

На правах рукописи

ПОЛЯКОВА АНАСТАСИЯ СЕРГЕЕВНА

**КОЛЛЕКТИВНЫЕ МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА
ДАННЫХ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ**

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Красноярск 2019

Работа выполнена в ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» (СибГУ им. 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

Сташков Дмитрий Викторович
канд. техн. наук
АО «СИНЕТИК» (г. Новосибирск),
ведущий специалист по АСУ ТП
горнорудной промышленности

Ведущая организация: ФГБОУ ВО «Томский государственный
университет систем управления и
радиоэлектроники»

Защита состоится «22» ноября 2019 г. в 15:00 часов на заседании диссертационного совета Д 212.249.05, созданного на базе Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева по адресу: 660037 г. Красноярск, проспект имени газеты «Красноярский рабочий», 31.

С диссертацией можно ознакомиться в библиотеке Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева и на сайте <https://www.sibsau.ru>

Автореферат разослан «__» _____ 2019 г.

Ученый секретарь
диссертационного совета,
кандидат технических наук, доцент

Панфилов
Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В современной науке существует значительное количество методов и инструментов интеллектуального анализа данных (ИАД). Достаточно очевидным обобщением является использование нескольких алгоритмов одновременно, т.е. применение подходов коллективного принятия решений. Это позволяет повысить качество решения задачи ИАД и значительно упрощает поиск компромисса между точностью, простотой и интерпретируемостью каждой отдельной модели.

С точки зрения теории искусственного интеллекта (ИИ) коллектив можно представить в виде набора автономных агентов, работающих над какой-то общей задачей, например, в мультиагентной системе. Можно сказать, что коллектив является интеллектуальным, если он может эффективно использовать интеллект своих членов в ходе решения задачи.

На практике задачи ИАД, имеющие различные характерные особенности (неполнота и неточность исходных данных, высокая вычислительная сложность получения результатов и их формализации, и т.п.), довольно трудно решить с помощью отдельной технологии.

Коллектив из технологий интеллектуального анализа данных представляет собой набор моделей, каждая из которых способна решить поставленную задачу, а их комбинация позволяет повысить эффективность коллектива в целом.

Простейшими методами формирования коллектива являются методы простого и взвешенного голосования. Более сложными методами являются бэггинг (Breiman, 1996), бустинг (Freund & Schapire, 1996; Quinlan, 1996; Breiman, 1996; Kearns, 1988), случайные леса (Breiman, 2001) и стэкинг (Wolpert, 1992). Известными разновидностями бустинга являются AdaBoost (Schapire, 2001) и градиентный бустинг (Friedman, 2002). Похожим на бэггинг методом является метод случайных подпространств (Skurichina, 2002).

Важным моментом при формировании коллектива является не только выбор его членов и их обучение, но также и способ объединения в коллектив. Бустинг и бэггинг являются методами, которые обучают членов коллектива независимо друг от друга, в этом смысле они не производят искусственную специализацию членов коллектива. Альтернативой является, например, перераспределение обучающих примеров между обучающими методами для членов коллектива, приводящие к специализации их на определенной части задачи. Кроме того, возможен вариант, при котором члены коллектива и способ их комбинирования настраиваются одновременно.

В то же время системы коллективного принятия решения на основе нечеткой логики (НЛС) имеют ряд преимуществ:

1. За счет специальной подсистемы (базы правил) НЛС позволяет накапливать опыт решения, дообучаться и применять правила на других задачах.
2. Модели на базе НЛС формализуются на языке, близком к экспертному, и в тех случаях, когда точность решения критична (неприемлемы

потери), процедуры поддержки принятия решения, близкие к естественные языку, за счет их интерпретируемости являются более предпочтительными с точки зрения сертификации и анализа специалистами предметной области.

3. Системы с применением нечеткой логики могут в автоматическом режиме отбрасывать неудачные модели или не учитывать влияние слабых агентов при коллективном выводе.
4. Стандартные формы коллективов испытывают затруднения в случае, когда один агент очень сильный, а все остальные – слабые, и объединение в коллектив дает не улучшенное решение, а ослабление решения сильного агента, в то время как процедура на нечеткой логике может формировать решение не хуже лучшего агента.

Отличительной чертой НЛС является то, что модель строится по принципу «белого ящика». НЛС позволяют координировать и объединять опыт экспертов предметной области, а также способны моделировать нелинейные функциональные зависимости произвольной сложности. Все эти свойства дают возможность рассматривать использование НЛС в качестве метода коллективного принятия решений, что позволило бы существенно повысить качество принимаемых решений, а также их интерпретируемость.

За счет автоматизации процессов проектирования ИИТ отпадает необходимость привлечения экспертов и снижаются вычислительные затраты в ходе тестирования для определения наиболее эффективного метода. В то же время, возникающие при этом задачи выбора эффективных вариантов коллективов требуют применения мощных и универсальных оптимизационных процедур адаптивного типа. Для этого целесообразным представляется использование адаптивных стохастических алгоритмов решения задач глобальной оптимизации алгоритмически заданных функций смешанных переменных, в частности – эволюционных алгоритмов (ЭА). ЭА позволяют в автоматическом режиме выбирать конфигурацию и настраивать параметры коллективных моделей принятия решений на основе нечеткой логики.

Таким образом, разработка и исследование методов автоматизированного формирования коллективных моделей принятия решений на основе нечеткой логики с использованием эволюционных алгоритмов является **актуальной научно-технической задачей**.

Целью диссертационного исследования является повышение эффективности интеллектуальных технологий анализа данных путем автоматизированного формирования коллективов алгоритмов с помощью специальных систем на нечеткой логике.

Для достижения поставленной цели необходимо решить комплекс **задач**.

1. Провести обзор современных методов анализа данных и форм их коллективного взаимодействия.
2. Разработать и реализовать алгоритм коллективного вывода на основе нечеткой логики для решения задач классификации и регрессии.

3. Разработать и реализовать процедуру выбора алгоритмов классификации или регрессии для включения в состав коллектива.
4. Разработать и реализовать процедуру автоматизированного выбора показательных примеров в опорное множество для формирования коллективного вывода.
5. Реализовать в виде программной системы процедуры коллективного принятия решения на основе нечеткой логики.
6. Исследовать работоспособность предложенного алгоритма на тестовых и практических задачах.

Область исследования. Работа выполнена в соответствии со следующими пунктами паспорта специальности 05.13.01:

- разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации;
- методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических, экономических, биологических, медицинских и социальных системах.

Методы исследования.

При выполнении работы использовались методы и подходы теории вероятностей, методы статистической обработки данных, эволюционных вычислений, оптимизации, нечеткой логики, системного анализа данных, выявления закономерностей в исходных данных.

Научная новизна работы.

1. Разработана новая схема формирования коллективного вывода на основе нечеткой логики, отличающаяся иерархической процедурой интеграции правил коллективного вывода.
2. Разработана новая эволюционная процедура выбора агентов для формирования эффективных коллективов, отличающаяся от известных использованием нескольких критериев эффективности.
3. Разработана новая эволюционная процедура автоматизированного формирования базы правил, отличающаяся от известных применением двух уровней эволюции и способом представления решения в бинарном пространстве поиска.
4. Разработана новая система на основе нечеткой логики для формирования коллективов моделей и алгоритмов анализа данных для решения задач классификации и регрессии, отличающаяся от известных адаптивной процедурой формирования коллективного решения.
5. Разработана комплексная процедура автоматизированного формирования системы коллективного вывода на основе нечеткой логики, отличающаяся возможностью эффективного перераспределения вычислительных ресурсов.

Значение для теории состоит в разработке комплексного подхода к решению задачи ИАД с помощью нового коллективного метода принятия решения на основе нечеткой логики, который является эффективным

обобщением отдельных методов интеллектуального анализа данных. Сформированное итоговое решение, получаемое на основе коллектива моделей, эффективнее, так как коллектив всегда работает не хуже самой лучшей модели. Результаты, полученные при выполнении диссертационной работы, создают теоретическую основу для разработки новых технологий распределения ресурсов и распараллеливания процессов в ходе решения сложных трудно формализуемых задач анализа данных.

Практическая ценность.

Разработанные алгоритмические схемы, которые реализованы в виде программной системы на языке программирования Python, являются полноценной библиотекой. Программная система позволяет формировать коллективный вывод при решении задач интеллектуального анализа данных и проектирование коллективов моделей на основе нечеткой логики для задач классификации и регрессии. Программная система протестирована на задачах распознавания лиц по изображению и прогнозирования аффективного (эмоционального) поведения человека по голосу, а также на задаче моделирования технологического процесса металлургического производства.

Реализация результатов работы. В диссертационной работе была разработана программная система, которая прошла регистрацию в Роспатенте.

Диссертационная работа выполнена в рамках проектов:

1. Проект №2.1680.2017/ПЧ «Разработка теоретических основ автоматизации комплексного моделирования сложных систем методами вычислительного интеллекта», в рамках Государственного задания СибГУ на 2017-2019 гг.

2. Проект №14.574.21.0037 «Распределенные самоконфигурируемые многоагентные технологии проектирования и управления интеллектуальными информационными сетями» в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014 - 2020 годы».

3. Российско-словенский проект "Application of artificial intelligence methods on small field robot" ARRS-MS-BI-RU-JR-Prijava/2018/65 на 2019-2020 год.

4. Проект «Разработка системы автоматического распознавания и классификации дефектов элементов панелей солнечных батарей», Фонд содействия развитию малых форм предприятий в научно-технической сфере по программе «У.М.Н.И.К» 2015-2017гг.

5. Российско-германский проект «Advanced feature selection techniques for multimodal emotion recognition» в рамках конкурса германской службы академических обменов «Программа Эйлера» 2015-2016 гг.

Основные защищаемые положения:

1. Предложенная схема иерархической процедуры коллективного вывода на основе системы нечеткой логики позволяет повысить эффективность коллективного вывода в сравнении одноуровневым принятием решения.

2. Разработанная процедура эволюционного выбора агентов для формирования эффективных коллективов моделей позволяет повысить точность работы коллектива.
3. Разработанная процедура автоматизированного формирования базы правил с применением двух уровней эволюции и предложенным способом представления решения в бинарном пространстве поиска позволяет формировать эффективные базы правил с минимальным количеством правил и высоким уровнем обобщения самих правил без потери точности.
4. Разработанная интегрированная процедура автоматизированного формирования системы коллективного вывода на основе нечеткой логики позволяет в автоматизированном режиме управлять вычислительными ресурсами при обучении коллектива и получать коллективы с высокой точностью решения задач классификации и регрессии.

Апробация работы. Основные положения и результаты работы прошли всестороннюю апробацию на Всероссийских и Международных конференциях: «8th International Congress on Advanced Applied Informatics - "7th International Conference on Smart Computing and Artificial Intelligence" SCAI 2019» (2019 г., Toyama, Japan), «The International Workshop "Advanced Technologies in Material Science, Mechanical and Automation Engineering"» (2019 г., г. Красноярск), «The International Workshop on Mathematical Models and their Applications (IWMA)» (2014, 2016 г., г. Красноярск), «Международная научно-практическая конференция "Решетневские чтения"» (2012-2016 гг., г. Красноярск), «Международная научно-практическая конференция "Актуальные проблемы авиации и космонавтики"» (2011-2016 гг., г. Красноярск), «Всероссийская научно-практическая конференция "Информационно-телекоммуникационные системы и технологии" ИТСиТ-2014» (2014 г., г. Кемерово), «Всероссийская научно-техническая конференция "Приоритетные направления развития науки и технологий"», (2013 г., г. Тула).

Публикации. По теме диссертации опубликовано 19 печатных работ, из них три статьи в журналах перечня ВАК РФ и три в изданиях, индексируемых в международных базах цитирования Web of Science и/или Scopus. Получено одно свидетельство о регистрации программной системы в Роспатенте.

Структура и объем работы. Диссертация содержит 150 страниц текста и состоит из введения, четырех глав, заключения, списка литературы из 138 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении представлена общая характеристика проблемы, обоснована актуальность выбранной темы, определены цель и задачи исследования. Сформулированы основные положения, выносимые на защиту, научная новизна и практическая значимость полученных результатов.

В первой главе рассматриваются существующие коллективные формы принятия решения и комбинации отдельных технологий в коллективе.

Предлагается и описывается альтернативный подход к коллективному принятию решения на основе нечеткой логической системы (Fuzzy Rule-Based System, FRBS), а также приводится его алгоритмическая схема для решения задач классификации и регрессии (рисунок 1).

Также предлагается и исследуется модификация FRBS с помощью гибридизации FRBS и процедуры итогового формирования решения с помощью среднего (*mean*) и взвешенного среднего (*Wmean*), позволяющая получить решение точнее, чем решение лучшего агента из коллектива. Предложенный подход называется “*FRBS+Wmean*” или “*FRBS+mean*” в зависимости от способа формирования итогового решения.

“*FRBS+Wmean*” формируется таким образом, чтобы эффективно объединять алгоритмы (агентов) в коллектив. Нечеткий контроллер (НК) принимает решение о выборе классификатора или алгоритма регрессии в зависимости от близости тестового объекта к объектам обучающей выборки и успешностью классификатора на ближайшем объекте.

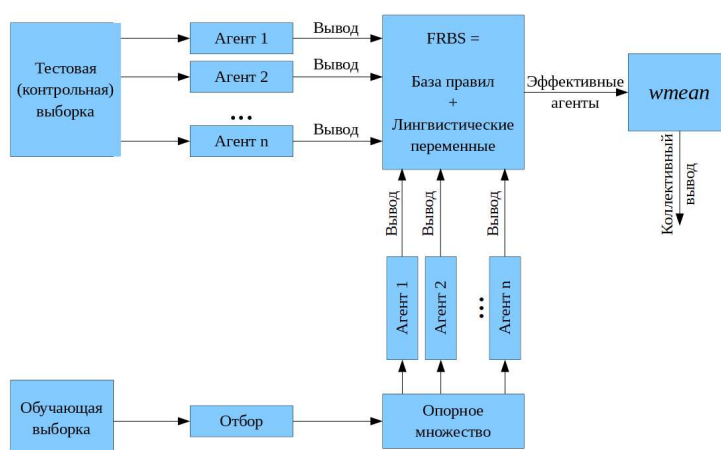


Рисунок 1. Общая схема коллективного вывода на основе подхода “*FRBS+Wmean*”

Для НК формируются три лингвистические переменные (ЛП) для входа и одна для выхода:

1. *Distance*: близость объекта тестовой выборки к ближайшей точке из обучающей выборки.
2. *Error*: разница между выходом модели (агента) на тестовой выборке и в ближайшей точке (точках) обучающего множества (ошибка агента на объекте выборки).
3. *Weight_agent*: вес агента (вычисляется на основе ошибок агентов на обучающем множестве).
4. *Confidence*: степень доверия к агенту, которая вычисляется с помощью нечеткой процедуры вывода, учитывая 3 ЛП входа.

Выходом работы НК для каждого объекта выборки из тестового множества является степень доверия к агенту. Нечеткий вывод степени доверия осуществляется по каждому агенту. Выбирается тот агент, степень доверия к которому выше.

Варьируемыми параметрами алгоритма “*FRBS+Wmean*” являются два показателя: количество ближайших точек из обучающего множества к объекту из тестовой выборки (*nPoints*), а также количество агентов, композиция которых будет применяться для получения решения для объекта из тестовой выборки (*nAgent*).

Особенностью данной системы является то, что FRBS на основе опорного множества, принимает решение о выборе множества эффективных агентов *nAgent*. При построении решающего правила для каждой тестовой точки, решение будет приниматься тем агентом, степень доверия к которому выше. Эта схема работает в том случае, если принятие решения происходит по одному агенту, то есть $nAgent = 1$.

В случае $nAgent > 1$ для текущего решения с помощью системы на нечеткой логике отбирается *nAgent* лучших агентов (*nAgent* – параметр алгоритма), а из них итоговое решение принимается средним “*FRBS+mean*” или взвешенным средним “*FRBS+Wmean*”. При этом веса агентов определяются пропорционально их уверенности на данном примере.

Таким образом, система на нечеткой логике, позволяющая получить степень доверия агента, зависит от пяти параметров, переданных в НК:

$$FRBS(Distance, Error, Weigh_agent, nAgent, nPoints) = \mu$$

Задача FRBS состоит в максимизации точности решения задачи классификации или регрессии путем эффективного распределения вычислительных ресурсов между агентами. На рисунке 1 представлена общая схема описанного выше подхода, для которой исходная выборка разбивается на 3 части: обучающая, тестовая и контрольная.

Из обучающей выборки производится отбор точек в опорное множество. Опорное множество позволяет оценить эффективность агентов в различных точках тестового пространства. То, как использовать информацию об эффективности агентов на точках опорного множества, и то, как эту информацию проецировать на точки тестового пространства, определяется с помощью FRBS. Формирование опорного множества может выполняться различными методами отбора. На основе обучающей выборки происходит обучение агентов в отдельности. На тестовой выборке происходит оценка эффективности обучения агентов и обучение базы нечетких правил в FRBS. На контрольной выборке происходит итоговая оценка эффективности работы системы в целом.

База правил является достаточно простой и интерпретируемой, и выглядит следующим образом:

IF *error* - высокая THEN *confidence* – низкая

IF *error* - средняя AND *distance* – близко AND *weight_agent* - высокий THEN *confidence* – высокая

IF *error* - средняя AND *distance* – средне THEN *confidence* – средняя

IF *error* - низкая AND *distance* – близко AND *weight_agent* - высокий THEN *confidence* – высокая

IF *error* - низкая AND *distance* – близко AND *weight_agent* - низкий THEN *confidence* – средняя

IF *error* - низкая AND *distance* – средне AND *weight_agent* - высокий THEN *confidence* – высокая

IF *distance* – далеко THEN *confidence* – низкая

Предложенная процедура формирования коллективного решения на основе гибридизации FRBS и формирования итогового решения с помощью среднего и взвешенного среднего, позволяет получить решение точнее, чем решение лучшего агента из коллектива.

Обучение алгоритмов производится на всем обучающем множестве $X_{\text{обуч}}$.

Эффективность разработанного подхода была исследована на основе известных наборов тестовых задач регрессии и классификации.

Во второй главе предложена эволюционная процедура для автоматизированного формирования состава коллектива и выбора опорного множества.

Высокая временная и вычислительная сложность коллективного подхода на основе нечеткой логики для решения задачи классификации или регрессии является следствием того, что для вычисления лингвистической переменной Error («близость объекта обучающей выборки к тестовой выборке») приходится использовать весь набор данных, то есть необходимо выполнять сравнение между каждой парой экземпляров в наборе данных для поиска ближайшего объекта.

В предложенную схему коллективного вывода на основе подхода “FRBS+Wmean” (рисунок 1) вводится понятие “опорное множество” (опорная выборка), под которым подразумевается подмножество обучающего множества. Опорное множество будет использоваться в коллективном выводе с помощью системы на нечеткой логике, в соответствии с чем для точки из тестового множества определяется одна (в случае $nPoints=1$) или несколько (в случае $nPoints>1$) ближайших точек не из всего обучающего множества, а из опорного. В зависимости от того, насколько эта точка близка к объекту из тестового множества и насколько успешно справляется на ней алгоритм, определяется уверенность агента на данной тестовой точке. С помощью НК итоговое решение принимает либо агент с наибольшей уверенностью, либо $nAgent$ лучших агентов методом усреднения.

В качестве эволюционной процедуры для автоматизированного отбора экземпляров в опорное множество из обучающего при решении задач классификации или регрессии используется генетический алгоритм безусловной однокритериальной оптимизации (ГА) со специальной схемой кодирования.

Для организации процедуры отбора экземпляров выборки в опорное множество генетическим алгоритмом генотип хромосомы представляет собой двоичную последовательность, а фенотип - целочисленные значения, которые определяют номер точки из обучающего множества для ее включения в опорное множество. Количество точек, кодируемых в бинарную строку определяется пользователем.

Исследование эффективности предложенного способа кодирования хромосомы ГА проводилось на основе трех известных тестовых задач регрессии.

Вычислительные эксперименты показывают, что при эффективном выборе точек в опорное множество можно существенно снизить затраты вычислительных ресурсов при сохранении или повышении точности результата.

За счет правильного формирования опорного множества “слабые” агенты будут исключены. С другой стороны, за счет правильного выбора агентов даже в непоказательных точках опорного множества будет принято верное решение. Соответственно оба этапа (формирование опорного множества и отбор агентов) являются взаимосвязанными.

Эволюционные алгоритмы позволяют проектировать и отбирать эффективных агентов, исходя из того насколько успешно агенты справляются с поставленной задачей. Во время взаимодействия агентов осуществляется отбор наиболее успешных, которые затем используются для генерации нового поколения агентов, среди которых опять применяются те же самые процедуры оценки и отбора. В итоге наиболее успешное поколение решает задачу и достигает целей наиболее эффективным образом.

В главе 2 в качестве эволюционной процедуры для автоматизированного формирования состава коллектива предлагается многокритериальный эволюционный генетический алгоритм (Multi-Objective Evolutionary Genetic Algorithm, MOEGA) - генетический алгоритм недоминируемой сортировки (Nondominated Sorting Genetic Algorithm, NSGA-II). Этот алгоритм позволяет автоматизировать формирование состава коллектива, тем самым экономя вычислительные ресурсы (позволяют минимизировать количество агентов), и решать поставленные задачи достаточно качественно (повышая способность к обобщению результата).

При формировании коллектива необходимо выбирать, кто из агентов будет входить в состав коллектива при принятии конечного решения. Состав коллектива с использованием MOEGA кодируется в бинарную строку – один агент закодирован с помощью одного бита, где 0 означает, что этот агент не принимает участие в принятии решения, а 1 - участвует. Длина бинарной строки постоянна, так как кодируется для максимально доступного количества агентов.

В ходе исследований на основе тестовых задач регрессии и классификации показано, что выбор эффективных агентов с помощью многокритериальной эволюционной процедуры позволяет существенно повысить точность работы и обобщающую способность коллектива.

В третьей главе предложен двухступенчатый подход на основе эволюционных алгоритмов однокритериальной и многокритериальной оптимизации для автоматического формирования базы правил и дальнейшего их отбора с целью повышения эффективности решения задачи.

При формировании базы правил (БП) для нечеткой системы необходимо решить две задачи: формирование исходного множества нечетких правил и выбор правил из этого заданного (исходного) множества.

Если само генерирование нечетких правил может производиться даже случайным образом, то при отборе конечного множества нечеткой БП необходимо учитывать такие критерии как точность, выраженная путем вычисления среднеквадратичной ошибки правил (Mean Squared Error, MSE), сложность, выраженная количеством выбранных правил, и другие.

Общая схема работы формирования и оптимизации базы правил с применением двухступенчатого подхода в процедуре нечеткого вывода (Genetic Fuzzy Rule-Based Systems, GFRBS) представлена на рисунке 2.

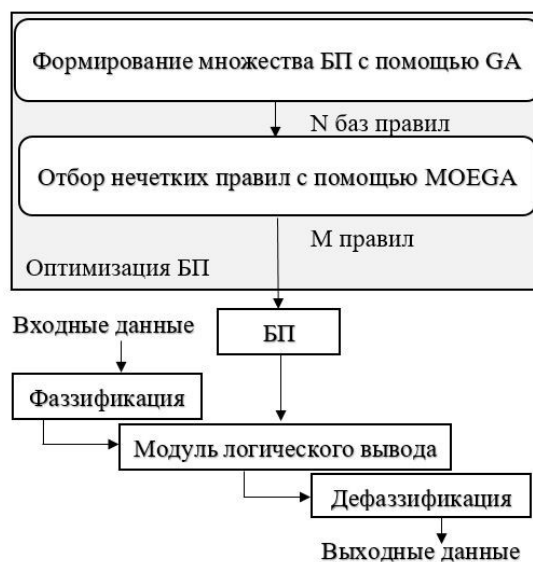


Рисунок 2. Общая схема коллективного принятия решения с отбором нечетких правил с помощью MOEGA.

Исходное генерирование правил производится с помощью ГА однокритериальной оптимизации. На следующем этапе для отбора правил был использован и адаптирован MOEGA.

Задача первой ступени, то есть ГА однокритериальной оптимизации – сгенерировать максимально выразительные правила. Это означает, что одно правило работает для большого количества случаев.

Для этого используется специальный вид кодирования, который, во-первых, позволяет кодировать практически произвольное правило, а во-вторых, БП на первой ступени ГА состоят из малого количества правил. В этом случае успешный индивид содержит хотя бы одно эффективное правило. В целом БП

может плохо справляться с задачей, но если она имеет хотя бы одно эффективное правило, то может быть полезной на второй ступени.

Задача второй ступени - отыскать эффективные правила и сформировать из них единую БП с помощью MOEGA. При этом обе ступени должны работать последовательно.

Для выбора нечетких правил предлагается использовать эволюционный алгоритм многокритериальной оптимизации NSGA-II, система кодирования которого представлена на рисунке 3, а для генерирования предлагается использовать генетический алгоритм однокритериальной безусловной оптимизации, система кодирования которого представлена на рисунке 4.

Бинарная система кодирования позволяет подобрать эффективные нечеткие правила для системы нечеткого вывода при коллективном принятии решений. Длина хромосомы (индивида), представляющая собой базу правил, равна $Z \times N$ бит, где Z – количество баз правил, отобранных после первой ступени. Из N баз происходит парсировка Z правил, каждому из которых соответствует один бит (значение 0 означает исключение правила из базы, а 1 – включение).

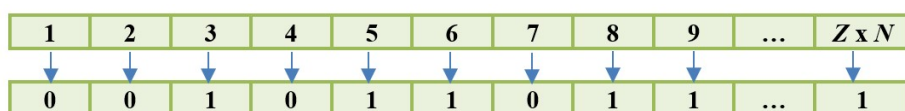


Рисунок 3. Структура кодирования индивида для MOEGA – NSGA-II

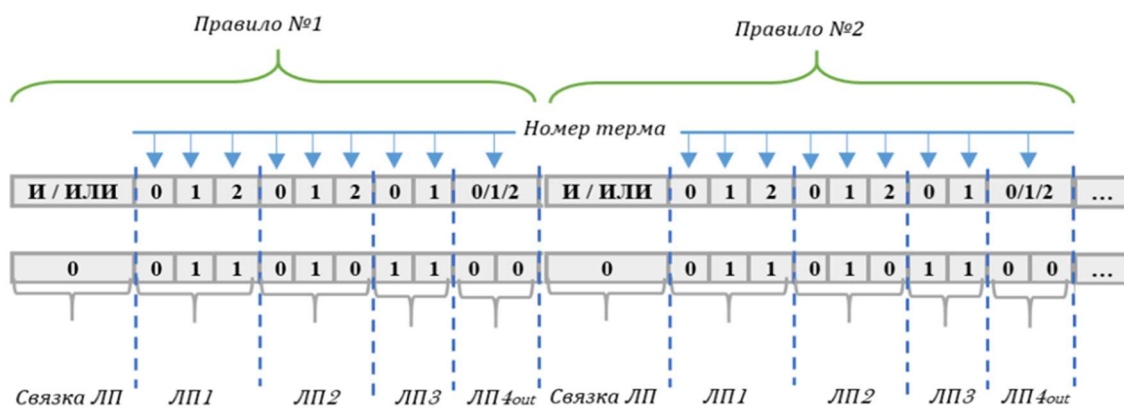


Рисунок 4. Кодирование для генетического алгоритма однокритериальной оптимизации при генерировании исходного множества правил

Исследование эффективности применения двухступенчатого подхода на основе однокритериального ГА и многокритериального ГА в предложенном подходе GFRBS проводилось на задаче оптимизации БП при решении задач классификации в сравнении с исходной БП.

Проведенные исследования показывают, что гибридизация FRBS и формирования итогового решения с помощью среднего и взвешенного среднего позволяет найти решение точнее лучшего из агентов даже без оптимизации состава коллектива, а также позволяет улучшить результаты GFRBE при

эффективном выборе количества агентов, принимающих решение. При этом, использование оптимизации БП в GFRBE позволяет получать базы правил эффективнее решающие текущую задачу с меньшим количеством правил относительно исходной базы правил. Также показано, что при уменьшении количества правил с помощью оптимизации БП в GFRBE сохраняется точность решения задачи.

Кроме того, в третьей главе рассматриваются несколько подходов и методов к построению нечетких БП и настройке семантики ЛП с применением различных эволюционных алгоритмов для формирования БП.

В ходе оптимизации БП и ЛП системы “FRBS+Wmean” проводится сравнительный анализ эффективности построения нечетких систем коллективного принятия решений с различными последовательностями реализации этапов автоматического формирования БП и ЛП. Оптимизация базы правил производится с помощью ГА однокритериальной безусловной оптимизации, а ЛП – с помощью алгоритма дифференциальной эволюции (ДЭ).

Построение нечеткой системы предлагается рассматривать следующим образом: на первом шаге генерируется начальная база правил и первоначальный набор лингвистических переменных. Затем производится формирование базы правил при помощи ГА однокритериальной оптимизации. На втором шаге под каждую сформированную базу правил производится оптимизация лингвистических переменных также с помощью ГА. Оценка БП и ЛП производится с помощью процедуры нечеткого коллективного вывода. БП или ЛП устанавливаются в систему GFRBS и проверяются на тестовой выборке.

Рассматриваются две схемы. В первой схеме, как представлено на рисунке 5, этап оптимизации ЛП производился после оптимизации БП.



Рисунок 5. Пример схемы, в которой сначала производится оптимизация БП, а затем ЛП (КЭ – критерий эффективности)

Во второй схеме построения нечеткой системы этап оптимизации БП выполняется после оптимизации ЛП.

Оценивание эффективности индивида производится с помощью ρ_c – критерия качества формируемых БП и ЛП.

На рисунке 6 представлено кодирование хромосомы ГА (БП) бинарной оптимизации, в которой каждое правило состоит из предпосылки и выхода правила. Первая строка указывает возможные варианты связки нечетких переменных, которая происходит с помощью операторов «И» или «ИЛИ», а

вторая строка показывает бинарное представление, где значение 0 – оператор «И», а 1 – «ИЛИ». Данная связка применяется к каждому правилу.

Также первая строка указывает количество термов для каждой ЛП, на каждый из которых отводится по одному биту, соответственно во второй строке 0 означает исключение терма из правила, а 1 – включение.

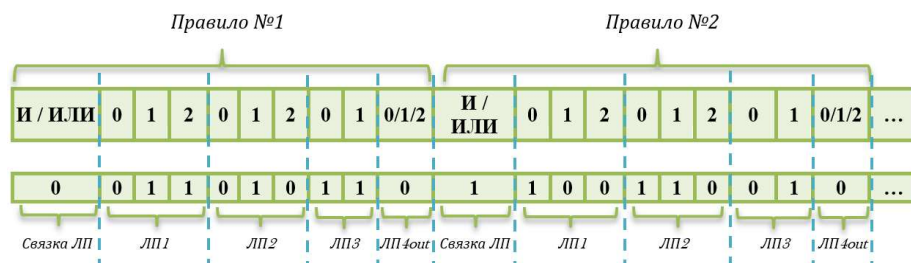


Рисунок 6. Структура кодирования хромосомы, которая представляет собой базу правил, для ГА однокритериальной оптимизации

ЛП входов *Distance* (ЛП1) и *Error* (ЛП2) описываются тремя термами, а *Weight_agent* (ЛП3) имеет два терма. На ЛП выхода *Confidence* (ЛП4out), которая имеет три терма, приходится два бита, то есть эта переменная может принимать следующие значения: 00 – малое, 01 – среднее, 10 – среднее, 11 – большое.

Количество бит, необходимое для представления одного правила равно 11. Соответственно длина хромосомы (индивида), представляющего собой базу правил, имеет размерность ($m \times 11$) бит, где m - количество правил.

При вычислении функции пригодности в ГА используются элементы условной оптимизации, а именно применение «смертельного» штрафа к индивиду, при котором недопустимые индивиды отбрасываются (т.е. пригодность данного индивида приравнивается нулю).

Штраф назначается индивиду (то есть всей базе правил), так как при случайном генерировании и в ходе эволюции могут получаться БП, которые невозможно использовать. Например, когда для некоторых входов (или для всех) в результате нечеткого вывода получается пустое нечеткое множество, к которому применить процедуру дефазификации невозможно.

Кодирование хромосомы дифференциальной эволюции для оптимизации ЛП (задача вещественной оптимизации) представлено на рисунке 7. В хромосоме содержатся параметры треугольной функции принадлежности (a , b , c) для каждого терма ЛП. Первая строка указывает номер терма для каждой ЛП, а вторая строка показывает вещественное представление параметров терма.

Таким образом последовательное применение генетического алгоритма для оптимизации БП и дифференциальной эволюции для оптимизации ЛП позволяет настраивать нечеткую систему коллективного вывода в целом.

Исследование эффективности прямой (сначала БП, затем ЛП) и обратной схем формирования коллективного вывода с различной последовательностью этапов проектирования нечеткой системы проводилось на основе известных наборов тестовых регрессионных задач.

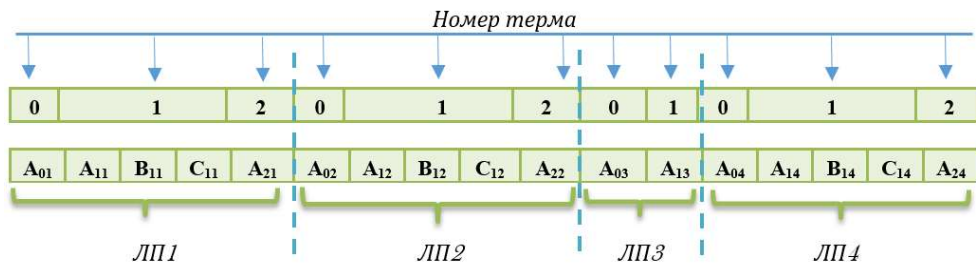


Рисунок 7. Структура кодирования хромосомы, которая представляет собой лингвистические переменные для дифференциальной эволюции

В результате исследования было показано, что порядок последовательного применения операторов оптимизации БП и ЛП существенного значения не имеет. Однако распределение ресурсов при структурной оптимизации базы правил и параметрической оптимизации ЛП имеет значение. На двух задачах из трех победили схемы, в которых больше вычислительных ресурсов выделялось на оптимизацию ЛП. Такой эффект в первую очередь объясняется различиями в структуре и размерности поисковых пространств в указанных задачах. На третьей задаче не выявлено существенного различия в схемах настройки нечеткой системы коллективного вывода. Это может быть связано с тем, что данные по этой задаче были синтетическими и природа зависимости между входами и выходами не является сложной.

Полученные результаты численных экспериментов показывают, что предложенный подход к построению системы на нечеткой логике для формирования ансамбля эффективен при решении задач классификации и регрессии. Решенные практические задачи подтверждают корректность и реализуемость предложенного метода.

Четвертая глава посвящена апробации предложенных подходов на практических задачах. Рассматриваются задачи распознавания лиц по изображению, задача прогнозирования эмоционального состояния человека по аудиоданным и задача моделирования технологического процесса металлургического производства (восстановление криолитового соотношения).

Алгоритм коллективного принятия решения сравнивается со специализированными методами решения рассматриваемых задач. В задачах распознавания лиц по изображению одним из самых эффективных подходов является алгоритм Eigenfaces, который показывает высокую точность в распознавании лиц, он точнее, чем алгоритмы машинного обучения общего назначения, такие как нейронные сети, машина опорных векторов и т.д. В данной работе рассматривалось две задачи распознавания лиц с исходными векторами признаков размерностью 10304 и 36000 соответственно. Изображения лиц для каждого спикера получены с различной освещенностью, положением головы, открытым и закрытым ртом, с очками и без, с разным выражением лица

На всех исследуемых задачах распознавания лиц предложенный в данной работе коллективный подход показывает результаты лучше, чем Eigenfaces. При этом Объединение агентов в коллектив позволяет получить решение лучше, чем у лучшего из агентов.

Основной сложностью в задаче прогнозирования эмоционального состояния человека по аудиоданным является то, что результирующие величины вычислены на основе экспертных оценок, которые не всегда согласуются между собой. Специализированного алгоритма для прогнозирования эмоционального состояния человека по аудиоданным в настоящий момент нет. Поэтому результаты исследования коллективного вывода сравнивались с алгоритмами машинного обучения общего назначения.

Эмоции человека выражены в виде комбинации двух показателей: «Valence» – направленность эмоции (отрицательные или положительные эмоции) и «Arousal» – выраженность эмоции (степень возбужденности). Эти показатели представлены вещественными числами.

База данных представляет собой аудиозаписи, каждая из которых длится 5 минут. На основе этого необходимо решить задачу прогнозирования аффективного (эмоционального) поведения. Для извлечения акустических характеристик использовалась система OpenSMILE с набором функций ComParE, которая извлекает 130 акустических признаков.

В данной работе временной интервал для формирования признакового пространства был выбран 1.5 секунды, что обеспечило размерность исходного вектора признаков равную 5200.

Показано, что применение предложенного подхода позволяет получать решение точнее лучшего из агентов. При этом, такие процедуры коллективного вывода как *mean* и *Wmean* позволяют получить результат на уровне среднего агента.

Задача восстановления криолитового соотношения является промышленной задачей и связана с влиянием большого количества неконтролируемых и неизмеряемых факторов. Объясняющие факторы не всегда полностью представляют результирующую переменную и не всегда измеряются достаточно точно. При этом, для прогнозирования криолитового соотношения специалистами металлургической отрасли была разработана специализированная модель, в которой учитываются технологические и химические зависимости между объясняющими факторами и результирующей переменной. В диссертационной работе для прогнозирования криолитового соотношения использовалось признаковое пространство с 9 признаками. Показано, что применение коллективного вывода с учетом данных модели производства позволяет существенно повысить точность прогноза, в том числе и в сравнении с самой моделью производства.

Таким образом, предложенный подход был успешно апробирован на ряде практических задач. Применение предложенного алгоритма коллективного вывода позволяет использовать как методы машинного обучения общего

назначения, так и специализированные модели и алгоритмы, и повышать точность вычислений в различных предметных областях.

В заключении сформулированы основные выводы и результаты, полученные в диссертационной работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Проведен обзор современных методов анализа данных и форм их коллективного взаимодействия.
2. Разработан и реализован коллективный алгоритм анализа данных на основе нечеткой логики для решения задач классификации и регрессии;
3. Разработана и реализована процедура выбора алгоритмов для включения в состав коллектива.
4. Разработана и реализована процедура автоматизированного выбора показательных примеров в опорное множество для формирования коллективного вывода.
5. Разработана программная система, реализующая предложенный алгоритм для формирования коллектива и подходы к проектированию различных его этапов.
6. Исследована работоспособность предложенного подхода на тестовых и практических задачах.

Таким образом, в данном диссертационном исследовании поставлена и решена задача повышения эффективности разработки и применения интеллектуальных технологий анализа за счет автоматизированного формирования коллективов алгоритмов с помощью специальных систем на нечеткой логике, что имеет существенное значение для теории и практики системного анализа и обработки информации.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК РФ:

1. Полякова А.С., Липинский Л.В., Семенкин Е.С. Эволюционный алгоритм автоматизированного формирования базы правил в процедуре нечеткого вывода при коллективном принятии решений // Системы управления и информационные технологии, 2019, №2(76). С. 29-36.
2. Полякова А.С. Формирование коллектива решающих правил многокритериальным эволюционным алгоритмом в задаче анализа эмоций человека по аудиоданным / А.С. Полякова, Л.В. Липинский // Вестник МГТУ им. Н.Э. Баумана. Серия «Приборостроение». 2018. Т.18. №4. С. 744-747.
3. Полякова А.С., Сидоров М. Ю., Семенкин Е. С. Комбинирование подходов кластеризации и классификации для задачи распознавания эмоций по речи // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. 2016. Т. 17. № 2(59). С. 335-342.

Публикации в изданиях, индексируемых в международных базах:

4. Polyakova A.S., Lipinskiy L.V., Semenkin E.S. Investigation of Reference Sample Reduction Methods for Ensemble Output with Fuzzy Logic-Based Systems // 8th International Congress on Advanced Applied Informatics "7th International Conference on Smart Computing and Artificial Intelligence" (SCAI 2019), Toyama, Japan, 2019 (Web of Science, Scopus).

5. Polyakova A. S., Lipinskiy L. V., Semenkin E. S. Investigation of resource allocation efficiency in optimization of fuzzy control system // IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2019. Т. 537. №. 5. С. 052036 (Web of Science, Scopus).

6. Polyakova A., Lipinskiy L. A study of fuzzy logic ensemble system performance on face recognition problem // IOP Conference Series: Materials Science and Engineering. – IOP Publishing, 2017. Т. 173. №. 1. С. 012013 (Web of Science, Scopus).

Публикации в сборниках трудов конференций:

7. Васьковская О. А., Полякова А. С. Исследование эффективности алгоритмов классификации для задач оценки качества диалогового взаимодействия типа «Человек-человек» // Материалы XX междуна. науч. конф. «Решетневские чтения», (Красноярск, 09-12 ноября 2016 г.). Т. 2, С. 22-23.

8. Полякова А. С. О применении библиотеки OpenCV в задаче распознавания лиц по их изображению // Материалы II междуна. науч.-практич. конф. «Актуальные проблемы авиации и космонавтики», (Красноярск, 11-15 апреля 2016 г.). Т. 1, С. 558-559.

9. Полякова А. С., Круглова У. Н. Краткосрочное прогнозирование цены на электроэнергию с помощью методов интеллектуального анализа данных // Материалы междуна. конф. «Перспектив Свободный», (Красноярск, 15-25 апреля 2015 г.). С. 49-51.

10. Полякова А. С., Сидоров М. Ю. Роль кластеризации при решении задачи распознавания эмоций по акустическим характеристикам // Материалы XIX междуна. науч. конф. «Решетневские чтения», (Красноярск, 10-14 ноября 2015 г.). Т. 2, С. 83-84.

11. Polyakova A. S., Kruglova U. N. Preliminary processing of data for the problem of short-term forecasting of electric energy cost // Материалы XIV междуна. науч. конф. «Молодежь. Общество. Современная наука, техника и инновации», (Красноярск, 14 мая 2015 г.). 2015. №. 14. С. 247-249.

12. Полякова А. С. Настройка лингвистических переменных с помощью генетического алгоритма // Информационно-телекоммуникационные системы и технологии. Материалы Всероссийской научно-практической конференции (ИТСиТ-2014). 2014. С. 413-414.

13. Полякова А. С. Формирование нечетких систем с помощью генетического алгоритма // Материалы XVIII междуна. науч. конф. «Решетневские чтения», (Красноярск, 11-14 ноября 2014 г.). Т. 2, С. 108-110.

14. Полякова А. С., Семенкин Е. С. Исследование эффективности методов NSGA-II, SPEA и SPEA2 в задачах многокритериальной оптимизации // Приоритетные направления развития науки и технологий. Тезисы докладов XIII Всероссийской научно-технической конференции, Тула. 2013. С.76-77.

Зарегистрированные программные системы

15. Полякова А.С., Липинский Л.В., Семенкин Е.С. Программная система формирования коллективного решения с помощью нечеткого логического вывода. Свидетельство о государственной регистрации программы для ЭВМ №2017662418 от 07.11.2017.

Полякова Анастасия Сергеевна
Коллективные методы интеллектуального анализа данных
на основе нечеткой логики
Автореферат
Подписано к печати 20.09.2019. Формат 60x84/16
Уч. изд. л. 1.0 Тираж 100 экз. Заказ № _____
Отпечатано в отделе копировальной и множительной техники
СибГУ им. М.Ф. Решетнева.
660037, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31