На правах рукописи

СТАНОВОВ ВЛАДИМИР ВАДИМОВИЧ

САМОНАСТРАИВАЮЩИЕСЯ ЭВОЛЮЦИОННЫЕ АЛГОРИТМЫ ФОРМИРОВАНИЯ СИСТЕМ НА НЕЧЕТКОЙ ЛОГИКЕ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск 2016

Работа выполнена в ФГБОУ ВО «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева», г. Красноярск

Научный руководитель: доктор технических наук, профессор
Семенкин Евгений Станиславович

Официальные оппоненты: **Кравец Олег Яковлевич**
доктор технических наук, профессор,
ФГБОУ ВО «Воронежский государственный технический университет»
профессор кафедры автоматизированных и вычислительных систем

Демидова Лилия Анатольевна
доктор технических наук, профессор,
ФГБОУ ВО «Рязанский государственный радиотехнический университет»
профессор кафедры вычислительной и прикладной математики

Ведущая организация: Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)

Защита состоится 23 декабря 2016 г. в 14 часов на заседании диссертационного совета Д 212.249.05, созданного на базе ФГБОУ ВО «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева» по адресу 660037 г. Красноярск, проспект имени газеты «Красноярский рабочий», 31.

С диссертацией можно ознакомиться в библиотеке Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнева на сайте СибГАУ: http://sibsau.ru

Автореферат разослан «_» ноября 2016 г.

Ученый секретарь
диссертационного совета Илья Александрович Панфилов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. На сегодняшний день разработка методов интеллектуального анализа данных является стремительно развивающимся направлением. Цель интеллектуальных систем анализа и обработки информации заключается не только в минимизации затрат исследователей или же пользователя интеллектуальной системы при решении сложных задач, но и полностью автоматический поиск закономерностей в исследуемой предметной области.

Среди всех задач интеллектуального анализа данных стоит выделить задачи классификации, так как к ним сводится множество реальных задач, в том числе классификация изображений, распознавание фрагментов текста, устной речи, классификация поисковых запросов, а также задачи медицинской диагностики. На сегодняшний день разработано множество интеллектуальных систем анализа данных (ИСАД), которые в зарубежной литературе, как правило, называются алгоритмами Data mining. Среди современных отечественных научных школ, занимающихся данной проблематикой, следует выделить Ю.И. Журавлёва, К.В. Рудакова (ВЦ РАН), Н.Г. Загоруйко (ИМ СО РАН), А.А. Дорофеюка (ИПУ РАН).

Недостатком большинства подходов ИСАД является то, что зачастую они работают по принципу «черного ящика», что значительно затрудняет интерпретацию результатов классификации и построенных классификатором закономерностей. По этой причине ряд отечественных и зарубежных исследователей занимается проблемами формирования классификаторов, которые могут быть легко поняты и представлены в форме естественного языка. Наиболее популярным направлением здесь является формирование нечетких систем. Среди отечественных исследователей данной проблематикой занимаются, например, И.А. Ходашинский (ТУСУР), А.П. Рыжов (МГУ), а среди зарубежных следует выделить работы Х. Ишибучи (Hisao Ishibuchi, Osaka University, Japan) и Ф. Херреры (Francisco Herrera, Granada University, Spain).

Системы на нечеткой логике (НЛС) позволяют строить лингвистические правила и объединять их в базы правил, которые представляют собой модель «белого ящика». Значительный вклад в разработку нечетких систем классификации сделали группы испанских специалистов во главе с Ф. Херрерой и японских исследователей во главе с Х. Ишибучи. Нечеткая база правил представляет собой набор независимых правил, каждое из которых выражает причинно-следственную связь между входными переменными и соответствующим классом. Нечеткие правила оперируют лингвистическими понятиями, вследствие чего могут быть непосредственно восприняты экспертом. Эта особенность позволяет использовать нечеткие базы правил не только как инструмент классификации, но и как метод интеллектуального анализа данных для извлечения новых знаний.

Формирование нечеткой системы классификации заключается в определении структуры базы правил – то есть в поиске значимых правил и выборе наилучшей комбинации этих правил. Данная задача может быть сформулирована как задача оптимизации. При этом целевая функция характеризуется значительной вычислительной сложностью, так как задана алгоритмически, имеет большую размерность и пространство поиска, характеризуется наличием дискретных переменных и т.д.

Эволюционные методы оптимизации хорошо зарекомендовали себя для решения сложных оптимизационных задач, вследствие чего их применение к формированию баз нечетких правил для задачи классификации является целесообразным.

Применение эволюционных методов для построения нечетких классификаторов может повлечь значительные временные затраты. С ростом объемов данных, которые необходимо подвергать интеллектуальному анализу вследствие развития интернет-технологий и отсутствия экспертов в некоторых областях, разработка быстрых и эффективных средств интеллектуального анализа становится всё более востребованной. Процедуры селекции обучающих примеров, подразумевающие выбор обучающих примеров в процессе работы алгоритма позволяют значительно снизить объем требуемых вычислительных ресурсов, и, помимо того, повысить качество и робастность получаемых интеллектуальных систем.

Таким образом, разработка и исследование методов автоматизированного формирования баз нечетких правил методами эволюционных алгоритмов с активным выбором обучающих примеров для классификации с извлечением скрытых знаний является **актуальной научно-технической задачей.**

Целью диссертационной работы является повышение качества и интерпретируемости нечетких классификаторов, а также снижение требуемых вычислительных ресурсов при их формировании за счет применения самонастраивающихся эволюционных алгоритмов.

Достижение поставленной цели предполагает решение следующих задач:

1. Выполнить обзор существующих методик и алгоритмов формирования нечетких правил и баз правил с целью выявления наиболее эффективных подходов и направлений.
2. Исследовать методы самонастройки эволюционных алгоритмов оптимизации на репрезентативном множестве тестовых задач.
3. Разработать алгоритм формирования баз нечетких правил для решения задач классификации с несбалансированными данными.
4. Разработать метод селекции обучающих примеров для нечеткого классификатора, позволяющий снизить временные затраты и повысить эффективность алгоритма.
5. Реализовать разработанные подходы в виде программных систем и протестировать их эффективность на репрезентативном множестве тестовых и реальных задач.

Методы исследования. В процессе выполнения данной диссертационной работы использовались методы статистической обработки данных, теории

вероятностей, эволюционных вычислений, оптимизации, нечеткой логики, системного анализа данных, моделирования динамических систем, выявления закономерностей в исходных данных.

Научная новизна работы включает следующие пункты:

1. Разработан новый самонастраивающийся эволюционный алгоритм формирования нечетких систем для решения задач классификации с представлением баз правил в форме матриц переменной размерности, отличающийся от известных использованием оценки достоверности правил при назначении их весовых коэффициентов и за счет этого превосходящий по эффективности другие методы эволюционного построения нечетких систем.
2. Разработан новый метод гибридизации Питтсбургского и Мичиганского подходов в эволюционном алгоритме формирования баз нечетких правил, отличающийся от известных использованием при построении новых правил вероятностной процедуры выбора релевантных нечетких термов и позволяющий существенно повысить точность классификации на первых поколениях работы эволюционного алгоритма.
3. Разработан новый метод селекции примеров для обучения классификаторов, отличающийся от известных адаптивной вероятностной процедурой организации подвыборок и назначения весовых коэффициентов и позволяющий одновременно повысить точность классификации и снизить объем требуемых для этого вычислительных ресурсов.
4. Разработан новый метод самонастройки эволюционных алгоритмов, отличающийся от известных схемой оценки успешности операторов, применяемых несколько раз к каждому индивиду, и позволяющий настраивать вероятности применения эвристик в Мичиганской части алгоритма.

Теоретическая значимость результатов диссертационной работы состоит в разработке новых эволюционных алгоритмов формирования нечетких систем, позволяющих получать компактные и точные базы правил посредством использования кодирования в форме матриц переменной размерности, гибридизации Питтсбургского и Мичиганского подходов и применения алгоритма самонастройки, для решения задач классификации и разработке нового метода активной селекции обучающих примеров для классификаторов, что представляет собой существенный вклад в теорию и практику исследования методов формирования нечетких систем посредством эволюционных алгоритмов.

Практическая ценность. Разработанные методы реализованы в виде программной системы, для решения задач классификации. Программная система позволяет быстро формировать базы нечетких правил за счет использования самонастройки, а также снижения количества пересчетов степеней принадлежности и весов правил. Программная система

протестирована на задачах классификации из области техники, распознавания изображений, банковского скоринга и медицинской диагностики.

Реализация результатов работы. Разработанные алгоритмы использованы при выполнении исследований в рамках российско-германских проектов (совместно с университетом г. Ульм) «Распределенные интеллектуальные информационные системы обработки и анализа мультилингвистической информации в диалоговых информационно-коммуникационных системах» (ФЦП ИР, ГК №11.519.11.4002) и «Математическое и алгоритмическое обеспечение автоматизированного проектирования аппаратно-программных комплексов интеллектуальной обработки мультилингвистической информации в распределенных высокопроизводительных системах космического назначения» (ФЦП НПК, ГК № 16.740.11.0742), российско-словенского проекта (совместно с университетом г. Марибор) «Manpower control strategy determination with self-adapted evolutionary and biologically inspired algorithms» (ARRS Project BI-RU/14-15-047), а также в рамках проекта №8.5541.2011 «Развитие теоретических основ автоматизации математического моделирования физических систем на основе экспериментальных данных» и проекта № 140/14 «Разработка теоретических основ эволюционного проектирования интеллектуальных информационных технологий анализа данных» тематического плана ЕЗН СибГАУ. Диссертационная работа была поддержана Фондом содействия развитию малых форм предприятий в научно-технической сфере по программе «У.М.Н.И.К» («Участник молодежного научно-инновационного конкурса») в рамках НИОКР «Разработка программного обеспечения интеллектуального анализа данных "FuzzyMiner"» на 2014-2016 гг., а также Российским Фондом Фундаментальных Исследований в рамках проекта № 16-31-00349 «Разработка алгоритмов и подходов к повышению качества и скорости формирования технологий интеллектуального анализа данных посредством снижения размерности данных» на 2016-2017 гг.

Три программные системы, разработанные в ходе выполнения диссертации, зарегистрированы в Роспатенте. Данные программные системы используются в учебном процессе Института информатики и телекоммуникаций СибГАУ при выполнении лабораторных и курсовых работ и переданы в две инновационные IT-компании.

Основные защищаемые положения:

1. Разработанный метод формирования нечетких систем для решения задачи классификации самонастраивающимся эволюционным алгоритмом позволяет формировать компактные и легко интерпретируемые базы правил.
2. Предложенная схема кодирования базы правил в эволюционном алгоритме позволяет снизить вычислительную сложность алгоритма.
3. Гибридный алгоритм формирования нечетких баз правил для решения задач классификации не уступает по точности другим подходам.

4. Разработанный метод селекции обучающих примеров позволяет существенно снизить объем требуемых вычислительных ресурсов.
5. Применение метода селекции обучающих примеров к гибридному эволюционному алгоритму формирования нечетких баз правил позволяет формировать более эффективные классификаторы в смысле точности, полноты и F-меры.

Публикации. По теме данной работы опубликовано более 35 печатных работ, в том числе 8 в журналах из Перечня ВАК, а также зарегистрировано в Роспатенте три программные системы.

Апробация работы. Результаты диссертационной работы были доложены на 12 всероссийских и международных научно-практических конференциях и конференциях с международным участием, в том числе на Пятой международной конференции «Системный анализ и информационные технологии» САИТ-2013 (Красноярск, 2013), Второй и Третьей международных конференциях по математическим моделям и их применениям (2^{nd} and 3^{rd} International Workshops on Mathematical Models and their Applications, Красноярск, 2013, 2014), III Всероссийской научной конференции с международным участием «Теория и практика системного анализа» (ТПСА, Рыбинск, 2014), 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD, Xaimen, China, 2014), 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO, Vienna, Austria, 2014), International Congress on Evolutionary Computations (CEC, Sendai, Japan, 2015), International Conference on Swarm Intelligence (ICSI, Peking, China, 2015), IEEE Symposium Series on Computational Intelligence (SSCI 2015, South Africa), 4th International Congress on Advanced Applied Informatics (AAI 2015), July 12-16, Okayama Convention Center, Okayama, Japan, 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016).

Структура работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулирована цель и поставлены задачи исследования, рассмотрены вопросы научной новизны и практической ценности проведенных исследований, изложены основные положения, выносимые на защиту.

В **первой главе** рассматривается проблема выбора параметров эволюционных алгоритмов на примере стандартного генетического алгоритма оптимизации. Проблема выбора параметров возникает из-за различной эффективности генетических операторов селекции, скрещивания и мутации при решении различных задачи классификации. Каждый генетический оператор обладает определенными свойствами и меняет характер поиска, причем для каждой новой задачи заранее определить наиболее подходящий набор операторов невозможно. Более того, в процессе решения

оптимизационной задачи на различных этапах высокую эффективность могут показывать разные конфигурации алгоритма.

В настоящее время для решения данной проблемы помимо многократного запуска алгоритма с различными конфигурациями, используются и другие методы, в том числе, коэволюционный подход и методы самоконфигурации. Суть коэволюционного подхода состоит в использовании нескольких конкурирующих либо кооперирующих популяций с различными настройками. Методы самоконфигурирования меняют вероятности применения генетических операторов в пределах одной популяции. При этом наиболее успешные операторы получают большие вероятности быть примененными на следующих поколениях. Таким образом, самоконфигурирование позволяет менять поведение алгоритма в автоматическом режиме в ходе решения задачи. Существующие методы самоконфигурирования отличаются схемами определения успешности операторов, а также способами назначения новых вероятностей. Общая схема самоконфигурирования включает следующие этапы, выполняемые на каждом поколении ГА:

- оценка эффективности операторов,
- определение оператора-победителя,
- увеличение вероятности применения оператора-победителя,
- уменьшение вероятностей применения проигравших операторов.

Для формирования нечетких баз правил для задачи классификации необходимо определить структуру представления решения и способ кодирования базы правил. Метод кодирования базы правил влияет на эффективность поиска, а также на структуру получаемых решений. Основными подходами являются кодирование номеров нечетких термов либо кодирование форм и расположений нечетких термов; также встречаются гибридные подходы. Помимо этого, выделяют Питтсбургский (решение – база правил) и Мичиганский подходы (решение – отдельное правило), каждый из которых обладает рядом преимуществ и недостатков. На сегодняшний день наиболее перспективными являются методы, комбинирующие два этих подхода.

Во **второй главе** рассматриваются несколько подходов методов к кодированию нечетких баз правил на примере нескольких алгоритмов, использующих Питтсбургский подход и самоконфигурируемый генетический алгоритм для формирования баз правил. Указаны основные недостатки подобных подходов, а также предложено несколько схем кодирования баз правил в хромосому. Помимо этого, описан гибридный алгоритм построения баз правил, использующий Мичиганский этап в качестве оператора мутации в алгоритме Питтсбургского типа. Исследована эффективность самоконфигурирования алгоритма на тестовых задачах классификации в сравнении со стандартным алгоритмом.

Классификация объектов состоит в приписывании объекту *F*-мерного пространства R^F некоторого класса C_j из заранее заданного множества $C = \{C_1, \dots, C_k\}$. Пусть X= $\{X_1, \dots, X_F\}$ –набор входных переменных задачи, а

U_f, $f = 1, \dots, F$ – область определения f-ой переменной. Пусть $P_f = \{A_{f,1}, \dots, A_{f,T_f}\}, f = 1, \dots, F$ – разбиение области определения U_f на T_f нечетких множеств.

Нечеткое правило R_m, $m = 1, \dots, M$, где M – число правил, обычно имеет вид:

$$R_m: \text{ЕСЛИ } X_1 \text{это } A_{1,j_{m,1}} \text{ и} \dots \text{и } X_F \text{это } A_{F,j_{m,F}} \text{ ТО } Y \text{ это } C_{j_m},$$

где Y это выход, $C_{j_m} \in \mathrm{C}$ это номер класса для m-го правила, а $j_{m,f} \in [1, T_f]$ - это номер нечеткого множества из разбиения P_f, выбранный для переменной X_f. Таким образом, базу правил можно задать матрицей $J \in N^{M \times (F+1)}$

$$J = \begin{bmatrix} j_{1,1} & \cdots & j_{1,F} & C_{j_1} \\ \cdots & \cdots & \cdots & \cdots \\ j_{m,1} & \cdots & j_{m,F} & C_{j_m} \\ \cdots & \cdots & \cdots & \cdots \\ j_{M,1} & \cdots & j_{M,F} & C_{j_M} \end{bmatrix},$$

где элемент $j_{m,f}$ обозначает, что для правила R_m и переменной X_f было выбрано нечеткое множество $A_{f,j_{m,f}}$ и правилу поставлен в соответствие класс C_{j_m}.

Все числа в данной матрице целые, так что задача поиска оптимальной матрицы, описывающей нечеткую базу правил при заданном числе правил сводится к задаче целочисленной оптимизации. Данная задача может быть решена генетическим алгоритмом, наиболее важным моментом в этом случае является способ кодирования матрицы в хромосому.

Простейшее кодирование базы правил в хромосому ГА заключается в определении максимально возможного числа правил в базе и задании числа нечетких термов для каждой из входных переменных, правила записываются в хромосому последовательно, каждый ген кодирует номер терма, соответствующий номер класса для каждого правила также кодируется. То есть, каждый элемент $j_{m,f} \in [1, T_f]$ кодируется в бинарную строку длины l, такую, что $2^l \geq T_f + 1$. Добавление единицы к числу нечетких множеств необходимо для включения терма игнорирования значения переменной. При этом нечеткие термы для каждой переменной заранее определены и не подвергаются модификации в ходе поиска. Недостатком такого подхода является большая вычислительная сложность.

Альтернативная схема вместо номеров термов кодирует положения функций принадлежности. Предложена гибкая схема кодировки с применением сигмоидальных функций, позволяющая строить нечеткие множества различных форм. Задача поиска базы правил в данной постановке сводится к задаче безусловной вещественной оптимизации на единичном гиперкубе. Среди недостатков данного метода следует отметить низкую интерпретабельность получаемых баз правил, а также существенную вычислительную сложность. Приводятся результаты тестирования описанных методов построения нечетких баз правил на множестве тестовых задач.

С целью сохранения интерпретабельности целесообразным является использование фиксированных нечетких термов, легко соотносимых с вербальными понятиями «низкий», «средний», «высокий» и т.д. Одним из путей достижения более высокой точности при данном подходе является использование нескольких разбиений на нечеткие множества для каждой переменной. При этом диапазон изменения переменной разбивается на 2, 3, 4 и 5 нечетких множеств.

В **третьей главе** предлагается гибридный алгоритм построения баз правил использующий Питтсбургскую схему в качестве основной. Число правил в базе не фиксировано, но ограничено сверху и может меняться в ходе работы алгоритма. Номер класса храниться отдельно и не кодируется в хромосому. Помимо номера класса, каждому правилу приписывается вес правила, используемый в процессе нечеткого вывода. При использовании весов, правила записываются следующим образом:

$$R_m: \text{ЕСЛИ } X_1 \text{это } A_{1,j_{m,1}} \text{ и } \dots \text{и } X_F \text{это } A_{F,j_{m,F}} \text{ ТО } Y \text{ это } C_{j_m} \text{ с весом } CF_{jm},$$

Каждое нечеткое правило может быть рассмотрено как ассоциативное правило, в котором номер класса следует из значений, которые приняли переменные, то есть $A_q \to C_q$, где $A_q = (A_{q1}, \dots, A_{qn})$.

Таким образом, базу правил можно задать матрицей $J \in N^{M\times(F+2)}$

$$J = \begin{bmatrix} j_{1,1} & \cdots & j_{1,F} & C_{j_1} & CF_{j_1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ j_{m,1} & \cdots & j_{m,F} & C_{j_m} & CF_{j_m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ j_{M,1} & \cdots & j_{M,F} & C_{j_M} & CF_{j_M} \end{bmatrix}.$$

Таким образом, правило представляет собой строку целых чисел от 1 до 15. Включение терма игнорирования необходимо для уменьшения средней длины правил и повышения их обобщающей способности. Класс C_q и вес правила CF_q вычисляются по обучающей выборке. Степень принадлежности конкретного измерения x_p к правилу R_q вычисляется посредством оператора умножения следующим образом:

$$\mu_{Aq}(x_p) = \mu_{Aq1}(x_{p1}) \times \mu_{Aq2}(x_{p2}) \times \dots \times \mu_{Aqn}(x_{pn}),$$

где $\mu_{Aqi}(x_{pi})$ – это значение функции принадлежности нечеткого множества A_{qi} при входе x_{pi}. Для того, чтобы определить соответствующий номер класса и вес правила, рассчитаем достоверность (*confidence*) правила:

$$Conf(A_q \to Class\ k) = \frac{\sum_{x_p \in Class\ k} \mu_{Aq}(x_p)}{\sum_{p=1}^{m} \mu_{Aq}(x_p)}.$$

Если значение $Conf > 0.5$, то генерируется нечеткое правило с вектором A_q и классом C_q. При этом C_q является классом с максимальным значением $Conf$. Вес правила CF_q определяется как в последнем случае в пункте 3.1, следующим образом:

$$CF_q = Conf(A_q \to Class\ k) - \sum_{k=1, k \neq C_q}^{M} Conf(A_q \to Class\ k).$$

Если значения *confidence* меньше или равны 0.5 для всех классов, то вес правила оказывается отрицательным, соответствующее нечеткое правило считается неперспективным и не формируется.

Классификация по базе правил RS производится с использованием одного правила-победителя. Правило-победитель в этом случае определяется с учетом весов:

$$w = arg\max_{m}\left(CF_{j_m} \cdot \prod_{f} \left(\mu_{j_m, f}(x_f) \right) \right).$$

Инициализация в алгоритме производится с использованием эвристики с целью повышения качества решений на первом поколении алгоритма. Эвристика использует измерения из выборки для построения правил, устанавливая наиболее подходящий выбранному измерению нечеткий терм для каждой переменной. Вероятность того, что будет установлено B_j-е нечеткое множество:

$$P(B_j) = \frac{\mu_{B_j}(x_{pi})}{\sum_{k=1}^{14} \mu_{B_j}(x_{pi})}.$$

Эвристики также применяются для определения номера класса, соответствующего каждому правилу, а также для определения веса правила. В основе последних лежит мера достоверности правила (confidence), позаимствованная из алгоритмов построения ассоциативных правил.

Для данного алгоритма определены специальные процедуры скрещивания и мутации баз правил с учетом особенностей представления решений. В результате скрещивания может получиться индивид как с большим числом правил, так и с меньшим. Вероятность мутации зависит от текущего числа правил в базе.

Мичиганский этап состоит в модификации базы правил и применяется ко всем индивидам наряду с оператором мутации. В Мичиганской части каждое правил в базе представляет собой отдельный индивид. Пригодность индивидов определяется по числу измерений, верно классифицированных правилом. Используются три операции изменения базы правил: удаление наименее пригодных правил, добавление новых правил и замена наименее пригодных правил на новые. При этом добавление новых правил производится двумя методами: эвристическим и генетическим. Эвристический метод строит новые правила, используя неверно классифицированные индивиды, с применением схемы, аналогичной инициализации правила по выборке. Генетический метод строит новые правила с помощью классических операторов селекции, скрещивания и мутации правил, и формирует потомков, добавляемых в ту же базу правил, таким образом, производя одно поколение.

Общая схема алгоритма выглядит следующим образом:

1. Инициализация популяции с использованием выборки.
2. Оценивание индивидов-баз правил.
3. Генерация потомков с использованием селекции, скрещивания и мутации.
4. Применение Мичиганской части к каждому индивиду.
5. Формирование нового поколения.
6. Если условие остановки не удовлетворено, переход к шагу 2.

Мичиганская часть состоит из следующих этапов:

1. Пусть каждое правило в текущей базе – отдельный индивид.
2. Классифицировать обучающую выборку при помощи базы правил и назначить пригодность.
3. Сгенерировать или удалить несколько правил из популяции.
4. Вернуть полученную популяцию в Питтсбургскую часть.

Оценка пригодности (*Fitness*) происходит как свертка трех основных критериев – ошибки на обучающей выборке $f_1(i)$, числа правил $f_2(i)$ и суммарной длинны всех правил $f_3(i)$. Ошибка на обучающей выборке берется в процентах и с весовым коэффициентом, равным $w_1 = 100$. Два других критерия берутся с весовыми коэффициентами, равными единице, т.е. $w_2 = 1, w_3 = 1$. Ошибка берется в процентах, чтобы исключить влияние объема выборки.

$$Fitness(i) = f_1(i) * w_1 + f_2(i) * w_2 + f_3(i) * w_3.$$

В ходе тестирования самоконфигурирования для данного алгоритма выяснено, что пропорциональная селекция всегда показывает худшие результаты, чем другие два типа селекции – ранговая и турнирная, вследствие чего принято решение отказаться от использования пропорциональной селекции. Самоконфигурируемый алгоритм без пропорциональной селекции показывает в среднем лучшие результаты, чем стандартный алгоритм, однако лучший стандартный алгоритм превосходит самонастраивающийся.

Таблица 1. Расшифровка номеров конфигураций алгоритма

№ конфигурации	Тип селекции	Тип мутации	Мичиганская часть	Добавление правил
1	Пропорц.	Слабая	Добавление с вероятностью 0.5; Удаление с вероятностью 0.5	Эвристич. и генетический подходы равное число раз
2		Средняя		
3		Сильная		
4	Ранговая	Слабая		
5		Средняя		
6		Сильная		
7	Турнирная (2)	Слабая		
8		Средняя		
9		Сильная		

Таблица 2. Доля верно классифицированных на тестовой выборке

Алгоритм	Australian	German	Segment	Phoneme	Pageblocks	Satimage
1	0.839	0.701	0.791	0.784	0.920	0.790

2	0.839	0.707	0.790	0.790	0.931	0.785
3	0.841	0.710	0.767	0.789	0.932	0.782
4	**0.872**	0.759	0.880	0.805	0.942	0.831
5	0.869	0.748	0.888	0.807	**0.950**	**0.840**
6	0.861	**0.768**	**0.893**	**0.810**	**0.947**	0.835
7	0.857	0.743	0.878	0.806	0.939	0.827
8	0.860	0.746	0.885	**0.810**	0.945	0.836
9	**0.874**	0.743	0.883	0.806	0.950	0.830
Средний Стандартный	0.857	0.736	0.851	0.801	0.934	0.817
Средний без пропорцион. селекции	0.865	0.751	0.884	0.807	0.947	0.833
Самоконфи-гурируемый	0.857	0.749	0.876	0.806	0.945	0.833
Самоконфи-гурируемый Без пропор-циональной	**0.871**	0.751	0.881	**0.812**	**0.950**	0.835

Гибридный алгоритм построения нечетких баз правил в сравнении с другими схемами формирования позволяет получать более точные базы правил.

Таблица 3. Доля верно классифицированных на тестовой выборке

Задача	Кодирование номеров термов	Кодирование положений термов	Гибридный алгоритм
Australian	0.8400	0.8517	**0.8574**
Banknote	0.9419	**0.9945**	0.9671
Breast cancer	0.9496	0.9559	**0.9653**
Column 2c	0.7777	0.8073	**0.8150**
Column 3c	0.6417	**0.8032**	0.7795
German credit	0.7074	0.7097	**0.7496**
Glass	0.5359	0.6002	**0.7286**
Heart	0.7708	0.7598	**0.8543**
Ionosphere	0.7987	0.7669	**0.8794**
Iris	0.9363	**0.9426**	0.9022
Liver	0.5679	0.6289	**0.6967**
Pima	0.7317	0.7323	**0.7785**
Seeds	0.8834	**0.9174**	0.9095
Wine	0.8777	0.8505	**0.9130**

Таким образом, самоконфигурируемый гибридный эволюционный алгоритм построения баз правил позволяет получать точные и легко интерпретируемые решения. Самоконфигурирование позволяет избавить исследователя от многократных запусков алгоритма, в то время как активное

использование эвристик существенно экономит вычислительные ресурсы и позволяет повысить качество получаемых баз правил.

В **четвертой главе** рассмотрен адаптивный вероятностный метод селекции обучающих примеров для классификаторов с использованием выборки сокращенного размера на примере самоконфигурируемого гибридного эволюционного алгоритма. Также рассматриваются методы снижения объемов данных (Data Reduction, DR), призванные уменьшить объем обрабатываемой информации и как следствие повысить быстродействие. Одним из направлений является выбор примеров (Instance Selection, IS), суть которого заключается в формировании сокращенной выборки, с целью её использования в процессе обучения. Процесс селекции обучающих примеров состоит в последовательном обучении классификатора на подвыборках, формируемых с учетом качества классификации и назначении счетчиков измерениям. Селекция обучающих примеров позволяет существенно снизить объем вычислительных ресурсов для обучения при этом не только не снижая качества классификации, но также и повышая его на некоторых задачах.

Помимо этого, показано, что данный метод позволяет получать не только качественные решения в смысле общей точности классификации, но также может быть модифицирован для получения сбалансированных решений, в равной степени точно описывающих каждый класс.

Общая схема алгоритма селекции обучающих примеров. Пусть n – число измерений в обучающей выборке. Поставим в соответствие каждому измерению счетчик U_i, $i = 1 \ldots n$. Алгоритм содержит следующие шаги:

1) Установить все $U_i = 1$,
2) Сформировать подвыборку размером $H\%$ от обучающей случайным образом в зависимости от значений U_i,
3) Запустить процесс обучения на подвыборке в течении G итераций.
4) Проверить все текущие решения на всей обучающей выборке и сохранить лучшее,
5) Для всех измерений в подвыборке пересчитать значения U_i,
6) Если достигнуто максимальное число итераций, выход, иначе переход к шагу 2.

Размер подвыборки H фиксирован и не меняется в ходе работы алгоритма. Пересчет значений U_i производится следующим образом: для всех j из подвыборки если измерение j классифицировано верно, $U_j = U_j + 1$, иначе $U_j = 1$. При этом для классификации используется лучший для всей обучающей выборки индивид. Вероятность выбора измерения j в подвыборку рассчитывается по формуле:

$$p_i = \frac{1/U_i}{\sum_{j=\overline{1,n}} 1/U_j} \qquad (1)$$

Схема изменения вероятностей следует двум основным принципам: повышение вероятностей включения в подвыборку ранее неиспользованных измерений либо неверно классифицированных, и снижение вероятностей

ранее использованных и верно классифицированных измерений. Такая схема направляет процесс поиска и фокусирует алгоритм обучения на проблематичных областях пространства признаков.

Для гибридного эволюционного алгоритма генерирования нечетких баз правил селекция обучающих примеров имеет особое значение вследствие того, что данный алгоритм имеет несколько эвристик, использующих измерения, содержащиеся в выборке. В течение периода адаптации может происходить переобучение алгоритма на текущую подвыборку. Для предотвращения переобучения на каждом поколении в популяцию добавляется лучшее текущее решение для всей обучающей выборки. На каждом поколении алгоритма лучший найденный индивид для подвыборки проверятся на всей выборке.

Для тестирования эффективности алгоритма использовалась процедура стратифицированной 10-частной кросс-валидации. В данной процедуре каждая из 10 частей сохраняла соотношение числа измерений каждого класса, равное соотношению в изначальной выборке. Несбалансированность числа измерений по классам может привести к смещению результатов классификации в сторону мажоритарных классов.

Для генерации подвыборок в методе селекции обучающих примеров в случае наличия несбалансированных данных предлагается две схемы формирования: стратифицированная и балансирующая. Стратифицированная схема сохраняет соотношение числа измерений каждого класса равное такому в изначальной выборке. Балансирующая схема генерирует подвыборку так, что количество измерений в каждом классе равно. В случае невозможности соблюдения баланса по какому-то классу, берутся все измерения данного класса из обучающей выборки. Искусственное балансирование подвыборки приводит к тому, что алгоритм обучается на более сбалансированной выборке, что приводит к повышению точности классификации.

Для апробации самоконфигурируемого гибридного эволюционного алгоритма формирования нечетких баз правил для задачи классификации с активным обучением было выбрано 9 задач с репозиториев KEEL и UCI. Таблица 4 содержит описание задач.

Таблица 4. Задачи классификации

Задача	Число измерений	Число переменных	Число классов
Magic	19020	10	2
Page-blocks	5472	10	5
Penbased	10992	16	10
Phoneme	5404	5	2
Ring	7400	20	2
Satimage	6435	36	6
Segment	2310	19	7
Texture	5500	40	11
Twonorm	7400	20	2

В таблице 5 приведены результаты тестирования алгоритма без селекции обучающих примеров на всех 9 задачах. Ошибки усреднялись по двухкратной процедуре кросс-валидации и приведены в процентах. Число индивидов устанавливалось равным 100, число поколений 10000, максимальное число правил 40.

Таблица 5. Результаты без селекции обучающих примеров.

Задача	Обуч. ошибка	Тест. ошибка	Число правил	Длина правила	Время (минуты)
Magic	15.06	15.73	12.6	3.82	370.62
Page-blocks	3.52	3.96	10.1	3.49	94.48
Penbased	7.06	7.46	30.5	6.23	385.2
Phoneme	15.03	16.48	18.25	3.01	84.95
Ring	4.64	5.82	26.6	3.83	226.70
Satimage	12.22	14.22	20.4	11.06	345.82
Segment	4.55	6.45	22.2	6.69	146.18
Texture	6.50	7.75	25.8	14.90	352.26
Twonorm	4.42	6.06	17.4	7.40	254.88

Таблица 6 содержит результаты для алгоритма с активным обучением и балансированием подвыборки. Размер подвыборки – 20% от обучающей, длина периода адаптации – 100 поколений.

Таблица 6. Результаты с активным обучением и балансированием подвыборки.

Задача	Обуч. ошибка	Тест. ошибка	Число правил	Длина правила	Время (минуты)
Magic	14.62	15.08	17.3	3.63	129.65
Page-blocks	2.71	3.25	18.9	4.82	18.53
Penbased	3.27	3.81	30.8	6.11	91.42
Phoneme	15.63	16.88	24	2.84	19.62
Ring	3.23	5.08	30.2	3.85	68.23
Satimage	10.57	12.93	27.2	5.84	85.12
Segment	3.55	5.19	25.1	6.24	32.40
Texture	3.37	4.45	27	12.81	114.79
Twonorm	4.03	4.81	15	7.74	38.11

Из сравнения таблиц 5 и 6 видно, что применение селекции обучающих примеров позволяет существенно снизить ошибку как на обучающей, так и на тестовой выборке. При этом время работы алгоритма сокращается в 5 и более раз.

Таблица 7 содержит сравнение предложенного подхода с другими методами, протестированными на тех же задачах.

Таблица 7. Сравнение ошибки на тестовой выборке с аналогами

Задача	Предложенный метод	Parallel Fuzzy GBML	GP-Coach	IVFS-Coop	IVFS-Amp	FARC-HD	Bio HEL
Magic	15.08	**14.89**	20.18	19.82	20.82	15.49	-
Page-blocks	**3.25**	3.62	8.77	6.57	5.84	4.99	-
Penbased	3.81	**3.30**	17.80	17.00	21.73	3.96	6.00
Phoneme	16.88	**15.96**	-	-	-	17.86	-
Ring	**5.08**	5.25	-	12.57	16.89	5.92	-
Satimage	12.93	12.96	27.50	-	-	12.68	**11.60**
Segment	5.19	5.90	24.04	-	-	-	**2.90**
Texture	**4.45**	4.77	-	-	-	7.11	-
Twonorm	4.81	**3.39**	15.17	-	-	4.72	-

Полученные результаты численных экспериментов доказывают, что предложенный подход с активным обучением показывает эффективность классификации на уровне лучших алгоритмов из известных в мировом сообществе, и в некоторых случаях превосходит их. Проведенный статистический анализ результатов и тестирование алгоритма для различных объемов обучающей подвыборки и длины периода адаптации показывает преимущества предложенного подхода. Решенные практические задачи классификации с несбалансированными данными подтверждают корректность и реализуемость предложенного метода.

Заключение диссертации содержит основные результаты работы и выводы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Обоснована необходимость разработки новых методов формирования нечетких баз правил для задачи классификации с помощью эволюционных алгоритмов.
2. Разработано две новые схемы формирования нечетких баз правил с помощью самоконфигурируемого генетического алгоритма глобальной оптимизации, использующие Питтсбургский подход.
3. Разработан новый самоконфигурируемый гибридный эволюционный алгоритм формирования нечетких баз правил использующий комбинацию Питтсбургского и Мичиганского подходов, и отличающийся от известных схемой организации Мичиганской части, набором применяемых эволюционных операторов и процедурой самоконфигурирования.
4. Проведено комплексное исследование эффективности гибридного эволюционного алгоритма в сравнении с алгоритмами, использующими только Питтсбургский подход и показано его

превосходство в быстродействии и точности на большинстве задач за счет использования ряда эвристик.

5. Обосновано использование процедуры самоконфигурирования для гибридного эволюционного алгоритма формирования нечетких баз правил. Самоконфигурируемый алгоритм в среднем показывает результаты не хуже, чем стандартный алгоритм в среднем.
6. Обоснована целесообразность использования метода селекции обучающих примеров для формирования нечетких баз правил для задач с большим числом измерений и несбалансированными данными.
7. Разработана новая схема адаптивного вероятностного метода селекции обучающих примеров для эволюционных алгоритмов для задачи классификации. Разработанная схема применена к самоконфигурируемому гибридному эволюционному алгоритму, статистически доказано не только существенное увеличение быстродействия, но также и повышение точности классификации.
8. Разработана новая схема селекции обучающих примеров с балансированием обучающей подвыборки. Разработанная схема применена к самоконфигурируемому гибридному эволюционному алгоритму, статистически доказано, что данная схема позволяет формировать нечеткие базы правил, демонстрирующие схожую точность, как на мажоритарных, так и на миноритарных классах.
9. Разработанные в ходе исследования программные системы, реализующие предложенные подходы, успешно применены для решения реальных практических задач.

Таким образом, в диссертации разработаны, исследованы и апробированы новый самокофигурируемый гибридный эволюционный алгоритм формирования баз нечетких правил и новый адаптивный метод селекции обучающих примеров с возможностью формирования сбалансированных подвыборок ограниченного объема для обучения классификаторов, что является существенным вкладом в теорию и практику обработки информации и интеллектуального анализа данных.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в ведущих рецензируемых научных журналах и изданиях

1. Становов В.В., Бежитский С.С., Бежитская Е.А., Семенкин Е.С. Исследование эффективности многоагентного алгоритма решения задач глобальной поисковой оптимизации большой размерности // Системы управления и информационные технологии, № 4(62). – 2015.

2. Становов В.В., Бежитский С.С., Бежитская Е.А., Попов Е.А. Многоагентный алгоритм проектирования баз нечетких правил для задачи классификации // Вестник СибГАУ, Т. 16, №4, С. 842-848. – 2015.

3. Stanovov V., Semenkina O. Self-configuring hybrid evolutionary algorithm for multi-class unbalanced datasets // Вестник СибГАУ. 2015. Т. 16, № 1. С. 131–136.

4. Stanovov V., Skraba A., Kofiac D., Znidarsic A., Maletic M. Rozman C., Semenkin E., Semenkina M. Application of Self-Gonfiguring Genetic Algorithm for Human Resource Management // Journal of Siberian Federal University. Mathematics and Physics 2015, 8(1), 98-107.

5. Становов В. В., Семенкина О.Э. Самоконфигурирующийся гибридный эволюционный алгоритм формирования нечетких классификаторов с активным обучением для несбалансированных данных // Вестник СибГАУ 2014. № 5(57). С. 128–135.

6. Становов В.В., Семенкин Е.С. Самонастраивающийся эволюционный алгоритм проектирования баз нечетких правил для задачи классификации // Системы управления и информационные технологии. 2014. Т. 57. № 3. С. 30-35.

7. Становов В.В., Семенкин Е.С. Self-adjusted evolutionary algorithms based approach for automated design of fuzzy logic systems // Вестник СибГАУ. 2013. № 5 (51), С. 148-152.

8. Шкраба А., Становов В.В., Жнидаршич А., Розман Ч., Кофьяч Д. – Рассмотрение стратегии оптимального управления строго иерархической системы управления человеческими ресурсами // Вестник СибГАУ, Т. 17, №1, С. 97-102. – 2016.

9. Stanovov V., Semenkin E., Semenkina O., Self-Configuring Hybrid Evolutionary Algorithm for Fuzzy Imbalanced Classification with Adaptive Instance Selection., Journal of Artificial Intelligence and Soft Computing Research 6(3) (JAISCR), June 2016, pp. 173-188. – 2016.

Публикации в изданиях, индексируемых в международных базах

10. Stanovov V., Semenkin E., Semenkina O. Instance selection approach for self-configuring evolutionary fuzzy rule based classification systems // 4th International Congress on Advanced Applied Informatics July 12-16, 2015, Okayama Convention Center, Okayama, Japan. **(Web of Science, Scopus).**

11. Škraba A., Semenkin E., Kofjac D., Semenkina M., Znidaršic A., Maletic M., Akhmedova Sh., Rozman C., Stanovov V. Modelling and Optimization of Strictly Hierarchical Manpower System. 12th International Conference on Informatics in Control, Automation and Robotics, ICINCO. – 2015. **(Scopus).**

12. Stanovov V., Semenkin E., Semenkina O. Instance Selection Approach for Self-configuring Hybrid Fuzzy Evolutionary Algorithm for Imbalanced Datasets // Advances in Swarm and Computational Intelligence, LNCS 9140, 2015, pp. 451–459. **(Web of Science).**

13. Stanovov V., Semenkin E., Semenkina O. Self-configuring hybrid evolutionary algorithm for fuzzy classification with active learning // 2015 IEEE Congress on Evolutionary Computation (CEC’2015, Japan). **(Scopus).**

14. Semenkin E., Stanovov V. Fuzzy rule bases automated design with self-configuring evolutionary algorithm // Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO’2014, Austria), pp. 318-323. **(Scopus).**

15. Stanovov V., Semenkin E. Hybrid self-configuring evolutionary algorithm for automated design of fuzzy logic rule base // 11th international conference on Fuzzy Systems and Knowledge Discovery (FSKD'2014, China), pp. 317-321. **(Scopus, Web of Science).**

16. Akhmedova Sh., Semenkin E., Stanovov V., Fuzzy Rule-based Classifier Design with Co-Operative Bionic Algorithm for Opinion Mining Problems // 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016, Lisbon, Portugal). – 2016. **(Scopus).**

17. Akhmedova Sh., Stanovov V., Semenkin E., Fuzzy Rule-Based Classifier Design with Co-operation of Biology Related Algorithms // Advances in Swarm Intelligence, pp.198-205. – 2016. **(Scopus).**

18. Stanovov V., Sopov E., Semenkin E. Multi-Strategy Multimodal Genetic Algorithm for Designing Fuzzy Rule Based Classifiers // IEEE Symposium Series on Computational Intelligence (SSCI 2015, South Africa). – 8-10 December. – 2015. **(Scopus, Web of Science).**

Публикации в сборниках трудов конференций

19. Становов В. В. Применение самоконфигурируемого эволюционного алгоритма построения нечетких баз правил для решения задач классификации с несбалансированными данными // Материалы XVIII междун. науч. конф. Решетневские чтения, (Красноярск, 11-14 ноября 2014 г.) . – Т. 2, С. 127-129.

20. Становов В. В. Решение задачи определения уровня озона в атмосфере при помощи самоконфигурирующегося эволюцинного алгоритма построения нечетких баз правил // Материалы XVIII междун. науч. конф. Решетневские чтения, (Красноярск, 11-14 ноября 2014 г.). – Т. 2, С. 355-357.

21. Становов В.В., Семенкин Е.С., Бежитский С.С. Гибридный эволюционный алгоритм формирования нечетких баз правил для задачи классификации // Теория и практика системного анализа. Труды III Всероссийской научной конференции с международным участием (ТПСА-2014). -2014. - Том 2. - С. 115-122.

22. Становов В. В., Семенкин Е.С. Особенности генерации случайных чисел при распараллеливании эволюционных алгоритмов // Системный анализ и интеллектуальные технологии. Труды V Международной конференции (САИТ-2013). - 2013. - Том 2. - С. 411-418.

Зарегистрированные программные системы

23. Становов В. В., Панфилов И.А., Сопов Е.А. Распределенная самонастраивающаяся система формирования нечетких баз правил при помощи генетического алгоритма. Свидетельство №2014610096 о гос. регистрации в Реестре программ для ЭВМ от 03.03.2014.

24. Становов В. В., Семенкин Е. С. Самонастраивающаяся система формирования символьных выражений для решения задач классификации и регрессии методом генетического программирования. Свидетельство № 2013619070 о гос. регистрации в Реестре программ для ЭВМ от 25.09.2013.

25. Становов В. В., Сергиенко Р. Б. Распределенная программная система автоматического формирования нечетких систем методом генетического

программирования. Свидетельство № 2013611035 о гос. регистрации в Реестре программ для ЭВМ от 9.01.2013.

Становов Владимир Вадимович
Самонастраивающиеся эволюционные алгоритмы формирования систем на нечеткой логике
Автореферат
Подписано к печати 21.10.2016. Формат 60х84/16
Уч. изд. л. 1.0 Тираж 100 экз. Заказ № _____
Отпечатано в отделе копировальной и множительной техники СибГАУ.
660037, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31