

На правах рукописи



МОИСЕЕВА Татьяна Александровна

**МЕТОДЫ ГЕНЕРАЦИИ БАЗ ЗНАНИЙ
НЕЧЕТКИХ ПРОДУКЦИОННЫХ СИСТЕМ
С ИСПОЛЬЗОВАНИЕМ ПРОЦЕДУР КЛАСТЕРИЗАЦИИ**

Специальность 1.2.1. Искусственный интеллект и машинное обучение

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Воронеж – 2025

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Воронежский государственный университет»

Научный руководитель доктор технических наук, профессор **Леденёва Татьяна Михайловна**,

Официальные оппоненты **Катасёв Алексей Сергеевич**,
доктор технических наук, профессор; федеральное государственное бюджетное образовательное учреждение высшего образования «Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ», кафедра систем информационной безопасности, профессор

Ходашинский Илья Александрович,
доктор технических наук, профессор; федеральное государственное автономное образовательное учреждение высшего образования «Томский государственный университет систем управления и радиоэлектроники», кафедра компьютерных систем в управлении и проектировании, профессор

Ведущая организация Федеральное государственное бюджетное образовательное учреждение высшего образования «Юго-Западный государственный университет»

Защита состоится 26 сентября 2025 года в 16 часов на заседании диссертационного совета 24.2.288.11, созданного на базе федерального государственного бюджетного образовательного учреждения высшего образования «Воронежский государственный университет», по адресу: 394018, г. Воронеж, Университетская площадь, 1, аудитория 226.

С диссертацией можно ознакомиться в научной библиотеке Воронежского государственного университета и на сайте <http://www.science.vsu.ru>.

Автореферат разослан «_____» 2025г.

Ученый секретарь
диссертационного совета
24.2.288.11
кандидат физико-математических
наук, доцент

Медведева
Ольга
Александровна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Одним из важнейших направлений искусственного интеллекта является создание систем, основанных на знаниях. К ним, в частности, относятся нечеткие производственные системы (НПС), являющиеся обобщением обычных производственных систем. Приложения НПС ориентированы на решение задач управления (нечеткие системы управления, нечеткие регуляторы встроены в огромное количество промышленных изделий), прогнозирования, диагностики, принятия решений в условиях неопределенности. Ядром НПС, «отвечающим» за сферу применения, является база знаний, включающая базу *если-то-правил*, основанную на производственной модели представления знаний. Формирование базы знаний – важнейший этап проектирования, при этом различают два основных подхода к интеграции знаний и данных в НПС. В первом случае для формирования базы знаний привлекаются эксперты, что порождает субъективность и неоднозначность в формулировках правил, невозможность обеспечить полноту знаний о реальной системе. С другой стороны, в настоящее время наблюдается рост объемов информации, описывающей прецеденты в той или иной предметной области, а также полученной в результате наблюдения или экспериментов. Алгоритмы автоматической генерации баз знаний на основе обучающих данных включают несколько групп методов, среди которых кластерный подход признан перспективным для некоторых типов НПС (*evolving fuzzy systems*). Выделяя группировки данных, он позволяет повысить уровень интерпретируемости базы знаний – свойства, которое лежит в основе объяснительной способности интеллектуальных систем. Для кластеризации обучающих данных целесообразно использовать метрические алгоритмы, к преимуществам которых относятся геометрическая интерпретация, возможность выделения нетипичных объектов, возможность осуществления квантизации пространства признаков, сравнительная простота реализации. Повышение качества обработки данных в процессе кластеризации позволит, в свою очередь, обеспечить качество базы знаний НПС на начальной стадии ее формирования и определить дальнейшую процедуру работы с ней (оптимизация, редукция, интерполяция правил и др.). Актуальность исследования обусловлена необходимостью совершенствования методов и алгоритмов обработки данных, обеспечивающих генерацию баз знаний НПС с улучшенными свойствами на основе процедур кластеризации при наличии обучающей выборки – важной научной задачи, имеющей значение для инженерии знаний и разработки систем искусственного интеллекта.

Степень разработанности темы исследования. Исследованиям в области формирования и оптимизации баз знаний НПС посвящены работы J. Dickerson, B. Kosko, J.C. Bezdek, P. P. Angelov, A. Lemos, H. Genther, E. D. Lughofer, а также отечественных ученых А.С. Катасёва, Д.В. Катасёвой, И.А. Ходашинского, А.Л. Тулупьева, А.А. Сорокина, М.А. Сергиенко и др. С целью совершенствования алгоритмов нечеткой кластеризации I. Kramosil и J. Michálek ввели понятие нечеткой метрики, которое получило развитие в работах A. George, P. Veeramani, V. Gregori, O. Grigorenko, N. M. Ralević. В связи с появлением обобщенных представлений¹ треугольных и конорм в классе рациональных функций появилась возможность построения конкретных

¹ Ledeneva, T.M. Additive generators of fuzzy operations in the form of a Linear Fractional Function / T.M. Ledeneva // Fuzzy sets and systems, 2020. – № 386. – Pp. 1-24.

Ledeneva, T.M. New Family of Triangular Norms for Decreasing Generators in the Form of a Logarithm of a Linear Fractional Function / T.M. Ledeneva // Fuzzy sets and systems, 2022. – № 427. – Pp. 37-54.

представлений нечетких метрик, которые позволяют улучшить качество кластеризации – важного этапа построения базы знаний НПС.

Цель исследования заключается в совершенствовании методов генерации баз знаний нечетких производственных систем на основе обучающих данных с использованием кластерных процедур.

Для достижения цели необходимо решить следующие задачи:

1. Проанализировать подходы к генерации баз знаний НПС и выявить возможности метрических алгоритмов кластеризации для решения данной проблемы.
2. Разработать процедуры для формирования баз знаний на основе обучающих данных с использованием эллипсоидальной кластеризации.
3. Предложить новые варианты нечетких метрик и исследовать их свойства.
4. Разработать и протестировать программное обеспечение для проведения вычислительного эксперимента и апробации предложенных подходов.

Объект исследования – база знаний в форме совокупности производственных правил. **Предмет исследования** – процедура генерации баз знаний НПС с использованием алгоритмов кластеризации.

Методы исследования базируются на принципах инженерии знаний. Теоретические результаты получены с использование методов оптимизации, теории нечетких множеств и методов нечеткого моделирования, методов кластерного анализа. Для программной реализации использовались современные технологии объектно-ориентированного программирования.

Научная новизна. В диссертации представлены следующие результаты, характеризующиеся научной новизной:

- метод формирования баз нечетких производственных правил на основе эллипсоидальной кластеризации, отличающийся использованием эллипсоидов минимального объема и позволяющий повысить качество аппроксимации в сравнении с известным подходом, основанным на использовании матриц ковариаций кластеров;
- ограничения на параметры непрерывных архimedовых треугольных норм из класса рациональных функций, обеспечивающие свойство строгости;
- семейство нечетких метрик, впервые полученных на основе аддитивных генераторов непрерывных архimedовых строгих треугольных норм из класса рациональных функций, отличающихся набором настраиваемых параметров, что позволяет учитывать структуру данных при использовании метрических алгоритмов кластеризации;
- модель комплексной оценки качества кластеризации, основанная на использовании функций порядкового взвешенного агрегирования и позволяющая учитывать «нечеткое большинство» значимых значений критериев качества кластеризации;
- алгоритм формирования базы знаний нечеткого классификатора энцефалограмм для асинхронного интерфейса «мозг-компьютер», отличающийся учетом наиболее значимых предикторов и позволяющий распознавать реальные и мысленные движения верхних конечностей;
- структура и программная реализация приложения для формирования баз правил нечетких производственных систем на основе метрических алгоритмов кластеризации с возможностью выбора нечетких метрик и критериев качества кластеризации, что обеспечит учет особенностей информационной среды конкретной прикладной задачи.

Соответствие Паспорту специальности. Полученные в диссертации научные результаты соответствуют следующим пунктам Паспорта специальности 1.2.1 «Искусственный интеллект и машинное обучение: п. 5 «Методы и технологии поиска, приобретения и использования знаний и закономерностей, в том числе – эмпирических, в системах искусственного интеллекта ...», п. 15 «Математические исследования в области статистики, логики, алгебры, топологии, анализа функций ...».

Теоретическая и практическая значимость. Метод формирования баз знаний НПС развивает технологии выявления закономерностей и знаний в системах искусственного интеллекта. Нечеткие метрики, построенные с использованием генераторов непрерывных строгих архimedовых треугольных норм из класса рациональных функций, не только расширяют возможности построения нечетких метрических пространств, но и значительно дополняют инструментарий метрических алгоритмов кластеризации, демонстрируя в рамках исследования превосходство по многим показателям качества кластеризации перед традиционными метриками.

Практическая значимость работы заключается в программной реализации предложенного подхода к решению важной для НПС проблемы формирования базы знаний на основе обучающего множества, который в сравнении с часто используемым экспериментальным методом повышает качество и обоснованность решений в интеллектуальных информационных системах, в которых НПС является ядром.

Теоретические результаты диссертационной работы используются в учебном процессе Воронежского государственного университета; программный комплекс применяется в финансовой компании «ООО ФПК «Альфа», а также для проведения исследований, связанных с разработкой интерфейсов «мозг-компьютер», в Лаборатории медицинской кибернетики Воронежского государственного университета.

Результаты и положения, выносимые на защиту:

1. Предложенный метод формирования *если-то*-правил на основе аппроксимации кластеров эллипсоидами минимального объема создает основу для генерации базы знаний с учетом обучающих данных, что повышает уровень объяснительной способности интеллектуальных информационных систем, базирующихся на НПС.

2. На основе аддитивных генераторов строгих непрерывных архimedовых треугольных норм, представимых рациональными функциями, получены нечеткие метрики, которые, с одной стороны, имеют теоретическое значение для построения нечетких метрических пространств, а с другой – обладают практической ценностью для метрических алгоритмов нечеткой кластеризации, проявляя адаптивные свойства за счет настройки параметров.

3. Разработанный многофункциональный программный комплекс, структура которого включает библиотеки метрик и критериев качества кластеризации, программную реализацию некоторых известных алгоритмов кластеризации и всех предложенных алгоритмов и процедур, позволил осуществить объемный вычислительный эксперимент, а также продемонстрировал возможность использования предложенного подхода для построения нечеткого классификатора в системе интерфейса «мозг-компьютер» с целью анализа электроэнцефалограмм.

Степень достоверности и апробация результатов. Достоверность результатов исследования основана на корректном использовании математического аппарата, обосновании выбора алгоритмических решений и их согласованностью с результатами вычислительного эксперимента. Результаты диссертации докладывались и обсуждались на следующих научных конференциях: международная научная конференция «Актуальные проблемы прикладной математики, информатики и механики»

(Воронеж, 2021-2023 гг.), International Conferences on Control Systems, международная научная конференция Mathematical Modeling, Automation and Energy Efficiency (Lipetsk, 2022-2024 гг.), межвузовская научная конференция молодых ученых и студентов «Математика, информационные технологии, приложения» (Воронеж, 2024 г.).

Публикации. По результатам исследования опубликовано 15 научных работ (4 без соавторов), в том числе 5 статей в журналах из Перечня ВАК, 4 статьи – в изданиях, индексируемых в Scopus, 2 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. В работах, опубликованных в соавторстве, лично автору принадлежат следующие результаты: [1, 8] – теоретическое и экспериментальное исследование нечетких метрик; [2, 4, 6, 9] – разработка алгоритмической и программной реализации метода автоматизированного формирования базы знаний; [5, 7, 13] – сравнительный анализ алгоритмов.

Объем работы. Диссертация состоит из введения, четырех глав, заключения, списка использованных источников, включающего 190 наименований, и двух приложений. Объем диссертации составляет 175 страниц, диссертация содержит 55 рисунков, 22 таблицы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы исследования, приводится общая характеристика полученных результатов.

В **первой главе** представлены общие сведения о НПС, приведена классификация подходов к генерации баз знаний – основного компонента, который отвечает за приложения. В рамках каждого подхода проведен обзор и анализ соответствующих алгоритмов. В качестве базы исследования выбран кластерный подход. Поставлена цель исследования и определены задачи, которые необходимо решить для ее достижения.

Вторая глава посвящена построению базы знаний НПС на основе эллипсоидальной кластеризации. Предобработка данных направлена на исключение выбросов и нормализацию данных. Идея метода представлена на рис. 1. Разбиение данных на кластеры и построение аппроксимирующих их эллипсоидов позволяет сгенерировать базу знаний с минимальным количеством правил, «покрывающих» данные.

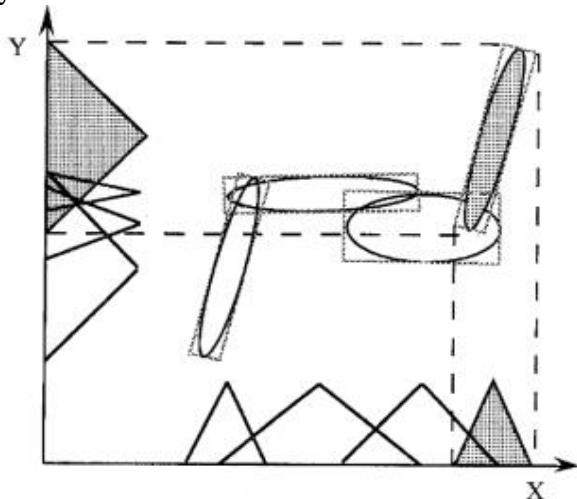


Рис. 1. Формирование производственных правил на основе эллипсоидальной кластеризации

Выбор эллипсоидов для аппроксимации кластеров обусловлен рядом их свойств, перечисленных в диссертации, основным из которых является то, что *существует*

Для построения функций принадлежности термов лингвистических шкал входных и/или выходной переменных используется либо проецирование матрицы разбиения на оси переменных, что позволяет получить точечно-определенные нечеткие множества, которые затем аппроксимируются подходящей функцией, либо эллипсоид вписывается в параллелепипед, который проецируется на оси переменных, тогда функции принадлежности задаются аналитически с использованием длины проекции и координат центра кластера.

эллипсоид минимального объема, «покрывающий» заданное множество точек в \mathbb{R}^n . Данный факт позволяет предположить возможность улучшения качества базы знаний с точки зрения существующих требований. Кроме того, для суммы эллипсоидов существует эллипсоид минимального объема, что дает основу для объединения термов лингвистических шкал переменных при необходимости.

В диссертации показано, что задача нахождения эллипсоида минимального объема, включающего все заданные точки $x_i \in \mathbb{R}^n$ ($i = \overline{1, k}$), путем преобразования исходной постановки сводится к следующей задаче выпуклого программирования с линейными ограничениями (объем эллипсоида пропорционален $1 / \sqrt{\det M}$):

$$\begin{cases} -\ln(\det M) \rightarrow \min \\ M > 0, 1 - (Mq_i, q_i) \geq 0 \quad (i = \overline{1, k}), \end{cases} \quad (1)$$

где M – матрица, определяющая эллипсоид с центром в нуле; $q_i = (x_i 1)^T$.

Для нахождения решения задачи (1) в диссертации строится функция Лагранжа, которая с использованием производной от матричной функции преобразуется к виду, позволяющему сформировать систему ограничений Куна-Таккера в матричном виде. Двойственная задача (λ – вектор множителей Лагранжа) после замены переменных $u = \lambda / n + 1, U = \text{diag}(u)$ имеет следующий вид (здесь $Q = (q_1, \dots, q_k)$):

$$\begin{cases} \ln \det(QUQ^T) \rightarrow \max \\ \mathbf{1}^T u = 1, u \geq 0. \end{cases} \quad (2)$$

Для решения задачи (2) в диссертации используется алгоритм Хачияна².

Пусть X – матрица, в столбцах которой стоят векторы наблюдаемых данных, u – решение задачи (2), тогда параметры минимального покрывающего эллипсоида вычисляются по формулам $c = Xu, P = \frac{1}{n} (XUX^T - Xu(Xu)^T)^{-1}$, где P – симметрическая положительно определенная матрица, определяющая эллипсоид с центром в c .

Если эллипсоиды получены, то проецируя матрицу разбиений или описывающие их параллелепипеды на оси (в диссертации приведены формулы для вычисления проекций), каждому из них можно поставить в соответствие продукционное правило. Заметим, что база знаний, полученная в результате процедуры автоматической генерации, как правило, нуждается в корректировке, которая направлена на решение задач структурной и параметрической оптимизации. Среди правил могут оказаться противоречивые и избыточные правила, следовательно, необходим дополнительный анализ посылок и заключений правил с точки зрения корректности лингвистических шкал. Если два правила имеют одинаковые посылки, но разные заключения, то этот факт свидетельствует о наличии конфликтов в базе правил. Избыточность проявляется в наличии схожих термов, которые значительно перекрывают друг друга. Скорректированную базу правил можно считать допустимой, если по сравнению с исходной не произошла значительная потеря точности.

В рамках исследования была проведена серия вычислительных экспериментов, посвященная сравнению предложенного подхода на основе минимальных

² Khachiyan, L. G. Rounding of polytopes in the real number model of computation // Mathematics of Operations Research, 1996. – Vol. 21. – No. 2. – P. 307-320.

эллипсоидов с алгоритмом, в котором используются матрицы ковариаций кластеров (матрица, обратная к матрице ковариаций, определяет эллипсоид (эллипс) в пространстве вход-выход). Выборка, подлежащая кластеризации, включала 2000 точек (на плоскости) и формировалась на основе стандартных тестовых функций (как правило, многоэкстремальных, со сложным ландшафтом, всего – 6) путем добавления шума (пример на рис. 2). На рис. 3 представлены минимальные эллипсы, построенные с помощью алгоритма Хачияна, а также лингвистическая шкала входной переменной.

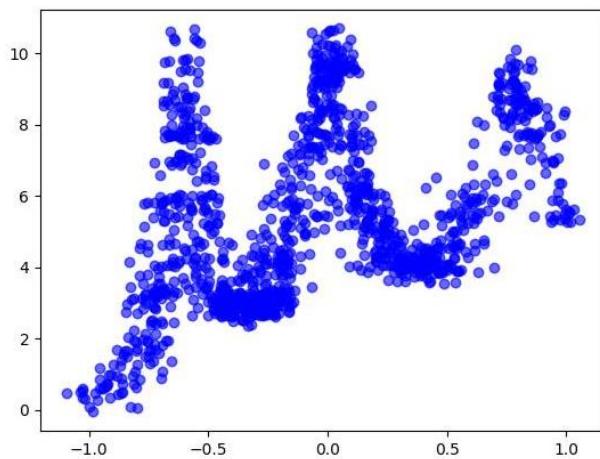


Рис 2. Выборка, построенная на основе одной из тестовых функций

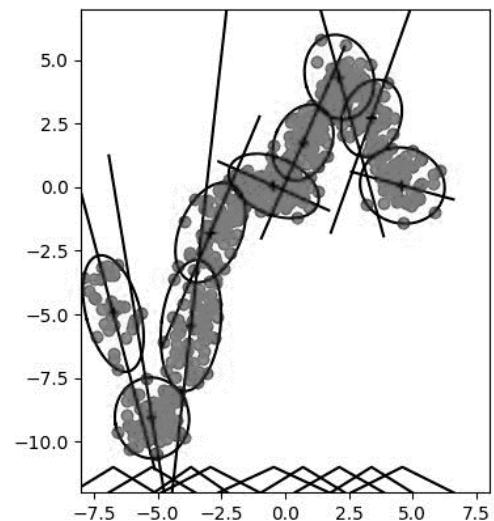


Рис. 3. Минимальные покрывающие эллипсы и термы лингвистической шкалы

В табл. 1 для каждой из тестовых функций при разбиении на кластеры приведено значение корня из среднеквадратичной ошибки для сформированной на основе кластеризации базы правил.

Таблица 1 – Значения корня из среднеквадратичной ошибки

Тестовая функция	5 кластеров		9 кластеров		17 кластеров	
	Эллипсы на основе матриц ковариаций кластеров	Эллипсы минимальной площади	Эллипсы на основе матриц ковариаций кластеров	Эллипсы минимальной площади	Эллипсы на основе матриц ковариаций кластеров	Эллипсы минимальной площади
f_1	13.1369	12.5552	11.8493	11.2804	12.8720	12.0148
f_2	3.9691	3.8884	3.3668	3.5584	3.3028	3.0823
f_3	2.7250	2.6267	2.3193	2.3067	2.2404	2.1668
f_4	5.2741	4.4178	1.8016	1.6961	1.7271	1.5669
f_5	3.7803	2.4458	2.9305	2.3775	2.5232	1.8707
f_6	15.7696	15.4959	15.2297	15.0289	15.2343	14.9577
Среднее значение ошибки	7.4420	6.9050	6.2495	6.0413	6.3166	5.9432

Анализ результатов проведенных экспериментов позволил сделать выводы, среди которых основными являются следующие:

1. При увеличении количества кластеров использование эллипсов минимальной площади в основном дает небольшое улучшение точности при генерации правил типа Такаги-Сугено, но при малом количестве кластеров выигрыш значительный.

2. Так как каждое правило связано с определенным кластером и имеется стратегия на уменьшение количества правил, то необходим обоснованный выбор процедуры кластеризации для повышения уровня интерпретируемости правил.

3. Если у полученных эллипсов большая ось по длине значительно превосходит меньшую, то целесообразно использовать правила типа Такаги-Сугено, иначе более подходящей является лингвистическая модель – этот выбор обеспечивает улучшение качества аппроксимации.

4. Оптимизация базы правил, полученной на основе эллипсоидальной кластеризации, может быть реализована в следующих направлениях: а) если получены эллипсы минимальной площади, то можно сократить количество правил за счет сложения соответствующих эллипсов; б) проекции эллипсов необходимо аппроксимировать функциями принадлежности подходящих нечетких чисел; в) если проецирование эллипсов на ось абсцисс приводит к сильно перекрывающимся термам, то необходим тщательный анализ термов с целью, например, их объединения.

В третьей главе рассмотрены основные понятия, связанные с нечеткими метрическими пространствами, в определение которых входит треугольная норма T , моделирующая пересечение нечетких множеств и конъюнкцию.

По определению, *треугольная норма* $T : [0,1] \times [0,1] \rightarrow [0,1]$ – это неубывающая, коммутативная и ассоциативная бинарная операция, которая для любого $x \in [0,1]$ удовлетворяет ограничению $T(x,1) = T(1,x) = x$. Если для любого $x \in [0,1]$ имеет место свойство $T(x,x) < x$, то треугольная норма называется *архimedовой*.

Двойственная относительно стандартного отрицания бинарная операция $S(x,y) = 1 - T(1-x,1-y)$, такая, что для любого $x \in [0,1]$ выполняется $S(x,0) = S(0,x) = x$, называется конормой (объединение, дизъюнкция).

К настоящему времени сформировались многочисленные семейства двойственных треугольных норм и конорм, однако в приложениях лишь немногие из них нашли применение (классические (\max, \min) и алгебраические ($xy, x + y - xy$)). Трудности при использовании параметрических операций обусловлены наличием параметров и необходимостью их настройки.

Важнейшей особенностью треугольных норм (и конорм) в силу их ассоциативности является представление в виде

$$T(x,y) = t^{(-1)}(t(x) + t(y)), \quad (3)$$

где функция $t : [0,1] \rightarrow [0,\infty)$, называемая аддитивным генератором, является непрерывной, строго убывающей со свойством $t(1) = 0$; $t^{(-1)}$ – псевдообратная функция такая, что для $x \in [0,t(0)]$ выполняется $t^{(-1)}(x) = t^{-1}(x)$, а при $x > t(0)$ $t^{(-1)}(x) = 0$. Если $t(0) = \infty$, то T – *строгая норма*, тогда для строгих треугольных норм имеем $t^{(-1)}(x) = t^{-1}(x)$.

В диссертации рассматриваются известные подходы к определению нечетких метрических и псевдометрических пространств. В общем случае тройка $fms = (X, M, T)$, где X – произвольное множество, T – непрерывная треугольная норма, M – нечеткое множество на $X \times X \times (0,\infty)$ с функцией принадлежности

$\mu_M : (0, \infty) \rightarrow [0, 1]$, удовлетворяющей определенным свойствам, называется *нечетким метрическим пространством*. Среди свойств функции принадлежности μ_M определяющим является неравенство, согласно которому

$$T(\mu_M(x, y, u), \mu_M(y, z, u)) \leq \mu_M(x, z, u).$$

Пусть U – произвольное множество, $d : U \times U \rightarrow [0, \infty)$ – обычная (четкая) функция расстояния, T – строгая архимедова треугольная норма с аддитивным генератором t , тогда существует *нечеткая метрика (нечеткая T -метрика)* в форме нечеткого множества M с функцией принадлежности³

$$\mu_M(x, y, u) = t^{(-1)} \left(\frac{d(x, y)}{\varphi(u)} \right). \quad (4)$$

Таким образом, выбрав d , φ и норму T с аддитивным генератором t , можно построить нечеткую метрику, в том числе для решения задачи кластеризации.

В диссертации рассматриваются непрерывные архимедовы треугольные нормы T из класса рациональных функций, представимые в виде многочлена или отношения двух многочленов, поскольку они представляют наиболее многочисленный класс. Известно, что в качестве их аддитивных генераторов могут выступать только функции, общий вид которых основан на дробно-линейной функции, а также логарифме и арктангенсе от нее с учетом аддитивных и мультипликативных констант. В диссертации получены ограничения на коэффициенты рациональных треугольных норм T , при выполнении которых они являются строгими. Для строгих треугольных норм на основе формулы (4) построены семейства нечетких метрик, при этом в качестве d рассматривались известные функции расстояния.

Для иллюстрации подхода рассмотрим треугольную норму

$$T_\rho(x, y) = xy / (\rho(x + y - xy) + 1 - \rho),$$

где $\rho < 1$, $\rho \neq 0$, которая является строгой для всех указанных значений ρ . Аддитивный генератор для T_ρ имеет вид $t_\rho(x) = -\frac{1}{1-\rho} \ln \frac{x}{\rho x + 1 - \rho}$, а псевдообратная функция определяется формулой

$$t_\rho^{(-1)}(x) = \left(e^{x(\rho-1)} (1-\rho) \right) / \left(1 - \rho e^{x(\rho-1)} \right), \quad x \geq 0.$$

Тогда, согласно (4), нечеткая метрика имеет вид

$$\mu_M(x, y, u) = \left(e^{\frac{d(x, y)}{\varphi(u)}(\rho-1)} (1-\rho) \right) / \left(1 - \rho e^{\frac{d(x, y)}{\varphi(u)}(\rho-1)} \right).$$

Заметим, что данная формула, по сути, задает множество нечетких метрик, конкретный вид которых определяется выбором параметров d , φ , ρ . Анализ свойств функции μ_M позволяет ее интерпретировать как функцию принадлежности нечеткого бинарного отношения сходства, тогда дополнение этого отношения есть несходство, а соответствующая функция принадлежности $(1 - \mu_M)$ задает некоторую функцию, которую целесообразно использовать для оценки расстояния

³ Grigorenko O.T., Miñana J.-J., Valero O. Two new methods to construct fuzzy metrics from metrics // Fuzzy Sets and Systems. 2023. – Vol. 467. – P. 108483.

$$r_\rho(d(x, y), u) = \left(e^{(1-\rho)\frac{d(x, y)}{u}} - 1 \right) / \left(e^{(1-\rho)\frac{d(x, y)}{u}} - \rho \right). \quad (5)$$

Визуализация семейства нечетких метрик позволила выделить несколько характерных поверхностей (рис. 4), как основы для рекомендаций при использовании в метрических алгоритмах кластеризации. Поверхность П1 соответствует функции, которую целесообразно использовать как индикатор при оценке сходства/несходства объектов. Поверхность П3 соответствует ситуации, когда оценка расстояния между объектами формируется с позиции пессимизма, а поверхность П4 – с позиции оптимизма. В случае П3 и П4 возможны значения параметров, когда возникают части поверхности, прилегающие к верхней грани. Это означает, что объекты с соответствующими оценками неразличимы. Наличие точки перегиба свидетельствует о смене позиции – с оптимизма на пессимизм или наоборот.

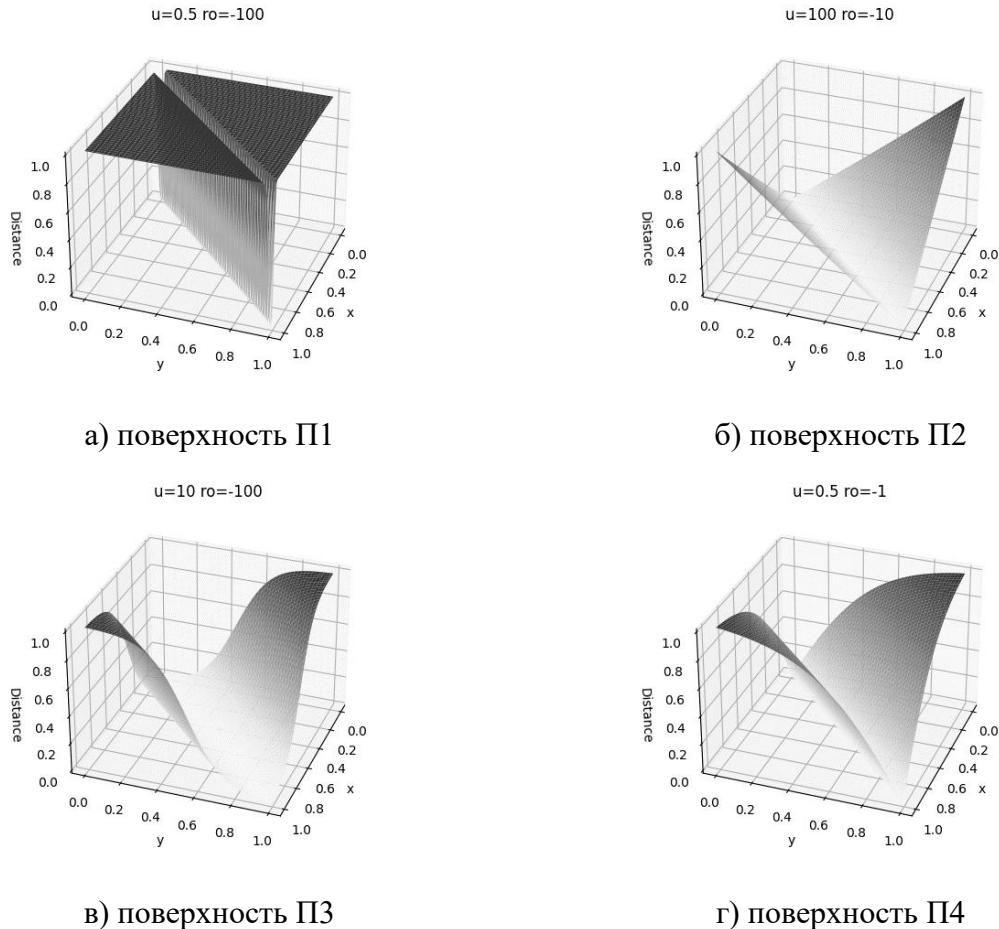


Рис. 4. Различные типы поверхностей нечеткой метрики

В диссертации приведены результаты серии вычислительных экспериментов, касающиеся свойств полученных нечетких метрик в приложении к задаче нечеткой кластеризации. В качестве базового метода был выбран алгоритм *нечетких k-медоид*, позволяющий использовать произвольную метрику для получения нечетких кластеров. Входная информация – это матрица расстояний между объектами. Помимо тестируемых метрик использовались известные функции расстояния Евклида (**E**), Минковского (**M**), Чебышева (**Ч**), а также манхэттенское расстояние (**Mx**) и метрика Канберра (**K**). Результат кластеризации (медоиды $C = \{c_i\}_{i=1, K}$, а также матрица нечеткого разбиения) оценивался с помощью часто используемых показателей качества

кластеризации. Для оценки влияния характера данных на качество кластеризации при выборе нечеткой метрики рассматривалось три типа данных: первый тип **I** – неперекрывающиеся кластеры с небольшой дисперсией (0.5); второй тип **II** – кластеры с большой дисперсией (1.0), которые могут частично перекрываться; третий тип **III** – кластеры с большой дисперсией (1.0), наложенные друг на друга. В качестве нечеткой метрики использовалась функция $r_\rho(d, u)$.

Для оценки качества кластеризации с использованием различных метрик рассчитывались следующие показатели качества кластеризации (их выбор обоснован в диссертации): коэффициент энтропии (**1**), индекс Кси-Бени с метрикой Евклида (**2**), индекс Кси-Бени с метриками, соответствующими использованным при кластеризации (**3**), коэффициент силуэта для четких кластеров с метриками, соответствующими использованным при кластеризации (**4**), коэффициент силуэта для четких кластеров с метрикой Евклида (**5**), коэффициент разбиения (**6**), эффективность разбиения с метрикой Евклида (**7**).

Для выявления влияния типа функции расстояния d в $r_\rho(d, u)$ на качество кластеризации генерировались различные наборы данных. В табл. 2 представлены значения критериев качества (было сгенерировано 7000 наборов данных, содержащих 5 кластеров, для которых строилось разбиение также на 5 кластеров; кластеры генерировались случайным образом, с различной дисперсией и степенью наложения).

Таблица 2. Оценки качества кластеризации (5 кластеров)

Метрика	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Евклида	0.3337	0.4577	0.4577	0.5610	0.5610	0.7371	0.1550
<i>f</i>	E	0.1218	0.4147	0.3292	0.7315	0.5667	0.8951
<i>u</i>	K	0.1431	0.9568	0.3221	0.6835	0.4733	0.8832
<i>z</i>	Mx	0.1296	0.4171	0.3324	0.7200	0.5622	0.8880
<i>z</i>	Ч	0.1256	0.4201	0.3309	0.7266	0.5627	0.8913
<i>y</i>	M	0.1247	0.4191	0.3313	0.7279	0.5637	0.8922
							0.1837

Подробный анализ результатов, приведенный в диссертации, показал, что по сравнению с кластеризацией на основе классического евклидова расстояния, кластеризация с нечеткой метрикой на основе того же евклидова расстояния в большинстве случаев дает лучшие значения для большинства используемых показателей качества кластеризации (табл. 2). Такой показатель качества, как коэффициент силуэта, вычисленный после приведения кластеров к четкому виду, для нечеткой метрики демонстрирует лучшие результаты для всех случаев. Также установлено, что лучшие значения показателей качества (1)-(7) зависят от функции d в $r_\rho(d, u)$, поэтому для выявления лучшей нечеткой метрики был предложен подход к построению комплексной (обобщенной) оценки по всему набору показателей на основе специальной операции агрегирования.

Пусть $x = (x_1, \dots, x_n) \in [0,1]^n$ – вектор частных оценок нечетких метрик по набору показателей качества кластеризации. Если важно учитывать значимость показателей, то целесообразно использовать типы сверток с весами, отражающими приоритет показателей. Если все показатели качества кластеризации одинаково важны, то рекомендуется использовать порядковый оператор взвешенного агрегирования. В отличие от предыдущего случая, здесь веса не связаны с показателями, они формализуют принцип «нечеткого большинства», который может быть положен в основу стратегии агрегирования частных оценок. Вектор весов определяется на основе

лингвистических кванторов, соответствующих принципу «нечеткого большинства» – таких как *большинство*, *как можно больше*, *по крайней мере половина* и т.д. Лингвистический квантор задается функцией квантификации Q . В диссертации рассматриваются различные варианты функции квантификации, порождающие соответствующие наборы весовых коэффициентов.

Пусть $w = (w_1, \dots, w_n) \in [0,1]^n$ – вектор весов, тогда n -мерный порядковый оператор взвешенного агрегирования $\Phi(x, w) = \sum_{i=1}^n w_i x_{\sigma(i)}$, где $\sigma: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ – перестановка, такая, что $x_{\sigma(1)} \geq \dots \geq x_{\sigma(n)}$, определяет комплексную оценку.

Поскольку было важно учитывать именно значения показателей качества кластеризации, то был реализован подход, согласно которому строго возрастающая и дифференцируемая функция квантификации Q_1 порождает вектор весов с компонентами $w_i = Q'_1(1 - x_{\sigma(i)}) / \sum_{j=1}^n Q'_1(1 - x_{\sigma(j)})$ ($i = \overline{1, n}$). В качестве альтернативного подхода использовались степенные функции квантификации $Q(x) = x^\alpha$. Если, например, $Q_2(x) = x^2$, то $w_i = (2i - 1)/n^2$ ($i = \overline{1, n}$). Установлено, что значение комплексной оценки зависит от выбранной функции квантификации. В табл. 3 представлены комплексные оценки нечетких метрик, полученные на основе данного подхода с использованием функций Q_1 и Q_2 (метрика с максимальным значением – лучшая).

Таблица 3. Комплексные оценки нечетких метрик по всему набору показателей качества кластеризации

Тип кластеров	Метрика		Комплексная оценка (1)	Комплексная оценка (2)
I	Евклида		0.0816	0.0
	Нечеткая	метрика Евклида	0.7423	0.2375
		метрика Канберра	0.6791	0.2131
		манхэттенское расстояние	0.7410	0.4736
		расстояние Чебышева	0.7544	0.0742
	метрика Минковского		0.7485	0.0873
II	Евклида		0.0816	0.0
	Нечеткая	метрика Евклида	0.7482	0.3520
		метрика Канберра	0.4024	0.2606
		манхэттенское расстояние	0.6515	0.5879
		расстояние Чебышева	0.6951	0.1912
	метрика Минковского		0.7019	0.1773
III	Евклида		0.0136	0.0351
	Нечеткая	метрика Евклида	0.9704	0.8885
		метрика Канберра	0.4329	0.3755
		манхэттенское расстояние	0.8575	0.8017
		расстояние Чебышева	0.8112	0.7813
	метрика Минковского		0.8529	0.8340

Анализ результатов вычислительного эксперимента позволил сформулировать ряд выводов, которые могут быть положены в основу рекомендаций по выбору нечетких метрик. Перечислим наиболее важные:

- 1) в случае неперекрывающихся кластеров лучшими являются метрики, которые основаны на учете разности между одноименными компонентами векторов;
- 2) нечеткая метрика в большинстве случаев демонстрирует лучшие классифицирующие способности;
- 3) чем в большей степени перекрываются кластеры, тем более востребованной становится нечеткая метрика, основанная на евклидовой метрике (так, в соответствии с двумя вариантами расчета комплексной оценки для сильно перекрывающихся кластеров нечеткая метрика является лучшей).

В **четвертой** главе приведено описание программного комплекса (ПК) «FuzzyLogicCore» (рис. 5), предназначенного для выполнения автоматизированного формирования баз знаний, в котором реализованы следующие функции: построение эллипсоидов различными методами, построение термов лингвистических шкал, формирование заключения правил Такаги-Сугено с использованием осей эллипсов, реализация механизма нечеткого логического вывода на основе построенной базы знаний, интерполяция нечетких правил для классификатора, вспомогательные инструменты для обработки и визуализации данных. Для разработки ПК были использованы следующие технологии и программное обеспечение: язык разработки Java и среда IntelliJ IDEA, язык разработки для визуализации Python и среда PyCharm, фреймворк для серверной разработки SpringBoot, СУБД PostgreSQL.

Предложенные в диссертации подходы использовались для построения нечеткого классификатора в системе интерфейса «мозг-компьютер» (ИМК) с целью анализа электроэнцефалограмм (ЭЭГ). Данные интерфейсы широко используются в медицинских целях, например, для восстановления двигательных функций у людей с нарушениями движений или параличом. Одной из актуальных является задача извлечения из ЭЭГ сигналов, связанных с моторными образами, которые представляют собой паттерны мозговой активности, отражающие формирование мысленных команд на движение, как реальное, так и воображаемое. Задача распознавания моторных образов является задачей классификации сигналов ЭЭГ. Данные ЭЭГ представляют собой многомерные временные ряды, которые соответствуют электрическим потенциалам, измеренным на поверхности головы испытуемого с помощью электродов.

Для проведения эксперимента была создана группа из 30 человек. Запись ЭЭГ осуществлялась с частотой дискретизации 5000 Гц нетренированных испытуемых в двух чередующихся произвольным образом состояниях: состоянии покоя, а также при реальном и воображаемом поднятии обеих рук. Данные были поделены на тестовую и обучающую выборки, объем тестовой выборки составлял 20%. Одним из основных результатов предложенного подхода является процедура для снижения размерности и построения признакового пространства для кластеризации/классификации, которая базируется на вычислении взаимной информации для пар отведений. В рамках исследования вектор признаков формируется для каждого испытуемого.

Для оценки качества классификации использовались такие метрики, как *доля правильных ответов, чувствительность и специфичность*.

Анализ результатов вычислительного эксперимента позволил выявить наиболее значимые параметры для построения классификатора как для реальных, так и для воображаемых движений.

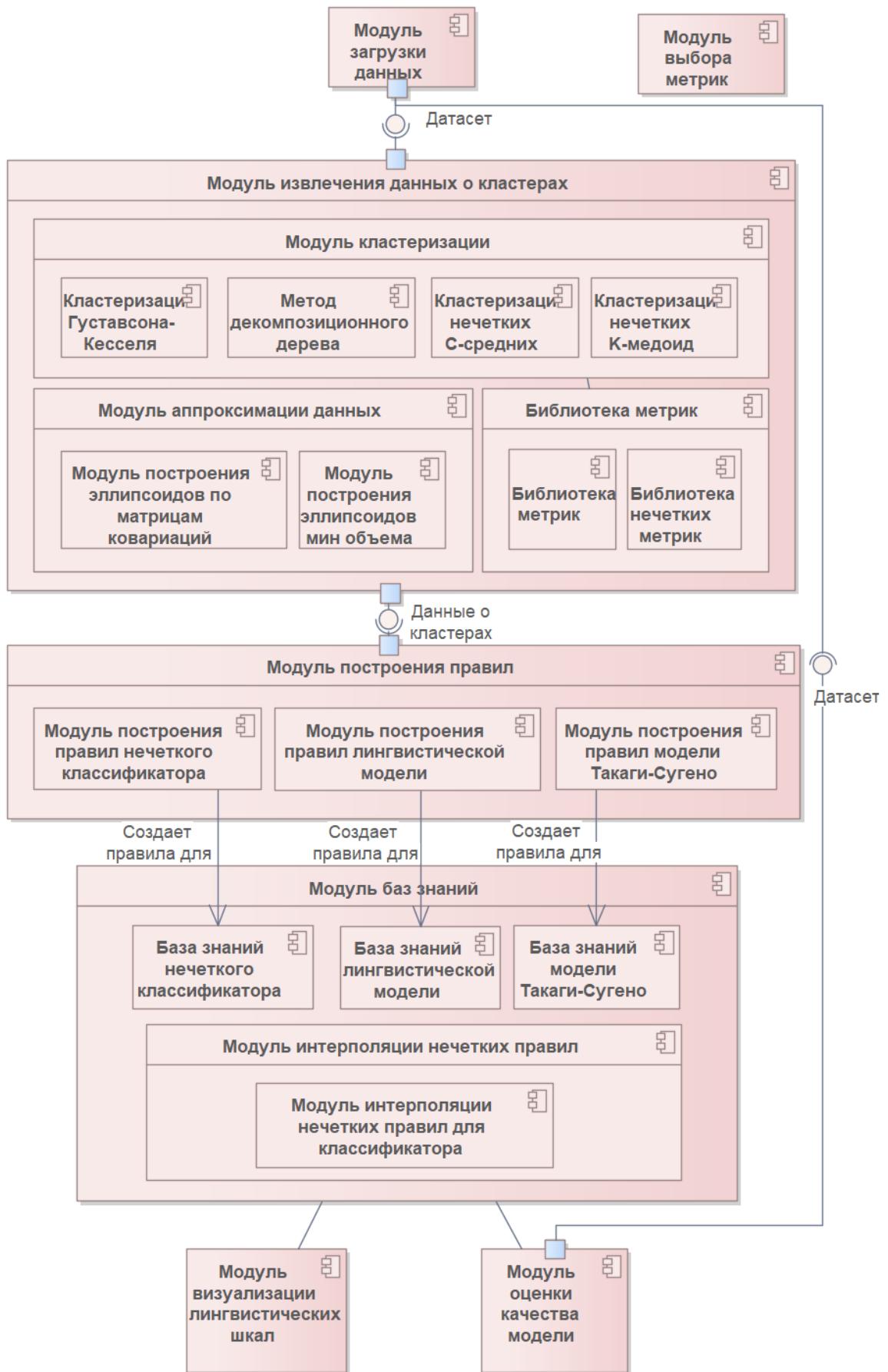


Рис. 5. Структура программного комплекса FuzzyLogicCore

Для 48% испытуемых удалось построить нечеткий классификатор для воображаемых движений, и для 72% испытуемых удалось построить классификатор для реальных движений. Средняя точность для 30 человек составляет 74% для реальных движений и 60% для воображаемых. Максимальная точность для отдельного человека составляет 96% при значении чувствительности 92% и значении специфичности 100% для реальных движений. Максимальная точность для отдельного человека для воображаемых движений составляет 71% при значениях чувствительности 78% и специфичности 62%. Важно, что в рамках исследования показана принципиальная возможность различения паттернов ЭЭГ реальных и воображаемых движений обеих рук с помощью нечеткого классификатора, анализирующего данные взаимной информации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Анализ подходов к построению баз знаний НПС показал, что при наличии наблюдаемых или экспериментальных данных, на основе которых можно сформировать обучающую выборку, для автоматизации данного процесса целесообразно использовать процедуры кластеризации, в том числе метрические алгоритмы.

2. Для автоматического построения баз знаний НПС на основе кластерных процедур предложен метод, основанный на построении эллипсоидов минимального объема (эллипсов минимальной площади), что позволило по сравнению с подходом на основе ковариационных матриц улучшить качество аппроксимации тестовых данных при том же уровне интерпретируемости примерно на 6-7 %.

3. Впервые предложено и исследовано семейство нечетких метрик, базирующихся на аддитивных генераторах непрерывных архimedовых треугольных норм из класса рациональных функций. Установлено, что нечеткие метрики имеют характерный уровень различимости, что важно при разработке рекомендации по их использованию в задачах кластеризации. Анализ результатов вычислительного эксперимента показал значительное превосходство нечетких метрик перед обычными метриками по большинству показателей качества кластеризации. Предложен подход к формированию комплексной оценки нечетких метрик с учетом этих показателей.

4. Разработан программный комплекс для автоматизированного формирования баз знаний НПС, который был применен для ряда задач, среди которых – построение нечеткого классификатора для интерфейса «мозг-компьютер» с целью анализа электроэнцефалограмм. Для 48% испытуемых удалось построить нечеткий классификатор для воображаемых движений, и для 72% испытуемых был построен классификатор для реальных движений. Кроме того, в процессе исследования были выявлены наиболее значимые признаки для классификации, что позволило сократить количество электродов в эксперименте до 6, а также установлены наиболее значимые параметры алгоритма для построения классификатора.

Рекомендации по использованию. Результаты, касающиеся новых нечетких метрик, расширяют теоретическую базу для построения метрических пространств на основе непрерывных рациональных архimedовых треугольных норм. Их целесообразно использовать в метрических алгоритмах кластеризации для управления степенью различимости объектов. Процедура формирования базы знаний на основе эллипсоидальной кластеризации может использоваться при разработке интеллектуальных информационных систем различного назначения в ситуации, когда имеются исторические или наблюдаемые данные значительного объема.

Перспективы развития проведенного исследования связаны, прежде всего, с оптимизацией сформированной базы правил, а также с интерполяцией нечетких правил, что позволит реализовать процедуру логического вывода с использованием неполной базы правил.

Публикации по теме диссертации

Публикации в рецензируемых научных журналах из Перечня ВАК РФ

1. Леденева, Т.М. Нечеткие метрики на основе генераторов архимедовых треугольных норм из класса рациональных функций / Т.М. Леденева, Т.А. Моисеева // Искусственный интеллект и принятие решений, 2024. – № 4. – С. 30-44. DOI: 10.14357/20718594240403
2. Моисеева, Т.А. Нечеткий классификатор электроэнцефалограмм для интерфейса «мозг-компьютер» / Т.А. Моисеева, Я.А. Туровский, Т.М. Леденева // Вестник Воронежского государственного университета. Системный анализ и информационные технологии, 2024. – № 4. – С. 129-142. doi.org/10.17308/sait/1995-5499/2024/4/129-142
3. Моисеева, Т.А. Влияние метрики на выявленную структуру временных задержек сигналов в задачах оценки электрогенеза мозга / Т.А. Моисеева, Я.А. Туровский // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, 2024. – № 2. – С. 103-112. DOI: 10.17308/sait/1995-5499/2024/2/103-112
4. Моисеева, Т. А. Генерация базы знаний на основе нечеткой кластеризации / Т. А. Моисеева, Т. М. Леденева // Информационные технологии и вычислительные системы. 2023. – № 1. – С. 97-108. DOI: 10.14357/20718632230110
5. Леденева, Т.М. Сравнительный анализ процедур кластеризации для обнаружения аномалий показателей, характеризующих функционирование контактной сети железных дорог / Т.М. Леденева, Т.А. Моисеева // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, 2020. – № 2. – С. 115–126. DOI: 10.17308/sait.2020.2/2921

Публикации в изданиях, индексируемых в базах Scopus и Web of Science

6. Moiseeva T. Knowledge Base Generation Based on Fuzzy Clustering / T. Moiseeva, T. Ledeneva // Programming and Computer Software, 2023. – Vol. 49. – No. 1. – Suppl. 2. – Pp. 99-107. DOI:10.1134/S0361768823100043
7. Moiseeva T. Missing Data Imputation Using Fuzzy System / T. Moiseeva, T. Ledeneva // 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russian Federation, 2022. – Pp. 350-354. DOI: 10.1109/SUMMA57301.2022.9974036
8. Ledeneva T. On One Approach to Constructing a Fuzzy Metric / T. Ledeneva, T. Moiseeva // 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russian Federation, 2023. – Pp. 175-179. DOI: 10.1109/SUMMA60232.2023.10349504
9. Ledeneva T. Generation of a Rule Base Based on Clustering of a Training Set / T. Ledeneva, T. Moiseeva // 2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russian Federation, 2024. – Pp. 102-107. DOI: 10.1109/SUMMA64428.2024.10803894.

Свидетельства о регистрации программ

10. Свидетельство о государственной регистрации программы для ЭВМ 2024661436 Российская Федерация. Программа для генерации нечеткой системы для решения задачи аппроксимации / Т.А. Моисеева; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный университет». – № 2024661436; заявление 13.05.2024 ; опубл. 17.05.2024.

11. Свидетельство о государственной регистрации программы для ЭВМ 2024661362 Российская Федерация. Программа для выполнения кластеризации методом нечетких к-медоид с использованием нечеткой метрики / Т.А. Моисеева; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный университет». – № 2024661362; заявление 13.05.2024 ; опубл. 17.05.2024.

Статьи и материалы конференций

12. Леденева Т.М. Обзор типов производственных правил в системах нечеткого логического вывода / Т.М. Леденева, Т.А. Моисеева // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научно-технической конференции, Воронеж, 2022 г. – С. 1793-1798.

13. Моисеева Т.А. Сравнительный анализ качества баз знаний, построенных на основе процедур кластеризации / Т.А. Моисеева, Т.М. Леденева // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции. Воронежский государственный университет. Воронеж, 2023. – С. 510-516.

14. Моисеева, Т. А. Реализация нечеткой системы для решения проблемы обработки пропусков в данных / Т.А. Моисеева // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции, Воронеж, 4-6 декабря 2023 г. – Воронеж, 2024. – С. 555-560.

15. Моисеева, Т.А. Разработка коэффициента неразличимости для нечеткой метрики / Т.А. Моисеева // Межвузовская научная конференция молодых ученых и студентов «Математика, информационные технологии, приложения», Воронеж, 24-25 апреля 2024 г. – Воронеж, 2024. – С. 573-580.