОВ НЕСБАЛАНСИРОВАННЫХ ДАННЫХ НА ОСНОВЕ МЕТАЭВРИСТИК «ГРАВИТАЦИОННЫЙ ПОИСК» И «ПРЫГАЮЩИЕ ЛЯГУШКИ»

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Томск – 2021

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Томский государственный университет систем управления и радиоэлектроники» (ТУСУР).

Научный руководитель –	**Ходашинский Илья Александрович**, доктор технических наук, профессор
Официальные оппоненты:	**Пимонов Александр Григорьевич**, доктор технических наук, профессор, заведующий кафедрой прикладных информационных технологий, ФГБОУ ВО «Кузбасский государственный технический университет им. Т.Ф. Горбачева» (г. Кемерово) **Гергет Ольга Михайловна**, доктор технических наук, доцент, профессор отделения информационных технологий, ФГАОУ ВО «Национальный исследовательский Томский политехнический университет»
Ведущая организация –	ФГБОУ ВО «Белгородский государственный технологический университет имени В.Г. Шухова»

Защита состоится «9» декабря 2021 г. в 15 часов 15 минут на заседании диссертационного совета Д212.268.05 ТУСУРа по адресу: 634050, г. Томск, пр. Ленина, 40, ауд. 201.

С диссертацией можно ознакомиться на официальном сайте https://postgraduate.tusur.ru/urls/zx4mrmzt и в библиотеке ТУСУРа по адресу: 634045, г. Томск, ул. Красноармейская, 146.

Автореферат разослан __ октября 2021 г.

Ученый секретарь
диссертационного совета  Костюченко Евгений Юрьевич

Общая характеристика работы

Актуальность темы. Машинное обучение применяется для построения автоматических систем анализа данных, которые позволяют ускорить и облегчить работу специалистов в различных сферах: медицине, информационной безопасности, экономике и других. Однако важным условием эффективного взаимодействия между интеллектуальной системой и её пользователем является доверие, которое достигается не только уверенностью в правильности результата автоматического анализа, но и в понимании, какие процессы внутри системы привели к этому результату. Системы нечеткого вывода отличаются от прочих алгоритмов машинного обучения тем, что в их основе лежат принципы человеческого мышления и логики. Нечеткие правила и функции принадлежности легко поддаются интерпретации, позволяя обеспечить понимание пользователем системы без глубокого погружения в специфику машинного обучения.

Однако при построении систем нечеткого вывода с целью решения задач классификации могут возникнуть трудности при работе с несбалансированными данными. Нечеткие классификаторы подвержены переобучению на классах с наибольшим числом экземпляров, что ведет к получению высокой общей точности при низкой доле правильной классификации объектов, принадлежащих наименьшим классам. Так как именно редкие классы зачастую оказываются наиболее важными для прогноза, требуются подходы, способные улучшить качество их распознавания. Задача создания алгоритмов построения нечетких классификаторов, позволяющих построить точные, компактные и интерпретируемые модели на несбалансированных данных, является актуальной.

Анализ существующих подходов по улучшению точности нечетких классификаторов несбалансированных данных показывает, что основным методом преодоления дисбаланса является применение предобработки данных. Классы меньшинства дополняются путем генерации искусственных экземпляров, что облегчает процесс обучения классификатора. Но использование таких алгоритмов при наличии шумов в данных ведет к повторению ошибок в новых образцах. Кроме того, добавление данных сложно использовать при количестве классов, большем двух, или в условиях крайне малого исходного числа экземпляров наименьших классов.

Использование для повышения точности нечетких классификаторов таких этапов обучения, как формирование структуры, настройка параметров и отбор признаков, является устоявшейся практикой. Их эффективность многократно подтверждена публикациями P. Angelov, V. Bolon-Canedo, S.L. Chiu, O. Cordon, A. Fernandez, H. Hagras, F. Herrera, H. Ishibuchi, M.J. del Jesus, V. Lopez, M. Sugeno, T. Takagi, L. Xu, R.R. Yager. Внесение модификаций в эти этапы может позволить нечеткому классификатору достигать высокой точности на несбалансированных данных.

Перечисленные задачи обучения классификатора могут быть решены с помощью метаэвристических алгоритмов. Метаэвристики – это класс алгоритмов, осуществляющих поиск удовлетворительных решений задач оптимизации без доказательства оптимальности найденных вариантов. Качество решения может быть выражено через некоторую метрику, например точность, стабильность, время. Применение метаэвристик с соответствующей задаче фитнес-функцией позволит достигнуть улучшения качества классификации несбалансированных данных с помощью нечетких систем без использования этапа предобработки данных. В качестве такой функции выбрана средняя геометрическая точность, учитывающая долю правильной классификации каждого класса.

Кроме упомянутых выше ученых, наиболее значимых результатов в изучении нечетких систем достигли А.Н. Аверкин, И.З. Батыршин, М.В. Бобырь, М.И. Дли, Ю.Н. Золотухин, А.С. Катасёв, С.М. Ковалев, Л.Г. Комарцова, В.В. Круглов, Ю.И. Кудинов, А.О. Недосекин, Ф.Ф. Пащенко, Д.А. Поспелов, Ю.П. Пытьев, Е.С. Семенкин, А.В. Язенин, Н.Г. Ярушкина, Г.Э. Яхъева, R. Babuska, A. Bastian, J.C. Bezdek, J. Casillas, J.L. Castro, D. Dubois, D. Filev, J. Gonzalez, S. Guillaume, U. Kaymak, B. Kosko, R. Krishnapuram, R. Kruse, E.H. Mamdani, S. Oh, W. Pedrycz, H. Prade, H. Tanaka, I. B. Turksen, T. Yasukawa, L. Zadeh.

Целью диссертационной работы является повышение средней геометрической точности нечетких классификаторов несбалансированных данных за счет использования метаэвристических алгоритмов на различных этапах построения классификатора.

Для достижения поставленной цели поставлены следующие задачи:

1) обзор существующих методов обработки несбалансированных данных и методов построения систем нечеткого вывода;

2) разработка и исследование алгоритма формирования структуры нечеткого классификатора, позволяющего улучшить среднюю геометрическую точность;

3) разработка и исследование гибридного алгоритма оптимизации параметров нечеткого классификатора несбалансированных данных;

4) разработка и исследование алгоритма настройки весовых коэффициентов, учитывающих важность признаков в базе нечетких правил;

5) проверка разработанных алгоритмов на контрольных примерах и сравнение с существующими аналогами.

Объектом исследования является процесс построения нечетких классификаторов несбалансированных данных.

Предметом исследования являются алгоритмы построения и оптимизации нечетких классификаторов для несбалансированных данных.

Методы исследования. В диссертационной работе применялись методы оптимизации, анализа данных и теории информации, а также теория нечетких множеств и нечеткой логики.

Достоверность результатов обеспечивается корректностью применения математических методов, результатами проведенных экспериментов, статистически сопоставимых с результатами, полученными исследователями других научных групп.

Научная новизна полученных результатов. В диссертации получены следующие новые научные результаты.

1. Разработан авторский алгоритм формирования базы правил нечеткого классификатора несбалансированных данных, отличительной особенностью которого является применение метаэвристики "прыгающие лягушки" для итерационно повторяющейся процедуры генерации и настройки дополнительного правила для класса с наименьшей долей правильной классификации.

2. Разработан новый гибридный алгоритм для оптимизации параметров нечетких классификаторов несбалансированных данных, особенность которого заключается в дополнении метаэвристики «гравитационный поиск» локальным поиском из метаэвристики «прыгающие лягушки» для улучшения эффективности оптимизации.

3. Разработан авторский алгоритм настройки весовых коэффициентов признаков при классификации несбалансированных данных, отличительной особенностью которого является применение гибридного метаэвристического алгоритма для поиска оптимального вектора весов признаков в базе нечетких правил.

Теоретическая значимость работы заключается в развитии технологии построения нечетких систем интеллектуального анализа несбалансированных данных. Алгоритм формирования базы правил нечеткого классификатора и алгоритм настройки весов признаков могут использовать любые аналогичные метаэвристики вместо предложенных. Гибридный алгоритм оптимизации может применяться для решения других задач параметрической оптимизации.

Практическая значимость работы подтверждается применением полученных в ней результатов для решения практической задачи оценки свертываемости крови у беременных женщин. Результаты внедрены в ОГАУЗ «Родильный дом №1» города Томска.

Разработанные алгоритмы использованы при выполнении следующих проектов:

– научный проект при поддержке РФФИ «Методы и инструментальные средства построения самообучающихся систем, основанных на нечетких правилах» (№16-07-00034-а), 2016-2018 гг. (№ госрегистрации АААА-А16-116021210312-4);

– научный проект при поддержке РФФИ «Методы построения нечетких классификаторов несбалансированных данных на основе алгоритма гравитационного поиска» (№19-37-90064-аспиранты), 2019-2021 гг. (№ госрегистрации АААА-А19-119101790046-5);

– государственное задание Министерства образования и науки Российской Федерации на 2017–2019 гг., проект № 2.8172.2017/БЧ «Методы и модели определения уровня защищенности информационных систем» (№ госрегистрации АААА-А17-117073110015-3);

– государственное задание Министерства образования и науки Российской Федерации на 2017-2019 гг., проект № 8.9628.2017/8.9 «Теоретические основы человеко-машинных интерфейсов» (№ госрегистрации АААА-А17-117073110013-9).

Разработанные алгоритмы применимы при построении нечетких классификаторов для решения практических задач и в научно-исследовательских целях при анализе данных.

На защиту выносятся следующие положения.

1. Разработанный алгоритм формирования базы нечетких правил, основанный на итеративном процессе генерации и настройки правила метаэвристикой «прыгающие лягушки», в комбинации с алгоритмом генерации структуры на основе экстремумов признаков классов[1] позволяет создавать классификатор, демонстрирующий при меньшем числе правил большую среднюю геометрическую точность по сравнению с классификаторами, полученными общеизвестными алгоритмами генерации структуры Ishibuchi+SMOTE[2] и E-алгоритмом[2], а также сопоставимую точность при сравнении с комбинациями Chi+SMOTE[2] и HFRBCS+SMOTE[2]. На исследуемых несбалансированных наборах данных средняя геометрическая точность возросла в среднем на 23 процента относительно точности, полученной при использовании только алгоритма экстремальных значений признаков классов.

Соответствует пункту 5 паспорта специальности: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях.

2. Разработанный гибридный алгоритм настройки параметров нечеткого классификатора на основе комбинации метаэвристик «гравитационный поиск» и «прыгающие лягушки» позволил увеличить среднюю геометрическую точность классификации на исследуемых несбалансированных наборах данных в среднем на 24 процента по сравнению с точностью до оптимизации. Статистическое сравнение подтвердило существование значимой разницы в точности по сравнению с исходными метаэвристиками при оптимизации параметров нечетких классификаторов несбалансированных данных. Построенные нечеткие классификаторы продемонстрировали большую среднюю геометрическую точность по сравнению с Chi+SMOTE,

[1] Алгоритмы структурной идентификации компактных и точных нечетких систем / И. А. Ходашинский, И. В. Горбунов, К. С. Сарин, С. Р. Субханкулова // Информационные и математические технологии в науке и управлении. – 2016. – № 1. – С. 82-93.

[2] Fernandez, A. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets / A. Fernandez; M. J. Del Jesus; F. Herrera // International Journal of Approximate Reasoning. – 2009. – Vol. 50. – P. 561–577.

Ishibuchi+SMOTE и E-алгоритмом, и сопоставимое качество классификации при сравнении с HFRBCS+SMOTE.

Соответствует пункту 13 паспорта специальности: Применение бионических принципов, методов и моделей в информационных технологиях.

3. Разработанный алгоритм настройки весовых коэффициентов признаков позволил увеличить среднюю геометрическую точность классификации в среднем на 16 процентов относительно точности до введения весов. При существенно меньшем количестве используемых правил алгоритм позволил продемонстрировать сопоставимую точность с комбинациями Chi+SMOTE и Ishibuchi+SMOTE и большую точность по сравнению с E-алгоритмом.

Соответствует пункту 5 паспорта специальности: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях.

Внедрение результатов диссертационного исследования.

Результаты исследовательской работы легли в основу программного обеспечения для оценки состояния свертывающей системы крови у беременных женщин, используемого в ОГАУЗ «Родильный дом №1». Разработанные алгоритмы были использованы в ФГБОУ ВО «ТУСУР» при выполнении проекта № 8.9628.2017/8.9 «Теоретические основы человеко-машинных интерфейсов» в рамках государственного задания Министерства науки и высшего образования РФ и в проекте № 2.8172.2017/8.9 «Методы и модели определения уровня защищенности информационных систем» в процессе исполнения государственного задания ТУСУР. Результаты исследования используются при изучении дисциплины «Информатика» на кафедре комплексной информационной безопасности электронно-вычислительных систем ТУСУР.

Апробация работы. Основные положения работы докладывались и обсуждались на конференциях различного уровня. Среди них:

– международная конференция IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) (2021, онлайн, IEEE);

– международные научно-практические конференции «Электронные средства и системы управления» (2015, 2017-2020 гг., Томск, ТУСУР);

– международные научно-технические конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР» (2015-2021 гг., Томск, ТУСУР);

– международные конференции студентов, аспирантов и молодых ученых «Перспективы развития фундаментальных наук» (2018-2021 гг., Томск, ТУСУР);

– всероссийские молодежные научные форума «Наука будущего – наука молодых» (12-14 сентября 2017 г., Нижний Новгород; 2019 гг.; 14-17 мая 2019, Сочи, Министерство науки и высшего образования РФ);

– всероссийский конкурс-конференция студентов и аспирантов по информационной безопасности «SIBINFO-2018» (19 апреля 2018 г., Томск, ТУСУР);

– всероссийский форум молодых ученых (27-28 апреля 2017 г., Екатеринбург, УрФУ);

– всероссийская научно-практическая конференция «Нечеткие системы, мягкие вычисления и интеллектуальные технологии» (3–7 июля 2017 г., Санкт-Петербург, СПИИРАН);

– международная летняя школа-семинар по искусственному интеллекту «Интеллектуальные системы и технологии: современное состояние и перспективы» (30 июня – 3 июля 2017 г., Санкт- Петербург, СПИИРАН);

– международной научно-практической конференции «Молодежь и современные информационные технологии» (7-11 ноября 2016 г., Томск, ТПУ);

– всероссийская научно-практическая конференция в рамках конгресса «Здравоохранение России. Технологии опережающего развития» (4-7 ноября 2015 г., Томск, Министерство здравоохранения РФ).

Публикации по теме диссертации. По результатам исследований опубликовано 28 печатных работ, из которых в рекомендованных ВАК РФ периодических изданиях – 6. Десять работ проиндексированы в международной базе SCOPUS, четыре – в Web of Science. Получены 4 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Постановка цели и задач научного исследования, интерпретация экспериментальных данных, подготовка публикаций по выполненной работе проводилась совместно с научным руководителем. Автором самостоятельно разработаны и реализованы алгоритмы формирования структуры нечеткого классификатора несбалансированных данных, настройки весовых коэффициентов признаков, настройки параметров термов на основе комбинации двух метаэвристик; получены результаты экспериментов, проведена апробация разработанных алгоритмов. Разработка программного обеспечения для ОГАУЗ «Родильный дом №1» проведена автором совместно с сотрудниками родильного дома.

Объем и структура работы. Диссертационная работа состоит из введения, четырех глав основной части, заключения, списка литературы из 155 наименований и 4 приложений. Основная часть работы содержит 116 страниц, в том числе 14 рисунков и 39 таблиц.

Основное содержание работы

Во введении описана актуальность работы, сформулированы цель и задачи исследования, изложены основные результаты, их теоретическая и практическая значимость, приведена новизна исследования и защищаемые положения.

В первой главе содержится обзор проблемы построения интеллектуальных систем при несбалансированном характере исследуемых данных. Приведен и проанализирован перечень

типовых методов преодоления проблемы дисбаланса данных. Представлен обзор основных алгоритмов построения и оптимизации нечетких классификаторов.

Постановка задачи. Цель классификации заключается в определении для входного объекта **x** из множества объектов $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|X|}\}$ наиболее подходящего класса из множества классов $C = \{c_1, c_2, \ldots, c_m\}$. Каждый объект описывается вектором значений признаков $\mathbf{x}_p = (x_p^1, x_p^2, \ldots, x_p^n)$, где x_p^i – значение i-го признака объекта $\mathbf{x}_p$ ($i = \overline{1,n}$), $p = \overline{1,|X|}$.

В нечетком классификаторе вывод о принадлежности объекта классу строится на основе степени принадлежности этого объекта к правилам. База правил нечеткого классификатора состоит из перечня утверждений следующего образца:

$$\text{ЕСЛИ } x^1 = T_{1j}, \text{ И } x^2 = T_{2j}, \text{ И } \ldots, \text{ И } x^n = T_{ni}, \text{ ТО class} = c_k, \quad (1)$$

где T_{ij} – нечеткий терм, характеризующий признак x^i в j-ом правиле ($j = \overline{1,R}$), R – количество правил в базе, c_k – метка k-го класса ($k = \overline{1,m}$). Нечеткие термы описывают входные переменные и могут представлять собой различные функции принадлежности: треугольного, трапециевидного, гауссова типа. Последовательность параметров термов всех признаков составляет вектор антецедентов **θ**. Для термов гауссова типа вектор **θ** составляется путем последовательного перечисления двух параметров – координаты вершины по оси абсцисс a_{ij} и стандартного отклонения терма b_{ij}: $\boldsymbol{\theta} = (a_{11}, b_{11}, \ldots, a_{1R}, b_{1R}, a_{21}, b_{21}, \ldots, a_{nR}, b_{nR})$.

Количество и расположение термов, а также число и содержание правил определяет алгоритм генерации структуры. После завершения этапа генерации структуры осуществляется процедура вывода и оценка качества построенного классификатора. Для поступившего на вход классификатора объекта $\mathbf{x}_p$ вычисляется степень его принадлежности каждому классу:

$$\beta_k(\mathbf{x}_p) = \sum_{r_{jk}} \prod_{i=1}^{n} \mu_{T_{ij}}(x_p^i), \quad (2)$$

где r_{jk} – правила с выходной меткой класса k, $\mu_{T_{ij}}(x_p^i)$ – значение функции принадлежности терма T_{ij} в точке x_p^i. Выходом является класс с наибольшей степенью принадлежности:

$$class = c_{k^*}, \quad k^* = \arg\max_{1 \le k \le m} \beta_k(\mathbf{x}_p). \quad (3)$$

При построении классификаторов, обрабатывающих несбалансированные данные, важно выбрать адекватный критерий оценки полученной модели. В данной работе рассмотрены наборы данных, насчитывающие два класса, так как любую задачу классификации можно разделить на подзадачи вида «один класс против остальных». Наиболее объективным показателем качества является доля правильной классификации на каждом отдельном классе. В качестве объединенной метрики использована средняя геометрическая точность *GM*:

$$GM = \sqrt[m]{\prod_{k=1}^{m} (inst_k^* / inst_k)}\,, \tag{4}$$

где $inst_k$* – количество правильно определенных экземпляров k-го класса, $inst_k$ – количество всех экземпляров k-го класса, $k \in \overline{1,m}$. Чем меньше экземпляров некоторого класса представлено в данных, тем больший вклад в среднюю геометрическую точность будет вносить правильно определенный экземпляр этого класса. Цель обучения классификатора заключается в поиске максимума выбранной целевой функции.

Во второй главе приведено описание разработанного алгоритма формирования структуры нечеткого классификатора несбалансированных данных, гибридного алгоритма настройки параметров термов, алгоритма настройки весовых коэффициентов признаков.

Алгоритм формирования структуры нечеткого классификатора на основе метаэвристики «прыгающие лягушки». Алгоритм предназначен для итерационного процесса генерации и настройки новых правил для базы правил нечеткого классификатора, созданной любым алгоритмом генерации. Входными данными является текущая база правил *Base*, количество добавляемых правил *NR* и параметры метаэвристики: *GIt* – количество глобальных итераций, *LIt* – количество локальных итераций, *NM* – число мемплексов, *NA* – число агентов в мемплексе (подмножестве векторов), *const* – константа для генерации новых параметров термов, γ – коэффициент приоритета метрики в фитнес-функции.

Одна итерация создания и настройки правила состоит в следующей последовательности действий. На основе текущей базы правил *Base* осуществляется расчет доли правильной классификации каждого класса в отдельности, выбирается класс с наихудшим показателем. Формируется популяция векторов (агентов), каждый из которых представляет собой вариацию нового правила и состоит из перечня параметров термов и консеквента выбранного класса. Термы генерируются на основе экстремальных значений признаков с некоторым отклонением. Размер популяции *N* равняется произведению *NA* и *NM*.

Далее запускается глобальный поиск, в котором популяция сортируется по убыванию значения фитнес-функции, после чего счетчик глобальных итераций увеличивается на единицу. Фитнес-функция отражает улучшение классификации при добавлении нового правила **R*** к исходной базе правил по сравнению с качеством классификации, полученным только на текущей базе:

$$fit(Base \cup \mathbf{R}^*) = score(Base \cup \mathbf{R}^*) - score\left(Base\right), \tag{7}$$

где $score(\bullet)$ – комбинация средней геометрической точности *GM* и общей точности *Acc*:

$$score = \gamma \times GM + (1-\gamma) \times Acc\,. \tag{8}$$

Коэффициент γ, принадлежащий промежутку [0;1], управляет приоритетом между двумя метриками. Общая точность рассчитывается как сумма правильно определенных классификатором экземпляров к общему количеству экземпляров:

$$Acc = \sum_{k=1}^{m} inst_k^* \Big/ \sum_{k=1}^{m} inst_k \,. \tag{9}$$

После сортировки агенты популяции последовательно разбиваются на подгруппы – мемплексы, внутри которых независимо проводится локальный поиск заданное количество локальных итераций. В каждом мемплексе выбираются **best** и **worst** – векторы с лучшей и худшей фитнес-функцией. На основе выбранных векторов генерируется новое правило:

$$\mathbf{new} = \mathrm{rand} \times const \times (\mathbf{best} - \mathbf{worst}) + \mathbf{worst} \,. \tag{10}$$

Если фитнес-функция созданного вектора оказывается лучше, чем у **worst**, то **worst** заменяется на **new**. В противном случае генерация происходит повторно, но на этот раз в **best** помещается глобально лучший агент (первый в популяции). Если и в этом случае не удалось улучшить вектор **worst**, то на его место записывается вектор, сгенерированный путем наложения случайного отклонения на глобально лучший агент.

Когда локальные итерации истекают, алгоритм возвращается к глобальному поиску. После завершения глобальных итераций весь процесс повторяется заново, пока не будут добавлены все *NR* правил. Выходом алгоритма является дополненная база правил нечеткого классификатора. Далее приведен псевдокод алгоритма.

Вход: *Base*, *NR*, *GIt*, *LIt*, *NM*, *NA*, *const*, γ.

Выход: *Base*.

цикл пока (*NR* > 0):

Определение худшего класса C_{worst}.

Генерация популяции $\{\mathbf{R}_1^*, \mathbf{R}_2^*, \ldots, \mathbf{R}_{NM \times NA}^*\}$ для C_{worst}.

Вычисление фитнес-функции $fit(Base \cup \mathbf{R}^*)$ по формуле (7) для каждого $\mathbf{R}^*$

цикл пока (*GIt* > 0):

Сортировка популяции по убыванию фитнес-функции.

Последовательное разбиение популяции на *NM* групп.

цикл по *NM*:

цикл пока (*LIt* > 0):

Локальный поиск с использованием оператора (10).

LIt := *LIt* – 1.

конец цикла.

конец цикла.

GIt := *GIt* – 1.

конец цикла.

Добавление $\mathbf{R}^*$ с максимальной фитнес-функцией в *Base*.

NR := *NR* – 1.

конец цикла.

вывод *Base*.

Гибридный алгоритм оптимизации параметров нечетких термов на основе комбинации метаэвристик «гравитационный поиск» и «прыгающие лягушки». Так как «гравитационный поиск» представляет собой в большей степени глобальный поиск, а в «прыгающих лягушках» существеннее проработан локальный поиск, то их комбинация позволит повысить эффективность оптимизации. Подробно предлагаемый гибрид описан далее.

Входными параметрами являются: вектор параметров антецедентов $\boldsymbol{\theta}_0$, количество итераций глобального и локального поиска *GIt* и *LIt* соответственно, количество агентов в мемплексе *NA*, количество мемплексов *NM*, начальное значение гравитационной постоянной G_0, коэффициент уменьшения α, константы ε и *const*, число лучших агентов K_{best}. На основе исходного агента $\boldsymbol{\theta}_0$ создается популяция частиц $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N-1}\}$, где $N = NA \times NM$. Популяция сортируется по убыванию значения фитнес-функции.

Глобальный поиск можно представить в виде следующей последовательности шагов. На первом шаге оцениваются массы векторов:

$$mass_i(t) = \frac{fit(\boldsymbol{\theta}_i(t)) - fit(\boldsymbol{\theta}_{worst}(t))}{fit(\boldsymbol{\theta}_{best}(t)) - fit(\boldsymbol{\theta}_{worst}(t))}, \tag{11}$$

$$M_i(t) = mass_i(t) \Big/ \sum_{j=0}^{N-1} mass_j(t), \tag{12}$$

где $M_i(t)$ – масса *i*-го агента на текущей итерации *t*, $i \in [0, N\text{-}1]$, $t \in [1, GIt]$, $fit(\boldsymbol{\theta}_i(t))$ – значение фитнес-функции *i*-го агента, $fit(\boldsymbol{\theta}_{worst}(t))$ и $fit(\boldsymbol{\theta}_{best}(t))$ – значение фитнес-функции худшего и лучшего вектора. Равнодействующая сил тяготения между агентом и K_{best} лучшими агентами сообщает агенту ускорение. Благодаря сортировке, K_{best} лучших частиц – это вектора с индексами от 0 до K_{best}-1. Ускорение каждого *d*-го элемента *i*-го агента рассчитывается на втором шаге глобального поиска:

$$a_i^d(t) = G(t) \times \sum_{j=0,\, j \neq i,\, j \in K_{\text{best}}}^{N} \text{rand}(0,1) \times \frac{M_j \times (\theta_j^d(t) - \theta_i^d(t))}{\left\| \boldsymbol{\theta}_j(t) - \boldsymbol{\theta}_i(t) \right\| + \varepsilon}, \tag{13}$$

где $G(t)$ – значение гравитационной постоянной, обновляющееся на каждой итерации:

$$G(t) = G_0 \times \exp(-\alpha \times t / GIt). \tag{14}$$

На третьем шаге определяется скорость как сумма набранного ускорения и случайной компоненты текущей скорости, а также происходит обновление элементов векторов:

$$V_i^d(t+1) = \text{rand}(0,1) \times V_i^d(t) + a_i^d(t), \quad (15)$$

$$\theta_i^d(t+1) = \theta_i^d(t) + V_i^d(t+1), \quad (16)$$

где $V_i^d(t)$ - текущая скорость, равная на первой итерации нулю.

Далее происходит перерасчет фитнес-функции, сортировка агентов и выполнение заданное количество итераций локального поиска из «прыгающих лягушек».

Локальный поиск заключается в замене худших векторов в мемплексе на новые, при этом фактическое разбиение на подгруппы не проводится, принадлежность к тому или иному мемплексу определяется по индексу агента. Для того, чтобы не заменять постоянно один и тот же вектор, вводится счетчик замены *f*. Счетчик увеличивается на единицу каждый раз, когда происходит замена на новый вектор, кроме замены случайным образом, и уменьшается до единицы каждый раз, когда достигает значения, равного *NA*-1. Для создания нового вектора **new** на каждой локальной итерации t_l ($t_l \in [1, LIt]$) выбираются два агента из одного и того же мемплекса *mem* ($mem \in [0, NM]$): вектор с индексом *mem* записывается в $\mathbf{best}(t_l)$, вектор с индексом *wr* ($wr = N - f \times NM + k$) записывается в $\mathbf{worst}(t_l)$. Новый агент **new** генерируется на основе следующего оператора:

$$\mathbf{new} = \text{rand}(0,1) \times const \times \left(\mathbf{best}(t_l) - \mathbf{worst}(t_l)\right) + \mathbf{worst}(t_l). \quad (17)$$

Для **new** оценивается значение фитнес-функции; если оно больше, чем у $\boldsymbol{\theta}_{wr}(t_l)$, то агент **new** заменяет $\boldsymbol{\theta}_{wr}(t_l)$, а вектор скорости V_{wr} обнуляется. В противном случае **new** создается заново, но в $\mathbf{best}(t_l)$ записывается глобально лучший вектор $\boldsymbol{\theta}_0$. Если фитнес-функция **new** по-прежнему не превышает $fit(\boldsymbol{\theta}_{wr}(t_l))$, то на месте $\boldsymbol{\theta}_{wr}(t_l)$ генерируется новый вектор на основе $\boldsymbol{\theta}_0$ с некоторым отклонением. После истечения локальных итераций осуществляется возврат к глобальному поиску. Выходом алгоритма является агент с максимальным значением фитнес-функции.

Псевдокод алгоритма представлен ниже.

Вход: $\boldsymbol{\theta}_0$, *GIt*, *LIt*, *NM*, *NA*, G_0, α, ε, *const*.

Выход: $\boldsymbol{\theta}_0$.

Генерация популяции $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{NM \times NA-1}\}$.

Вычисление фитнес-функции, сортировка популяции по убыванию фитнес-функции.

$t := 0$.

цикл пока ($t < GIt$):

Обновление $G(t)$ по формуле (14).

Расчет вектора масс $\mathbf{M}(t)$ по формулам (11), (12).

Вычисление ускорения $\mathbf{a}(t)$ на основе выражения (13).

Расчет скорости $\mathbf{V}(t)$ по формуле (15).

Обновление векторов популяции $\boldsymbol{\Theta}$ согласно выражению (16).

Расчет фитнес-функции, сортировка популяции по убыванию фитнес-функции.

цикл пока ($LIt > 0$):

Локальный поиск на основе формулы (17).

$LIt := LIt - 1$.

конец цикла.

$t := t + 1$.

конец цикла.

Сортировка популяции по убыванию фитнес-функции.

вывод $\boldsymbol{\theta}_0$.

Алгоритм настройки весовых коэффициентов признаков. Идея алгоритма заключается во введении весовых коэффициентов признаков, которые позволят варьировать степень важности признака при формировании выхода классификатора. Степень принадлежности экземпляра объекта *k*-му классу в таком случае будем определять по формуле:

$$\beta_k(\mathbf{x}_p) = \sum_{r_{jk}} \prod_{i=1}^{n} \left(w_i \times \mu_{T_{ij}}(x_p^i) \right), \tag{18}$$

где w_i – вес *i*-го признака, $w_i \in \mathbf{w} = (w_1, w_2, \ldots, w_n)$.

Популяция векторов весовых коэффициентов может быть сгенерирована различными способами. Самым простым и быстрым способом является случайная генерация. Другим подходом является создание популяции наложением случайного отклонения на первичный вектор весов, полученный путем оценки взаимной информации между признаками и выходом с нормированием полученных значений в промежутке от нуля до единицы. Такой способ позволяет получить отправную точку для алгоритма оптимизации, связанную с особенностями данных.

Полученная популяция подается на вход алгоритму оптимизации, задачей которого является поиск такого вектора весов признаков, который позволит получить лучшее значение фитнес-функции. В роли алгоритма оптимизации используется гибридный алгоритм из метаэвристик «гравитационный поиск» и «прыгающие лягушки», описанный выше. Алгоритм настройки весовых коэффициентов признаков в виде псевдокода приведен далее.

Вход: GIt, LIt, NM, NA, G_0, α, ε, $const_0$, ω.

Выход: $\mathbf{w}_0$.

Генерация популяции $\mathbf{W} = \{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{NA \times NM-1}\}$.

Вычисление фитнес-функции, сортировка популяции по убыванию фитнес-функции.

$t := 0$; $const := const_0$.

цикл пока ($t < GIt$):

Обновление $G(t)$ по формуле (14).

Расчет вектора масс $\mathbf{M}(t)$ по формулам (11), (12).

Вычисление ускорения $\mathbf{a}(t)$ на основе выражения (13).

Расчет скорости $\mathbf{V}(t)$ по формуле (15).

Обновление векторов популяции $\mathbf{W}$ в соответствии с выражением (16).

Нормализация векторов популяции.

Расчет фитнес-функции, сортировка популяции по убыванию фитнес-функции.

цикл пока ($LIt > 0$):

Локальный поиск на основе оператора (17).

$LIt := LIt - 1$.

конец цикла.

$t := t + 1$.

Сортировка популяции по убыванию фитнес-функции.

Обновление коэффициента *const*.

конец цикла.

вывод $\mathbf{w}_0$.

Третья глава посвящена экспериментальной проверке эффективности разработанных алгоритмов и статистическому сравнению полученных результатов с аналогами. Для экспериментов были использованы 36 наборов несбалансированных данных из репозитория KEEL (keel.es) с коэффициентом дисбаланса от 1,8 до 41,4. Все наборы данных содержат только два класса. Во всех экспериментах использовалась схема пятикратной кросс-валидации.

В качестве инструмента создания первичной базы правил нечеткого классификатора выбран алгоритм экстремальных значений признаков классов (АЭПК), создающий базы правил минимального объема, равного количеству классов. Для проверки эффективности сочетания АЭПК и разработанного алгоритма формирования структуры нечеткого классификатора на основе метаэвристики «прыгающие лягушки» (АЭПК+ПЛ), было проведено сравнение результатов классификации с общеизвестными алгоритмами создания структуры: Chi, Ishibuchi и E-алгоритмом, алгоритмом формирования иерархических баз правил HFRBCS, а также АЭПК без добавления правил. Алгоритмы Chi-3, Chi-5, Ishibuchi и HFRBCS осуществляли построение классификаторов на дополненных SMOTE данных. Как и E-алгоритм, эти алгоритмы использовали по тридцать правил для каждого класса. В таблице 1 представлены значения средней геометрической точности построенных нечетких классификаторов. Количество правил для комбинации АЭПК+ПЛ приведено в скобках.

Таблица 1 – Средняя геометрическая точность нечетких классификаторов

Данные	Chi-3	Chi-5	Ishibuchi	E-алг.	HFRBCS	АЭПК	АЭПК+ПЛ
gl1	64,9 ± 6,9	64,9 ± 6,9	59,3 ± 10,3	0,0 ± 0,0	**73,7** ± 4,7	40,5	68,9 ± 5,1 (7)
ecl0/1	92,3 ± 5,9	95,6 ± 5,2	**96,7** ± 2,4	95,3 ± 4,8	93,6 ± 6,5	88,8	96,4 ± 1,2 (4)
wis	88,9 ± 2,1	43,6 ± 5,9	95,8 ± 1,4	**96,0** ± 1,6	88,2 ± 1,6	73,4	95,1 ± 1,1 (4)
pm	66,8 ± 5,9	66,8 ± 2,3	71,1 ± 4,5	55,0 ± 4,6	68,7 ± 5,3	55,6	**72,2** ± 2,3 (9)
gl0	64,1 ± 3,5	63,7 ± 1,8	69,4 ± 7,7	0,0 ± 0,0	76,6 ± 8,1	60,1	**78,4** ± 3,8 (9)
yst1	67,7 ± 1,9	69,7 ± 1,5	51,4 ± 12,2	0,0 ± 0,0	**71,7** ± 2,4	39,6	69,9 ± 1,2 (9)
hbr	58,9 ± 6,0	60,4 ± 2,4	**62,7** ± 2,8	4,9 ± 11,1	57,1 ± 4,1	44,3	56,4 ± 4,8 (9)
vhc2	85,5 ± 3,4	87,2 ± 3,0	67,8 ± 5,0	43,8 ± 13,2	**90,6** ± 2,2	40,0	82,8 ± 4,0 (9)
vhc1	70,9 ± 4,3	**71,9** ± 1,3	64,9 ± 4,4	3,1 ± 6,9	71,8 ± 2,6	41,9	68,4 ± 1,9 (9)
vhc3	**69,2** ± 4,9	63,1 ± 2,0	63,1 ± 4,1	0,0 ± 0,0	66,8 ± 3,3	39,1	66,4 ± 2,6 (9)
gl0123/456	85,8 ± 3,0	85,9 ± 1,7	88,6 ± 5,2	82,1 ± 7,0	88,4 ± 4,0	87,6	**93,2** ± 1,5 (3)
vhc0	86,4 ± 3,1	84,9 ± 1,6	75,9 ± 1,4	39,1 ± 16,5	**88,9** ± 2,0	55,5	86,4 ± 2,2 (9)
ecl1	85,3 ± 9,8	86,1 ± 8,6	85,7 ± 2,9	77,8 ± 7,9	84,2 ± 12,7	80,8	**88,4** ± 1,6 (4)
nwth2	89,8 ± 10,8	96,3 ± 6,7	94,2 ± 4,2	88,6 ± 3,8	**99,7** ± 0,6	99,2	98,2 ± 0,1 (4)
nwth1	87,4 ± 8,1	95,4 ± 8,8	89,0 ± 13,5	88,5 ± 8,8	98,6 ± 2,5	**99,2**	97,2 ± 1,6 (3)
ecl2	88,0 ± 5,5	87,6 ± 5,0	87,0 ± 4,4	70,4 ± 15,4	87,6 ± 8,2	34,2	**90,5** ± 2,8 (9)
sgm0	95,0 ± 0,5	95,9 ± 1,2	42,5 ± 2,8	95,3 ± 1,1	97,5 ± 1,1	88,6	**97,6** ± 0,8 (9)
gl6	83,9 ± 9,8	78,1 ± 7,8	86,3 ± 8,2	**90,2** ± 3,8	87,0 ± 10,8	22,8	88,2 ± 4,7 (9)
yst3	90,1 ± 4,1	89,3 ± 3,3	77,1 ± 17,7	82,0 ± 2,3	**90,4** ± 2,3	85,5	**90,4** ± 1,8 (7)
ecl3	87,6 ± 4,1	**91,6** ± 5,0	85,4 ± 3,7	75,5 ± 8,7	90,8 ± 4,4	50,8	87,1 ± 3,0 (4)
pb0	79,9 ± 4,3	87,3 ± 1,9	32,2 ± 9,6	64,5 ± 2,8	**91,4** ± 0,7	63,6	88,2 ± 0,9 (9)
yst2/4	86,8 ± 5,5	86,4 ± 7,4	70,9 ± 23,5	80,9 ± 9,1	**89,3** ± 4,2	67,3	86,7 ± 4,9 (7)
yst05679/4	78,9 ± 6,0	76,0 ± 6,4	**79,5** ± 9,5	60,0 ± 16,4	73,2 ± 7,5	61,9	74,3 ± 3,6 (7)
vwl0	98,4 ± 0,6	97,9 ± 1,8	89,0 ± 6,6	89,6 ± 6,1	**98,8** ± 1,6	83,9	92,1 ± 3,4 (9)
gl2	47,7 ± 10,2	49,2 ± 8,2	43,6 ± 15,7	9,9 ± 22,1	54,8 ± 20,6	10,8	**61,7** ± 10,6 (4)
gl4	**85,0** ± 13,8	81,8 ± 11,2	78,3 ± 17,7	83,4 ± 19,9	70,4 ± 40,5	23,1	79,4 ± 10,4 (7)
ecl4	91,3 ± 7,4	92,1 ± 8,4	86,9 ± 8,7	92,4 ± 8,2	**93,0** ± 8,2	68,7	89,9 ± 3,1 (7)
pb13/2	91,9 ± 4,8	92,9 ± 9,5	94,5 ± 4,9	94,1 ± 10,3	**98,6** ± 0,7	75,1	90,1 ± 5,1 (7)
ab9/18	63,9 ± 11,0	66,5 ± 10,7	65,8 ± 9,2	32,3 ± 20,6	67,6 ± 14,0	58,7	**75,0** ± 2,3 (7)
yst1458/7	62,4 ± 4,6	58,8 ± 8,6	40,8 ± 16,6	0,0 ± 0,0	**62,5** ± 6,3	45,8	59,6 ± 3,2 (3)
yst2/8	72,8 ± 15,0	**78,8** ± 8,6	72,8 ± 15,0	72,8 ± 15,0	72,5 ± 15,1	68,8	70,2 ± 4,5 (7)
yst4	83,0 ± 3,1	**83,1** ± 2,6	71,4 ± 23,3	32,2 ± 20,6	82,6 ± 2,3	65,7	79,9 ± 2,1 (7)
yst1289/7	**76,1** ± 7,2	69,3 ± 4,6	48,6 ± 16,9	50,0 ± 13,6	69,4 ± 4,4	54,1	63,3 ± 7,2 (7)
yst5	93,4 ± 5,4	93,6 ± 2,1	**94,9** ± 0,4	88,2 ± 7,0	94,2 ± 2,6	72,9	93,5 ± 2,1 (7)
ecl0137/26	71,0 ± 41,4	49,6 ± 46,4	71,3 ± 41,7	**73,7** ± 43,1	71,5 ± 41,8	0,0	71,3 ± 3,8 (3)
yst6	87,5 ± 10,6	87,7 ± 9,3	**88,4** ± 6,1	51,7 ± 13,8	84,9 ± 12,9	51,2	84,0 ± 2,5 (4)
Среднее	80,0 ± 7,1	78,6 ± 6,4	73,4 ± 9,6	57,3 ± 9,6	**81,8 ±** 7,6	58,3	81,4 ± 3,3 (6,7)

В среднем «прыгающие лягушки» увеличили среднюю геометрическую точность на 23 процента по сравнению с исходным классификатором. Статистическое сравнение с применением критерия Уилкоксона показало: результаты АЭПК+ПЛ и Chi-3, Chi-5 и HFRBCS статистически не различимы (асимптотическая значимость (АЗ) 0,242, 0,315 и 0,396 соответственно), результаты в сравнении с Ishibuchi05 и E-алгоритмом обладают статистически значимым различием (АЗ меньше 0,001). Так как применение АЭПК+ПЛ не требует предобработки данных и позволяет получить сопоставимую или превосходящую аналогичные алгоритмы среднюю

геометрическую точность на меньшем количестве правил, его использование с целью получения точной и компактной системы классификации является предпочтительным.

Следующий эксперимент посвящен изучению качества работы гибридного алгоритма оптимизации параметров функций принадлежности. В таблице 2 представлены усредненные по десяти запускам результаты настройки параметров нечетких классификаторов с помощью одного из трех алгоритмов: метаэвристики «прыгающие лягушки» (ПЛ), метаэвристики «гравитационный поиск» (ГП) и их комбинации (ГП+ПЛ). Помимо средней геометрической точности (*GM*), использованной в качестве фитнес-функции всеми метаэвристиками, в таблице также представлено время работы алгоритмов в секундах (*Time*).

Таблица 2 – Результаты настройки параметров нечеткого классификатора

Данные	ГП+ПЛ		ПЛ		ГП	
	GM	*Time*	*GM*	*Time*	*GM*	*Time*
glass1	64,2 ± 5,3	14,4 ± 0,6	62,0 ± 5,4	11,4 ± 0,5	**64,7 ±** 5,2	77,4 ± 7,5
ecoli0vs1	96,0 ± 2,4	14,5 ± 0,2	96,1 ± 2,0	10,3 ± 0,9	**96,2 ±** 2,4	53,7 ± 7,2
wisconsin	**96,1 ±** 1,4	38,0 ± 1,1	94,8 ± 1,6	30,8 ± 1,1	95,9 ± 1,2	65,1 ± 47,1
pima	**72,6 ±** 2,1	34,8 ± 1,0	72,1 ± 2,4	35,4 ± 1,3	71,9 ± 1,9	208,5 ± 6,5
glass0	**78,4 ±** 3,1	16,9 ± 0,5	78,2 ± 4,4	15,3 ± 1,5	73,4 ± 5,7	34,0 ± 20,1
yeast1	71,3 ± 2,6	90,0 ± 2,8	**71,6 ±** 2,7	60,1 ± 3,5	69,4 ± 2,2	2,9 ± 0,0
haberman	59,2 ± 4,9	8,6 ± 0,1	52,9 ± 6,0	8,5 ± 0,4	**62,5 ±** 3,9	51,4 ± 2,3
vehicle2	**87,0 ±** 3,6	73,8 ± 2,8	84,9 ± 3,1	66,0 ± 2,7	77,8 ± 4,9	427,4 ± 12,8
vehicle1	**70,9 ±** 2,6	73,1 ± 2,9	69,6 ± 2,5	65,4 ± 3,1	66,9 ± 3,0	427,1 ± 11,4
vehicle3	68,8 ± 3,3	73,5 ± 3,0	**69,1 ±** 2,4	66,7 ± 3,7	65,3 ± 2,9	430,9 ± 13,1
gl.0123vs456	90,9 ± 3,5	14,5 ± 0,6	**91,2 ±** 3,2	10,1 ± 0,4	90,8 ± 3,1	77,1 ± 3,0
vehicle0	**93,2 ±** 1,3	72,7 ± 2,7	92,5 ± 1,8	74,6 ± 3,5	82,9 ± 3,6	498,9 ± 40,4
ecoli1	**88,2 ±** 3,2	20,4 ± 0,3	88,0 ± 3,2	13,4 ± 0,6	86,4 ± 4,3	10,1 ± 14,9
newthyroid2	**98,5 ±** 1,7	9,2 ± 0,2	97,8 ± 1,9	7,7 ± 0,4	97,8 ± 2,5	60,2 ± 1,0
newthyroid1	96,8 ± 3,2	9,2 ± 0,1	**98,2 ±** 1,7	8,0 ± 0,5	97,7 ± 2,6	59,9 ± 1,4
ecoli2	90,1 ± 2,9	20,8 ± 0,4	**91,0 ±** 2,8	13,4 ± 0,6	71,4 ± 28,6	0,8 ± 0,0
segment0	**98,7 ±** 0,7	205,6 ± 5,2	98,4 ± 0,5	204,7 ± 25,2	94,5 ± 0,9	823,7 ± 136,5
glass6	**86,3 ±** 4,6	16,5 ± 0,9	83,6 ± 5,6	13,4 ± 1,7	75,5 ± 17,1	33,2 ± 24,7
yeast3	90,6 ± 1,8	89,1 ± 3,2	**91,1 ±** 1,8	62,5 ± 6,6	90,4 ± 2,2	2,8 ± 0,0
ecoli3	**86,7 ±** 5,6	20,6 ± 0,3	83,7 ± 6,0	13,3 ± 0,5	66,5 ± 26,6	0,7 ± 0,0
page-blocks0	**85,8 ±** 1,8	276,9 ± 8,7	81,2 ± 2,2	250,8 ± 17,9	83,7 ± 3,0	1473,0 ± 17,2
yeast2vs4	86,3 ± 3,5	33,4 ± 0,9	**86,7 ±** 3,0	22,1 ± 1,0	84,8 ± 3,7	1,1 ± 0,0
yeast05679vs4	**76,3 ±** 7,3	34,0 ± 0,8	75,7 ± 8,1	23,6 ± 1,2	71,8 ± 7,0	1,2 ± 0,0
vowel0	94,5 ± 1,8	73,2 ± 2,0	**94,7 ±** 1,8	62,5 ± 3,5	91,5 ± 2,1	2,9 ± 0,1
glass2	70,3 ± 8,9	17,2 ± 0,4	**72,2 ±** 6,9	18,0 ± 1,6	48,0 ± 22,9	13,7 ± 13,5
glass4	82,6 ± 12,0	16,4 ± 0,6	**84,2 ±** 9,9	13,0 ± 1,1	72,6 ± 18,8	23,7 ± 20,5
ecoli4	**91,3 ±** 6,0	20,8 ± 0,3	90,1 ± 5,6	13,4 ± 0,4	90,9 ± 5,1	8,8 ± 12,9
page-bl.13vs2	94,2 ± 4,3	30,5 ± 0,7	**95,4 ±** 3,7	22,9 ± 1,2	92,4 ± 4,9	159,2 ± 4,3
abalone9-18	**84,7 ±** 3,2	42,3 ± 1,2	83,0 ± 5,2	28,3 ± 1,6	71,5 ± 6,0	1,3 ± 0,0
yeast1458vs7	**66,0 ±** 5,6	44,1 ± 1,7	60,9 ± 7,3	31,3 ± 2,7	52,9 ± 11,3	1,4 ± 0,0
yeast2vs8	**71,7 ±** 11,5	32,0 ± 0,6	69,4 ± 12,0	21,5 ± 0,9	67,9 ± 10,7	1,0 ± 0,0
yeast4	79,4 ± 3,5	89,3 ± 2,8	**79,8 ±** 3,6	63,2 ± 4,3	78,9 ± 3,6	2,8 ± 0,0
yeast1289vs7	**68,2 ±** 8,9	59,6 ± 1,6	67,0 ± 6,5	42,8 ± 4,4	62,3 ± 5,9	1,8 ± 0,0
yeast5	93,1 ± 4,8	90,5 ± 0,9	92,5 ± 4,8	59,5 ± 2,6	**94,3 ±** 3,1	2,8 ± 0,0
ecoli0137vs26	**60,5 ±** 37,3	14,2 ± 0,3	59,3 ± 38,8	12,1 ± 0,8	42,4 ± 44,6	64,7 ± 25,7
yeast6	85,5 ± 5,8	90,3 ± 1,6	**85,7 ±** 6,0	63,3 ± 4,8	84,8 ± 6,5	2,8 ± 0,0
Среднее	**82,6 ±** 5,2	52,2 ± 1,5	81,8 ± 5,2	42,8 ± 3,0	77,5 ± 7,9	141,9 ± 12,3

Сравнение критерием Уилкоксона показывает, что существует статистическое различие между результатами гибрида и исходными метаэвристиками: асимптотическая значимость при сравнении с «прыгающими лягушками» равняется 0,017, а при сравнении с «гравитационным поиском» – меньше 0,001. По времени работы лучший результат показали «прыгающие лягушки»: при сравнении с гибридом асимптотическая значимость меньше 0,001, что свидетельствует о статистическом различии в исследуемом показателе. Время работы «гравитационного поиска» и гибрида оказывается статистически неразличимым (АЗ равняется 0,239). При необходимости получить лучшую точность между тремя метаэвристиками стоит выбрать гибрид, но для более быстрого обучения нужно использовать «прыгающие лягушки».

Проведено статистическое сравнение полученной точности с результатами аналогов из таблицы 1. Нулевая гипотеза (НГ) гласит, что между результатами нет статистических различий. Итоги сравнения приведены в таблице 3.

Таблица 3 – Статистическое сравнение непараметрическим критерием Уилкоксона

Алгоритмы	Chi-3		Chi-5		Ishibuchi		E-алгоритм		HFRBCS	
	АЗ	НГ	АЗ	НГ	АЗ	НГ	АЗ	НГ	АЗ	НГ
ГП+ПЛ	0,029	Откл.	0,018	Откл.	<0,001	Откл.	<0,001	Откл.	0,671	Прин.

Нечеткие классификаторы несбалансированных данных, параметры которых настроены комбинацией ГП+ПЛ, всего лишь на двух нечетких правилах показывают статистически различимые результаты с алгоритмами Chi-3, Chi-5, Ishibuchi и E-алгоритмом, а также сопоставимые результаты с иерархическими нечеткими системами HFRBCS. По сравнению с результатами до оптимизации (таблица 1, столбец АЭПК), средняя геометрическая точность после оптимизации параметров алгоритмом ГП+ПЛ возросла в среднем на 24 процента.

Эффективность алгоритма настройки весовых коэффициентов признаков проверена путем сравнения нечетких классификаторов, веса признаков которых были настроены гибридом ГП+ПЛ, с нечеткими классификаторами, представленными в таблице 1. В качестве фитнес-функции ГП+ПЛ использовалась средняя геометрическая точность. В таблице 4 представлены значения этой метрики, полученные после усреднения результатов пятнадцати запусков. При оптимизации весов были использованы два способа создания популяции: на основе взаимной информации (ГП+ПЛ+ВИ) и случайная генерация (ГП+ПЛ+СГ).

Таблица 4 – Результаты классификации после настройки весов

Данные	ГП+ПЛ+ВИ	ГП+ПЛ+СГ	Данные	ГП+ПЛ+ВИ	ГП+ПЛ+СГ
glass1	61,0 ± 0,8	60,4 ± 0,8	yeast3	89,8 ± 0,4	89,5 ± 0,4
ecoli0vs1	96,8 ± 0,4	96,7 ± 0,2	ecoli3	86,3 ± 0,4	86,2 ± 0,4
wisconsin	92,0 ± 0,1	92,0 ± 0,0	page-blocks0	76,6 ± 2,7	75,4 ± 3,4
pima	64,8 ± 0,5	64,5 ± 0,4	yeast2vs4	86,7 ± 0,7	86,8 ± 0,7
glass0	77,6 ± 1,0	77,5 ± 0,7	yeast05679vs4	73,4 ± 1,6	72,5 ± 1,1
yeast1	59,8 ± 0,5	60,4 ± 0,5	vowel0	89,7 ± 0,2	89,7 ± 0,3
haberman	43,6 ± 0,0	43,6 ± 0,0	glass2	35,7 ± 3,8	38,3 ± 4,3

Продолжение таблицы №4

Данные	ГП+ПЛ+ВИ	ГП+ПЛ+СГ	Данные	ГП+ПЛ+ВИ	ГП+ПЛ+СГ
vehicle2	70,0 ± 1,3	70,3 ± 1,5	glass4	23,9 ± 6,4	28,3 ± 8,4
vehicle1	61,9 ± 1,2	61,3 ± 1,3	ecoli4	90,6 ± 0,3	91,3 ± 1,1
vehicle3	57,9 ± 1,3	58,6 ± 1,1	page-blocks13vs2	89,4 ± 2,1	89,4 ± 2,1
glass0123vs456	89,7 ± 1,5	89,6 ± 1,1	abalone9-18	75,7 ± 1,5	76,1 ± 1,3
vehicle0	69,9 ± 0,9	70,0 ± 0,7	yeast1458vs7	50,9 ± 1,6	50,6 ± 1,3
ecoli1	89,2 ± 0,2	89,3 ± 0,1	yeast2vs8	74,2 ± 1,6	74,8 ± 1,1
newthyroid2	98,6 ± 0,4	98,9 ± 0,2	yeast4	68,9 ± 0,8	68,5 ± 0,4
newthyroid1	98,8 ± 0,1	98,8 ± 0,1	yeast1289vs7	61,2 ± 0,3	61,2 ± 0,3
ecoli2	86,2 ± 0,4	86,4 ± 0,2	yeast5	93,9 ± 0,6	93,6 ± 0,9
segment0	95,8 ± 0,4	95,7 ± 0,6	ecoli0137vs26	52,3 ± 8,6	58,9 ± 9,2
glass6	73,2 ± 4,3	78,9 ± 5,6	yeast6	77,3 ± 0,7	77,8 ± 0,9

Результаты сравнения с аналогами из таблицы 1 представлены в таблице 5.

Таблица 5 – Статистическое сравнение непараметрическим критерием Уилкоксона

Алгоритмы	Chi-3		Chi-5		Ishibuchi		E-алгоритм		HFRBCS	
	АЗ	НГ	АЗ	НГ	АЗ	НГ	АЗ	НГ	АЗ	НГ
ГП+ПЛ+ВИ	0,017	Откл.	0,017	Откл.	0,649	Прин.	<0,001	Откл.	0,001	Откл.
ГП+ПЛ+СГ	0,018	Откл.	0,034	Откл.	0,540	Прин.	<0,001	Откл.	0,001	Откл.

Результаты двух подходов к генерации популяции при настройке весовых коэффициентов признаков статистически не различимы (АЗ равняется 0,381). Точность обеих версий сопоставима с алгоритмом Ishibuchi, статистически различима с E-алгоритмом в пользу предлагаемого алгоритма настройки весов, и различима с результатами Chi-3, Chi-5 и HFRBCS в пользу последних. Поскольку главная ценность средств отбора и взвешивания признаков – это уменьшение сложности модели, увеличение точности является не таким существенным, как при использовании двух ранее предложенных в данной работе алгоритмов. Поэтому этап взвешивания признаков стоит комбинировать как минимум с этапом настройки термов. Так, алгоритм настройки весов признаков на основе гибрида из «прыгающих лягушек» и «гравитационного поиска» увеличил среднюю точность на 16 процентов по сравнению с исходным классификатором; при проведении оптимизации параметров термов на 500 итераций алгоритмом перед настройкой весов, увеличение точности достигло 20 процентов.

В четвертой главе представлено описание применения разработанных алгоритмов для построения системы оценки системы свертывания крови у беременных женщин. Совместно с сотрудниками ОГАУЗ «Родильный дом №1» было создано программное обеспечение, позволяющее на основе результатов анализа реологических свойств крови АРП-01 «МЕДНОРД» оценивать текущее состояние системы свертывания крови пациенток по трем показателям: общее состояние, хронометрические характеристики, структурные характеристики. Каждый показатель может принимать три значения: гипокоагуляция, норма, гиперкоагуляция. Для каждого из трех показателей построены нечеткие классификаторы с трапециевидными функциями принадлежности. При создании классификаторов были использованы все три описанных в работе

алгоритма. Средняя геометрическая точность полученной системы (99% для общего состояния, 97% для хронометрических и структурных характеристик) позволила применять разработанную программу для оценки системы свертывания крови для беременных женщин в клинико-диагностической лаборатории ОГАУЗ «Родильный дом №1».

В заключении сделаны выводы о полученных в процессе работы результатах.

Заключение

В процессе выполнения диссертационной работы разработаны алгоритмы построения и нечетких классификаторов, способствующие повышению средней геометрической точности на несбалансированных данных при построении изначальной структуры алгоритмом экстремальных значений признаков классов. Результаты выполнения задач описаны далее.

1. Проведен обзор методов обработки несбалансированных данных при построении классификаторов. Выдвинуто предположение, что изменение способа оценки качества нечетких классификаторов и использование сбалансированных метаэвристик позволит работать с несбалансированными данными без применения этапа предобработки.

2. Разработан алгоритм формирования структуры нечеткого классификатора несбалансированных данных на основе метаэвристики «прыгающие лягушки», осуществляющей итеративное создание и настройку правил для классов с наименьшей долей правильной классификации. Предложена фитнес-функция, объединяющая среднюю геометрическую и общую точность. Разработанный алгоритм в комбинации с алгоритмом экстремальных значений признаков классов позволил улучшить среднюю геометрическую точность на 23 процента относительно результатов до добавления правил. Алгоритм продемонстрировал при меньшем числе правил сопоставимую или большую среднюю геометрическую точность по сравнению с общеизвестными алгоритмами построения нечетких классификаторов Chi+SMOTE, Ishibuchi+SMOTE, HFRBCS+SMOTE и E-алгоритмом.

3. Разработан гибридный алгоритм настройки параметров нечеткого классификатора, основанный на комбинации метаэвристики «гравитационный поиск» с локальным поиском из метаэвристики «прыгающие лягушки». По сравнению с результатами до оптимизации алгоритм позволил увеличить среднюю геометрическую точность классификации в среднем на 24 процента. При существенно меньшем количестве используемых правил алгоритм получил сопоставимую точность с HFRBCS+SMOTE и большую точность по сравнению с Chi+SMOTE, Ishibuchi+SMOTE и E-алгоритмом.

4. Разработан алгоритм настройки весовых коэффициентов признаков, отражающих важность признака при формировании вывода нечеткого классификатора. Настройка вектора весов осуществляется гибридным алгоритмом из метаэвристик «гравитационный поиск» и

«прыгающие лягушки». Алгоритм увеличил среднюю геометрическую точность в среднем на 16 процентов относительно точности до введения весов. В сравнении с аналогами алгоритм показывает лучшие результаты, чем E-алгоритм, сопоставимую точность с Ishibuchi+SMOTE, и уступает комбинациям Chi-3, Chi-5 и HFRBCS с алгоритмом SMOTE. При комбинации с оптимизацией параметров термов увеличение средней геометрической точности составило 20 процентов. Предложенная комбинация показала большую точность по сравнению с Ishibuchi и E-алгоритмом, сопоставимую точность с Chi-3 и Chi-5, и уступила HFRBCS.

5. Разработанные алгоритмы применены при создании нечетких классификаторов для оценки текущего состояния свертывающей системы крови у пациенток ОГАУЗ «Родильный дом №1». При оценке общего состояния системы свертывания крови на тестовых данных средняя геометрическая точность составила 99 процентов, при анализе хронометрических и структурных характеристик – 97 процентов. Созданное совместно с сотрудниками родильного дома программное обеспечение используется в клинико-диагностической лаборатории.

Публикации автора по теме диссертации

Работы, опубликованные в журналах, рекомендованных ВАК:

1. Ходашинский, И.А. Комплексная оценка параметров коагуляции у беременных женщин с помощью нечеткого классификатора / И.А. Ходашинский, И.Б. Бардамова, М.Б. Бардамова // Медицинская техника. – 2017. – N 3(303). – С. 52–55 (Scopus).

2. Ходашинский, И. А. Построение нечеткого классификатора алгоритмом гравитационного поиска / И. А. Ходашинский, М. Б. Бардамова, В. С. Ковалев // Доклады ТУСУР. – 2017. – Т. 20, N 2. – С. 84–87.

3. Ходашинский, И. А. Отбор признаков и построение нечеткого классификатора на основе алгоритма прыгающих лягушек / И. А. Ходашинский, М. Б. Бардамова, В. С. Ковалев // Искусственный интеллект и принятие решений. – 2018. – N 1. – С. 76–84 (Scopus).

4. Аутентификация пользователя по динамике подписи на основе нечёткого классификатора / И. А. Ходашинский, Е. Ю. Костюченко, М. Б. Бардамова [и др.] // Компьютерная оптика. – 2018. – N 42(4). – С. 657–666 (Scopus, WoS).

5. Бардамова, М. Б. Способы адаптации алгоритма прыгающих лягушек к бинарному пространству поиска при решении задачи отбора признаков / М. Б. Бардамова, А. Г. Буймов, В. Ф. Тарасенко // Доклады ТУСУР. – 2020. – Т. 23, N 4. – С. 57–62.

6. Бардамова, М.Б. Формирование структуры нечеткого классификатора комбинацией алгоритма экстремумов классов и алгоритма «прыгающих лягушек» для несбалансированных данных с двумя классами / М.Б. Бардамова, И. А. Ходашинский // Автометрия. – 2021. – Т. 57, N 4. – С. 54-64.

Помимо статей под номерами 1, 3 и 4, в Scopus и WoS проиндексированы статьи:

7. Hodashinsky, I. A. Tuning fuzzy systems parameters with chaotic particle swarm optimization / I. A. Hodashinsky, M. B. Bardamova // Journal of Physics: Conference Series. – 2017. – Vol. 803, N 1. – P. 012053 (Scopus, WoS).

8. A Fuzzy Classifier with Feature Selection Based on the Gravitational Search Algorithm / M. Bardamova, A. Konev, I. Hodashinsky, A. Shelupanov // Symmetry. – 2018. – Vol. 10, N 11. – P. 609 (Scopus, WoS).

9. Fuzzy classifier design for network intrusion detection using the gravitational search algorithm / M. Bardamova, A. Konev, I. Hodashinsky, A. Shelupanov // Journal of Physics: Conference Series. – 2019. – Vol. 1145, N 1. – P. 012008 (Scopus).

10. Gravitational search for designing fuzzy rule-based classifiers for handwritten signature verification / M. Bardamova, A. Konev, I. Hodashinsky, A. Shelupanov // Journal of Communications Software and Systems. – 2019. – Vol. 15, N 3. – P. 254–261 (Scopus).

11. Application of the Gravitational Search Algorithm for Constructing Fuzzy Classifiers of Imbalanced Data / M. Bardamova, A. Konev, I. Hodashinsky, A. Shelupanov // Symmetry. – 2019. – Vol. 11, N 12. – P. 1458 (Scopus, WoS).

12. Bardamova, M. B. Optimization of fuzzy classifier parameters with a combination of gravitational search algorithm and shuffled frog leaping algorithm / M. B. Bardamova, I.A. Hodashinsky // Journal of Physics: Conference Series. – 2020. – Vol. 1611, N 1. – P. 012068 (Scopus).

13. Bardamova, M. Hybrid Algorithm for Tuning Feature Weights in a Fuzzy Classifier / M. Bardamova, I. Hodashinsky // 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). – IEEE, 2021. – P. 0354-0357 (Scopus).

Другие работы, опубликованные по теме диссертации.

14. Прогнозирование результатов исследования реологических свойств крови у беременных женщин для оценки свертывающей системы с использованием нечеткого классификатора / И. А. Ходашинский, И. Б. Бардамова, М. Б. Бардамова [и др.] // Материалы Всероссийской научно-практической конференции в рамках I Конгресса «Здравоохранение России. Технологии опережающего развития». – Томск: STT, 2015. – С. 95–98.

15. Бардамова, М. Б. Применение нечеткого классификатора для прогнозирования риска возникновения и развития сердечно-сосудистых заболеваний / М. Б. Бардамова, В. С. Ковалёв, И. В. Горбунов // Материалы докладов XI Международной научно-технической конференции «Электронные средства и системы управления». – Томск: В-Спектр, 2015. – Ч. 1. – С. 248-252.

16. Бардамова, М. Б. Построение нечеткого классификатора для оценки состояния системы гемостаза у беременных женщин / М. Б. Бардамова // Сборник трудов XIV

Международной научно-практической конференции «Молодежь и современные информационные технологии». – Томск: ТПУ, 2016. – Т. 2. – С. 294–295.

17. Метаэвристические методы оптимизации параметров нечетких классификаторов / И. А. Ходашинский, А. Е. Анфилофьев, М. Б. Бардамова [и др.] // Информационные и математические технологии в науке и управлении. – 2016. – N 1. – С. 73-80.

18. Сравнительный анализ эффективности метаэвристических алгоритмов при построении нечетких классификаторов / М. Б. Бардамова, А. Е. Анфилофьев, В. С. Ковалев, И. В. Филимоненко // Сборник научных трудов IV Международной летней школы-семинара по искусственному интеллекту «Интеллектуальные системы и технологии: современное состояние и перспективы». – СПб.: Политехника-сервис, 2017. – С. 22–31.

19. Бардамова, М. Б. Бинаризация непрерывных метаэвристик в задачах отбора признаков для нечетких классификаторов / М. Б. Бардамова, И. А. Ходашинский // Труды VII всероссийской научной-практической конференции «Нечеткие системы, мягкие вычисления и интеллектуальные технологии». – СПб.: Политехника-сервис, 2017. – Т. 2. – С. 18–25.

20. Метаэвристические методы отбора информативных классифицирующих признаков / И. А. Ходашинский, А. Е. Анфилофьев, М. Б. Бардамова, К. С. Сарин // Информационные и математические технологии в науке и управлении. – 2017. – N 2 (6). – С. 11–18.

21. Bardamova, M. B. Designing fuzzy classifiers with feature selection by the binary gravitational search algorithm for imbalanced data / M. B. Bardamova // Материалы докладов XIV Международной научно-практической конференции «Электронные средства и системы управления». – Томск: В-Спектр, 2018. – Ч.2 – С. 266–269.

22. Ходашинский, И. А. Применение ранжирования и схем кроссвалидации при отборе признаков для нечеткого классификатора / И. А. Ходашинский, Ф. Е. Анфилофьев, М. Б. Бардамова [и др.] // Информационные и математические технологии в науке и управлении. – 2018. – N 2(10). – С. 31–41.

23. Ходашинский, И. А. Исследование эффективности бинарного гравитационного алгоритма при построении нечетких классификаторов с отбором признаков / И. А. Ходашинский, М. Б. Бардамова // Материалы IV Всероссийской Поспеловской конференции с международным участием «Гибридные и синергетические интеллектуальные системы». – Калининград: Изд-во БФУ им. Иммануила Канта, 2018. – С. 448–455.

24. Бардамова, М. Б. Нечеткий классификатор несбалансированных медицинских данных с применением алгоритма прыгающих лягушек / М. Б. Бардамова // Сборник избранных статей научной сессии ТУСУР. – Томск: В-Спектр, 2019. – Т. 1, N 1-2. – С. 41–44.

25. Бардамова, М. Б. Формирование структуры нечеткого классификатора алгоритмом на основе экстремумов классов, дополненного алгоритмом прыгающих лягушек / М. Б. Бардамова

// Сборник избранных статей по материалам международной научно-технической конференции «Научная сессия ТУСУР». – Томск: В-Спектр, 2020. – Ч. 2. – С. 49–51.

26. Бардамова, М. Б. Оптимизация параметров нечеткого классификатора комбинацией алгоритмов гравитационного поиска и прыгающих лягушек / М. Б. Бардамова // Сборник трудов XVII Международной конференции «Перспективы развития фундаментальных наук». – Томск: Изд-во Томск. гос. ун-та систем упр. и радиоэлектроники, 2020. – Т. 7. – С. 23–25.

27. Ходашинский, И. А. Модификации алгоритма прыгающих лягушек для отбора признаков в нечетком классификаторе при аутентификации пользователя по рукописной подписи / И. А. Ходашинский, М. Б. Бардамова // Информационные и математические технологии в науке и управлении. – 2020. – 4(20). – С 75–83.

28. Bardamova, M. B. Binarization of the Shuffled frog leaping algorithm for feature selection in fuzzy classifiers / M. B. Bardamova // Электронные средства и системы управления: материалы докладов XVI Международной научно-практической конференции. – Томск: В-Спектр, 2020. – Ч. 2. – С. 232–235.

Свидетельства о государственной регистрации программы для ЭВМ

1. Бардамова М.Б., Ходашинский И.А. Программа настройки параметров нечеткого классификатора на основе гравитационного поиска // Свидетельство о государственной регистрации программы для ЭВМ N 2018614316. Дата регистрации в реестре: 04.04.2018.

2. Бардамова М.Б., Ходашинский И.А. Программа отбора признаков для нечеткого классификатора на основе бинарного алгоритма гравитационного поиска со статическими функциями трансформации // Свидетельство о государственной регистрации программы для ЭВМ N 2018612574. Дата регистрации в реестре: 22.02.2019.

3. Бардамова М.Б., Ходашинский И.А. Программа добавления правил на основе алгоритма прыгающих лягушек для нечеткого классификатора несбалансированных данных. Свидетельство о государственной регистрации программы для ЭВМ N 2021611060. Дата регистрации в реестре: 21.01.2021.

4. Бардамова М.Б., Ходашинский И.А. Программа настройки параметров нечеткого классификатора несбалансированных данных комбинацией гравитационного алгоритма и алгоритма прыгающих лягушек. Свидетельство о государственной регистрации программы для ЭВМ N 2021611138. Дата регистрации в реестре: 22.01.2021.