Министерство образования и науки Украины
Харьковский национальный университет радиоэлектроники

На правах рукописи

САМИТОВА ВИКТОРИЯ АЛЕКСАНДРОВНА

УДК 004.032.26

# КЛАССИФИКАЦИЯ И КЛАСТЕРИЗАЦИЯ ДАННЫХ, ЗАДАННЫХ В НЕЧИСЛОВЫХ ШКАЛАХ

05.13.23 – системы и средства искусственного интеллекта

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
Бодянский Евгений Владимирович,
доктор технических наук, профессор

Цей примірник дисертаційної роботи ідентичний за змістом з іншими, що подано до спеціалізованої вченої ради.

Вчений секретар спецради Д 64.052.01 О.А. Винокурова

Харьков – 2016

СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

Конец XX столетия характеризуется выходом человечества на новый этап своего технологического развития. Активное внедрение вычислительных технологий в различные области человеческой деятельности привело к пониманию важности такого понятия, как информация. Соответственно, дисциплины, которые изучают методы обработки информации, добычи знаний и принятия решений на основе полученных данных и знаний, становятся все более актуальными. Учитывая активную автоматизацию сложных процессов во многих отраслях, на первое место выходят дисциплины, связные с интеллектуальными методами обработки данных и принятия решений. В основе интеллектуального анализа данных лежит поиск скрытых, нетривиальных и полезных закономерностей в данных, позволяющих получить новые знания об исследуемых объектах.

Известные статистические методы покрывают лишь часть нужд по обработке данных и для их использования необходимо иметь четкое представление об искомых закономерностях. В такой ситуации методы интеллектуального анализа данных приобретают особую актуальность. Их основная особенность заключается в установлении наличия и характера скрытых закономерностей в данных, тогда как традиционные методы занимаются главным образом параметрической оценкой уже установленных закономерностей.

На сегодня эти методы получили широкое распространение для решения различных задач обработки сигналов, оптимизации, оптимального и адаптивного управления, распознавания образов, идентификации, прогнозирования в реальном времени и т.п. Созданы реальные системы обработки изображений и компьютерного зрения, управления аэрокосмическими объектами, технической и медицинской диагностики, в экономике и финансах, в военном деле, управления движением, в энергетике, в криминалистике, анализа сигналов различной природы и др., причем этот перечень постоянно расширяется. По мере того, как увеличивается область применения интеллектуальных систем, растут требования к

их универсальности, в частности, устойчивость на любых данных, адаптивность к меняющимся условиям, прозрачность интерпретации результатов.

**Актуальность темы.** Актуальными методами интеллектуального анализа являются классификация и кластеризация. Классификация, основываясь на заранее размеченной (тестовой) выборке, устанавливает закономерность, согласно которой данные разбиваются на классы. В свою очередь, кластеризация разбивает данные на кластеры, руководствуясь следующим принципом, – объекты внутри одного кластера должны быть максимально схожи между собой, тогда как объекты, принадлежащие разным кластерам, отличны. Для решения задач кластеризации используются различные методы и алгоритмы [1 - 7].

При этом обычно предполагается, что каждое наблюдение из выборки данных может принадлежать только одному кластеру.

Однако, часто встречается ситуация, когда анализируемый вектор наблюдений с определенными уровнями принадлежности может относиться сразу к нескольким кластерам. В данном случае целесообразно применять методы нечеткого кластерного анализа [8-10], широко используемые в настоящее время во множестве приложений, связанных с биологией, экономикой, социологией, образованием, видеообработкой и т.п. Отметим, что, обрабатывая информацию со сложной внутренней структурой, классические методы нечеткой кластеризации данных сталкиваются с определенными трудностями. Это связано, прежде всего, с рядом допущений, лежащих в основе этих методов. Разбиение на кластеры, которые имеют определенную форму и специфическую внутреннюю точку – центр кластера, осуществляется исходя из взаимосвязей между анализируемыми данными и прототипом кластера. Неверно подобранная метрика может существенно повлиять на точность кластеризации.

Часто в таких областях, как социология, медицина, образование и т.п. данные представлены в нечисловых шкалах. Характерными примерами являются социальные опросы, описание симптомов заболевания (порядковая шкала) или данные о покупках в магазине (категориальная шкала). Отметим, что человеку гораздо привычнее оперировать порядковой шкалой, чем числовой.

Несмотря на большое количество научных работ, все еще существует проблема работы с данными, представленными в нечисловых шкалах, вызванная потребностью в методах кластеризации, работающих в условиях нечетких (пересекающихся или перекрывающихся) кластеров, а также таких методов, которые могут работать в последовательном режиме.

В связи с этим разработка методов нечеткой кластеризации и классификации данных, заданных в нечисловых шкалах, является актуальной задачей, что определяет перспективность как теоретических, так и практических результатов.

**Связь работы с научными программами, планами, темами.**

Диссертационная работа выполнена в рамках госбюджетных НИР: №273 «Нейро-фаззи системы для текущей кластеризации и классификации последовательностей данных в условиях их искаженности отсутствующими и аномальными наблюдениями» (№ДР 0113U000361); № 307 «Динамический интеллектуальный анализ последовательностей нечеткой информации в условиях существенной неопределенности на основе гибридных систем вычислительного интеллекта» (№ДР 0116U002539). В рамках указанных НИР соискателем разработаны следующие методы и архитектуры: методы нечеткой кластеризации данных, заданных в порядковой шкале; методы нечеткой кластеризации данных, заданных в категориальной шкале; архитектура гибридной системы вычислительного интеллекта для классификации порядковых данных.

**Цель и задачи исследования.** Целью настоящей диссертационной работы является разработка методов нечеткой кластеризации и классификации данных, заданных в нечисловых шкалах.

Согласно поставленной цели необходимо решить следующие научные задачи:

- проанализировать существующие методы и подходы к интеллектуальной обработке данных, представленных в нечисловых шкалах;
- синтезировать адаптивные и робастные методы нечеткой кластеризации порядковых данных;
- синтезировать метод возможностной нечеткой кластеризации данных, заданных в категориальной шкале;

– синтезировать архитектуру гибридной системы вычислительного интеллекта для классификации порядковых данных и метод её обучения;

– решить при помощи разработанных методов и моделей ряд практических задач интеллектуального анализа данных.

*Объект исследования*. Процесс нечеткой кластеризации и классификации данных, представленных в нечисловых шкалах.

*Предмет исследования*. Методы нечеткой кластеризации и классификации данных, заданных в нечисловых шкалах, с использованием гибридных систем вычислительного интеллекта.

**Методы исследования.** Основные результаты работы получены на основе теории теории оптимизации и статистического анализа – для нахождения скрытых закономерностей в информации; теории нечеткой кластеризации – для разработки адаптивных и робастных методов кластеризации данных, заданных в нечисловых шкалах, в условиях пересекающихся кластеров; теории искусственных нейронных сетей для построения архитектуры нейро-фаззи системы, обладающей возможностью классификации порядковых данных; имитационного моделирования для определения эффективности применения разработанных методов и архитектур.

**Научная новизна полученных результатов.** К новым, полученных лично автором, относятся следующие результаты:

1. Впервые предложен метод нечеткой кластеризации порядковых данных путем использования частотных прототипов и функций принадлежности, что позволило обрабатывать порядковые данные, не подчиняющиеся нормальному распределению.

2. Впервые предложен адаптивный метод нечеткой кластеризации порядковых данных на основе порядково-цифрового отображения, что позволило обрабатывать порядковые данные в последовательном режиме.

3. Усовершенствован метод нечеткой кластеризации данных, заданных в порядковой шкале, путем совместного использования функций принадлежности и

функции правдоподобия, что позволило повысить точность кластеризации порядковых данных.

4. Усовершенствован метод возможностной нечеткой кластеризации массивов категориальных данных путем совместного использования частотных прототипов и мер несходства, что позволило преодолеть недостатки классических методов и повысить точность кластеризации данных.

5. Усовершенствована нейро-фаззи система на основе нео-фаззи нейрона путем введения дополнительного выходного слоя, что позволило обрабатывать данные, заданные в порядковой шкале.

6. Получили дальнейшее развитие методы адаптивной робастной нечеткой кластеризации порядковых данных путём введения критерия специального вида (меры схожести), что позволило обрабатывать порядковые данные, содержащие выбросы, в последовательном режиме.

**Практическое значение полученных результатов.** Предложенные в работе методы нечеткой кластеризации и классификации данных, заданных в нечисловых шкалах, доведены до программной реализации, что позволило автоматизировать процесс обработки данных в порядковой и категориальной шкалах для решения задач интеллектуального анализа данных. Разработанные методы и архитектуры позволяют проводить анализ данных, заданных в нечисловых шкалах, в последовательном режиме и повысить точность их обработки. Имитационное моделирование полученных теоретических результатов показало их преимущество над существующими методами.

Результаты диссертационной работы были внедрены в ООО «Южэлектропроект», г. Харьков, где использовались при решении задачи анализа клиентской базы данных предприятия, что подтверждено соответствующим актом внедрения.

Также, один из разработанных методов внедрен в ООО НПФ «Мидиэл», г. Харьков, где был использован для автоматической обработки термограмм при решении задачи диагностики электрооборудования, что подтверждено соответствующим актом внедрения.

Изложенные в диссертации научные положения, выводы и рекомендации были использованы в научно-исследовательских работах Харьковского национального университета радиоэлектроники, а также при подготовке курсов «Искусственные нейронные сети: архитектуры, обучение и применение» и «Нейросетевые методы вычислительного интеллекта», что подтверждается соответствующими актами.

**Личный вклад соискателя.** Все результаты диссертационной работы, которые выносятся на защиту, получены автором самостоятельно. Вклад автора в публикациях, написанных в соавторстве такой: [11] – разработан метод нечеткой кластеризации данных, заданных в порядковой шкале, на основе частотных прототипов и функций принадлежности; [12] – предложен метод нечеткой кластеризации данных, представленных в порядковой шкале, на основе совместного использования функций принадлежности и функции правдоподобия; [14] – разработана архитектура нейро-фаззи системы на основе двойного нео-фаззи нейрона для обработки данных в порядковом виде; [15] – предложены адаптивные методы робастной нечеткой кластеризации данных в порядковой шкале на основе меры схожести; [16] – предложен возможностной метод нечеткой кластеризации массивов категориальных данных с использованием частотных прототипов и мер несходства.

Работа [13] опубликована без соавторов.

**Апробация результатов диссертации.** Основные положения и результаты диссертационной работы были представлены, докладывались и обсуждались на международных научных конференциях: 12-й, 19-й, 20-й Международных молодежных форумах «Радиоэлектроника и молодежь в XXI веке» (Харьков, Украина, 2007-2016 гг.); Международной научно-технической конференции «Системный анализ и информационные технологии» (Киев, Украина, 2007 г.); Международной научно-технической конференции «Полиграфические, мультимедийные и web-технологии» (Харьков, Украина, 2016 г.); XII Международной научной конференции «Интеллектуальные системы принятия

решений и проблемы вычислительного интеллекта (ISDMCI’2016)» (Железный порт, Украина, 2016 г.).

**Публикации.** Основные положения диссертационной работы опубликованы в 12 научных работах (из них 4 единолично): 5 статей в научных периодических изданиях Украины по техническим наукам, 1 статья в зарубежном издании и 6 тезисов докладов на международных конференциях.

# РАЗДЕЛ 1
# ОБЗОР СОСТОЯНИЯ ПРОБЛЕМЫ И ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

## 1.1 Данные и шкалы их измерения

Данные – это получаемые человеком факты, сообщения, события, измеряемые характеристики, регистрируемые сигналы.

Данные получаются в результате измерений. Под измерением понимается приписывание символьных форм объектам или событиям в соответствии с определенными правилами. Эти символы могут быть как числовыми, так и любая другая формальная знаковая система. Правила, на основании которых числа приписываются объектам, определяют шкалу измерения.

Как правило, шкалы измерений [17 - 18] классифицируют по типам измеряемых данных, определяющих допустимые для данной шкалы математические преобразования, а также типы отношений, отображаемых этой шкалой.

Выделяют следующие виды шкал:

- шкала наименований;
- шкала порядка;
- шкала интервалов;
- шкала отношений;
- шкала абсолютных значений.

Шкала наименований (номинальная шкала). Это самая простая из всех шкал. В ней числа выполняют роль меток и предназначены для обнаружения и различения изучаемых объектов. В номинальной шкале могут измеряться номера паспортов, страховок, машин, телефонов. Кроме того, такие признаки, как пол человека, его раса, национальность, цвет волос и т.п. также относятся к этой шкале. Числа в этой шкале нельзя складывать и вычитать, но можно подсчитывать, как часто встречается то или иное значение.

Шкала порядка. Символы, составляющие шкалу порядка, принято называть рангами, а саму шкалу – ранговой. В данной шкале признаки упорядочены по рангам (занимаемым местам), но интервалы между ними точно измерить невозможно.

Порядковые шкалы применяются в таких отраслях человеческой деятельности, как: медицина, экономика, психология, социология, экология, образование и т.п.

Примером порядковой шкалы могут служить шкала Мооса в минералогии, шкала силы землетрясений Рихтера и бофортова шкала ветров в географии.

В медицине порядковые шкалы используют для определения степени выраженности коронарной недостаточности (классификация по И. Фогельсону), стадий гипертонической болезни (классификация по А.Л. Мясникову), степени сердечно-сосудистой недостаточности (классификация по Г.Ф. Лангу, Н. Д. Стражеско, В. Х. Василенко), групп инвалидности и т.п.

Также, примерами шкал порядка могут служить различные системы оценивания знаний в образовании (пяти бальная, двенадцати бальная, сто бальная система оценивания) или военные и академические звания.

Человек более свободно оперирует качественными характеристиками, чем количественными. То есть, ему легче сказать, какой из двух людей выше, чем указать их точный рост в сантиметрах.

Поэтому мнения экспертов часто выражаются именно в порядковой шкале.

Шкала интервалов. В данной шкале числа упорядочены по рангам, а также разделены определенными интервалами. Точка отсчета и единица измерения в этой шкале выбираются произвольно. Примерами могут служить температурные шкалы, потенциальная энергия поднятого груза, летоисчисление и т.п.

Допустимыми преобразованиями в шкале интервалов являются линейные преобразования.

Шкала отношений. Эта шкала похожа на шкалу интервалов, однако в ней четко определено положение точки отсчета. В шкале отношений действует

отношение «во столько-то раз больше», то есть экспериментально определяются отношения одной величины к другой подобной, принятой за единицу.

В этой шкале могут быть измерены большинство физических единиц: сила, длина, масса, цена.

Шкала абсолютных величин. В данной шкале происходит измерение величины чего-либо. Например, непосредственно подсчитывается количество учеников в классе, количество построенных домов, количество собранного урожая, количество прожитых лет и т.п. В подобной шкале существует абсолютный ноль.

Свойства шкалы абсолютных величин схожи со свойствами шкалы отношений, однако, величины, представленные на этой шкале, имеют абсолютные, а не относительные значения. Результаты измерений по шкале абсолютных величин чувствительны к неточностям измерений, однако, имеют наибольшую информативность и точность.

Все шкалы измерения делят на две группы - шкалы качественных признаков и шкалы количественных признаков.

Порядковая шкала и шкала наименований относятся к шкалам качественных признаков. Таким образом, во многих случаях результаты качественного анализа можно рассматривать как измерения по этим шкалам.

К шкалам количественных признаков относятся шкалы интервалов, отношений, разностей, абсолютная.

## 1.2 Нечеткая логика

Классическая (булева) логика имеет древнюю историю развития. Ее основателем является Аристотель. Булева логика оперирует только двумя понятиями: «истина» со значением истинности «1» и «ложь» со значением истинности «0».

Наличие приближенных и нечетких суждений в процессе постановки и решения различных задач человеком послужило толчком к развитию нечеткой логики (fuzzy logic). Впервые этот термин был предложен американским ученым

Лотфи Заде (Lotfi Zadeh) в его работе «Fuzzy Sets», опубликованной в 1965 году в журнале «Information and Control» [19].

Заде предложил обобщить множество значений истинности высказываний до интервала [0;1]. Такой подход позволяет рассматривать классическую бинарную логику как частный случай нечеткой логики, а также выполнять рассуждения с неопределенностью и оперировать высказываниями с различными значениями истинности.

## 1.3 Нечеткие множества

Одним из главных понятий систем, основанных на нечеткой логике, является понятие нечеткого множества (fuzzy set) [19]. Пусть существует $E$ – универсальное множество, $x$ – элемент $E$, а $S$ – определенное свойство. Обычное (четкое) подмножество $A$ универсального множества $E$, элементы которого удовлетворяют свойству $S$, определяется как множество упорядоченной пары $A=\{\mu_A(x)/x\}$, где $\mu_A(x)$ – характеристическая функция, которая принимает значение 1, в случае если $x$ удовлетворяет свойству $S$, и 0 – в противном случае.

Нечеткое подмножество $A$ универсального множества $E$ определяется как множество упорядоченных пар $A=\{\mu_A(x)/x\}$, где $\mu_A(x)$ – характеристическая функция принадлежности, которая показывает уровень принадлежности элемента $x$ подмножеству $A$. Чем выше степень принадлежности, тем в большей мере элемент универсального множества соответствует свойствам нечеткого множества.

### 1.3.1 Функции принадлежности и методы их построения

Существует большое количество типовых форм кривых для задания функций принадлежности, среди них треугольная и трапецеидальная функции, квадратичный и гармонический $Z$ - сплайны, квадратичный и гармонический $S$ - сплайны, $Z$ - сигмоидальная и $Z$ - линейная функции, $S$ - сигмоидальная и $S$ - линейная функции, колоколообразная и гауссова функции.

Наибольшее распространение получили: треугольная, трапецеидальная и гауссова функции принадлежности.

Треугольная функция принадлежности определяется тройкой чисел $(a,b,c)$, а ее значение в точке $x$ вычисляется согласно выражению

$$\mu(x)=\begin{cases}1-\dfrac{b-x}{b-a}, a\le x\le b,\\ 1-\dfrac{x-b}{c-b}, b\le x\le c,\\ 0, x\notin(a,c).\end{cases} \tag{1.1}$$

При $(b-a)=(c-b)$ треугольная функция становится симметричной.

Трапецеидальная функция принадлежности задается четверкой чисел $(a,b,c,d)$, а ее значение в точке $x$ вычисляется согласно выражению

$$\mu(x)=\begin{cases}1-\dfrac{b-x}{b-a}, a\le x\le b,\\ 1, b\le x\le c,\\ 1-\dfrac{x-c}{d-c}, c\le x\le d,\\ 0, x\notin(a,d).\end{cases} \tag{1.2}$$

При $(b-a)=(d-c)$ трапецеидальная функция принадлежности принимает симметричный вид.

Функция принадлежности гауссова типа оперирует двумя параметрами. Параметр $c$ является центроидом нечеткого множества, а параметр $\sigma$ отвечает за ширину функции. Описывается функция следующей формулой:

$$\mu(x) = \exp\left[ -\left( \frac{x-c}{\sigma} \right)^2 \right]. \quad (1.3)$$

Выделяют прямые и косвенные методы построения функций принадлежности [20 - 22].

Метод относительных частот, параметрический метод, интервальный метод относятся к прямым методам построения функций принадлежности и применяются для таких атрибутов как время, скорость, температура, влажность и т.п.

Когда отсутствуют измеримые свойства объектов, наиболее подходящими представляются косвенные методы построения функций принадлежности. Примером такого метода является метод парных сравнений.

Кроме того, все методы делятся, в свою очередь, на одиночные и групповые.

### 1.3.2 Нечеткие и лингвистические переменные

Понятия нечеткой и лингвистической переменных встречаются в случае описания наблюдений с помощью нечетких множеств [23 - 24]. Нечеткая переменная – это набор $(x, E, A)$,

где $x$ – это название переменной;

$E$ – универсальное множество;

$A$ – нечеткое множество на $E$.

Лингвистическая переменная состоит из набора $(b, T, E, G, M)$,

где $b$ – имя лингвистической переменной;

$T$ – множество его значений (терм-множество), которые представляют собой названия нечетких переменных;

$E$ – универсальное множество;

$G$ – синтаксическая процедура, позволяющая генерировать новые термы с использованием слов естественного или формального языка;

$M$ – семантическая процедура, которая каждому значению лингвистической переменной ставит в соответствие нечеткое подмножество множества $E$.

### 1.3.3 Операции над нечеткими множествами

В теории нечетких множеств применяются стандартные логические операторы AND, OR и NOT, осуществляемые над нечеткими множествами [23]. Операторы используются для записи комбинаций логических понятий нечеткой логики, чтобы вычислять степени истинности. Чаще всего используются операторы Заде, описанные далее:

1) пересечение (AND): $\mu_{(A\cap B)}(x) = \min\{\mu_A(x), \mu_B(x)\}$;

2) объединение (OR): $\mu_{(A\cup B)}(x) = \max\{\mu_A(x), \mu_B(x)\}$;

3) отрицание (NOT): $\mu_{(\neg A)}(x) = 1 - \mu_A(x)$,

где $A$ и $B$ – нечеткие множества.

Общий подход к выполнению операторов пересечения, объединения и дополнения в теории нечетких множеств реализуется в так называемых треугольных нормах и конормах. Известно, что любые нормы и конормы могут использоваться в качестве конъюнкции и дизъюнкции в системах нечёткого логического вывода. Наиболее распространенные случаи $t$ - нормы и $t$ - конормы – приведенные выше реализации операций пересечения и объединения, которые удовлетворяют свойствам монотонности, коммутативности, ассоциативности и законам Де Моргана.

## 1.4 Системы нечеткого вывода и нечеткое управление

Нечеткий вывод занимает основное место в нечеткой логике и системах нечеткого управления. Он представляет собой некоторую процедуру получения нечетких заключений на основе нечетких условий или предпосылок с использованием понятий нечеткой логики.

Типовая структура модели на основе нечеткого логического вывода, представленная на рисунке 1.1 [26], состоит из следующих блоков: нечеткая база знаний, база данных, машина нечеткого логического вывода, фаззификатор и дефаззификатор.

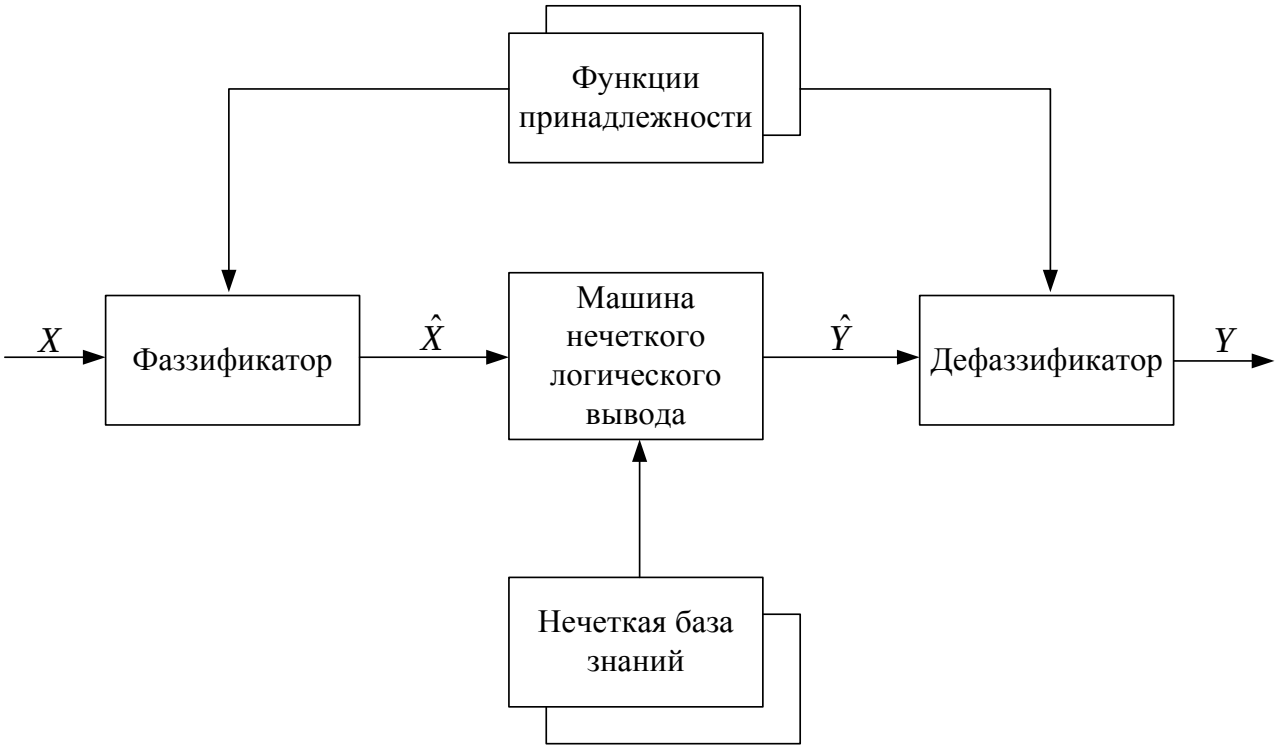



Рисунок 1.1 – Система нечёткого логического вывода

Используя нечеткие правила продукций, системы нечеткого вывода преобразуют значения входных переменных процесса управления в выходные переменные. Нечеткий вывод заключений реализуется на основе созданной базы правил нечетких продукций.

Основными этапами нечеткого вывода являются следующие шаги.

Шаг 1. Формирование базы правил систем нечеткого вывода.

Шаг 2. Фаззификация входных переменных.

Шаг 3. Агрегирование подусловий в нечетких правилах продукций.

Шаг 4. Активизация или композиция подзаключений в нечетких правилах продукций.

Шаг 5. Аккумулирование заключений нечетких правил продукций.

К настоящему времени предложено несколько алгоритмов нечеткого вывода. Среди них наиболее распространены алгоритмы Мамдани, Такаги - Сугено, Цукамото [24 - 26], которые наиболее часто применяются в системах нечеткого вывода.

## 1.5 Задача кластеризации

Кластеризация является одной из наиболее важных задач Data Mining. В настоящее время разработано большое количество методов и алгоритмов кластеризации.

Задача кластеризации часто может быть определена как задача оптимизации разбиения данных на группы. При этом под оптимальностью можно понимать минимизацию среднеквадратической ошибки разбиения

$$\mathrm{e}^2\left(x(k)\right) = \sum_{k=1}^{N}\sum_{i=1}^{r}\left\|x(k) - c_i\right\|^2,$$

где $r$ – количество кластеров;

$N$ – объем выборки данных;

$k = 1, 2, ..., N$ – номер наблюдения;

$c_i$ – прототип (центроид) $i$ - го кластера.

Важной частью процесса кластеризации является выбор подходящей функции сходства (подобия) для вычислительного процесса. Следует отметить, что вычисление функции схожести во многом зависит от того, в какой шкале представлены данные.

Формальная постановка задачи кластеризации выглядит следующим образом. Пусть задана выборка данных $X$, состоящая из $N$ $r$ - мерных векторов признаков $X = \{x(1), x(2), ..., x(N)\} \subset R^n$, $x(k) \in X$, $k = 1, 2, ..., N$ – это номер наблюдения или текущее дискретное время в задачах on-line обработки. Задано множество меток кластеров $Cl = \{cl_1, cl_2, ..., cl_r\}$, $i = 1, 2, ..., r.$ Для выборки $X$

определена некоторая функция расстояния между объектами $p(x, x')$.

Результатом работы алгоритма является разбиение исходного массива данных $X$ на $r$ классов (кластеров) с вычислением уровня принадлежности $U_i(k)$ $k$ - го вектора признаков $i$ - му кластеру таким образом, чтобы объекты внутри кластера были более похожи между собой, чем объекты разных кластеров.

Существует множество метрик [27, 37], с помощью которых определяется схожесть объектов внутри кластера.

Евклидово расстояние. Наиболее распространенная функция расстояния. Представляет собой геометрическое расстояние в многомерном пространстве

$$p(x, x') = \sqrt{\sum_{j=1}^{n} (x_j - x_j')^2}. \tag{1.4}$$

Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам, при этом

$$p(x, x') = \sum_{j=1}^{n} (x_j - x_j')^2. \tag{1.5}$$

Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является суммой модулей по координатам, при этом

$$p(x, x') = \sum_{j=1}^{n} | x_j - x_j' |. \tag{1.6}$$

В большинстве случаев эта мера расстояния приводит к таким же результатам, как и обычное расстояние Евклида. Однако, для нее характерно уменьшение влияния больших выбросов.

Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате, при этом

$$p(x, x') = \max(| x_j - x'_j |). \quad (1.7)$$

Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются, при этом

$$p(x, x') = \sqrt[\hat{z}]{\sum_{j=1}^{n} (x_j - x'_j)^{\hat{s}}}, \quad (1.8)$$

где $\hat{z}$ и $\hat{s}$ – параметры, определяемые пользователем.

Параметр $\hat{s}$ отвечает за постепенное взвешивание разностей по отдельным координатам, параметр $\hat{z}$ – за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра $\hat{z}$ и $\hat{s}$ равны двум, то это расстояние совпадает с расстоянием Евклида.

Решение задачи кластеризации сопряжено с закономерными трудностями такими, как:

– изначально неизвестно количество кластеров, на которые необходимо разбить выборку наблюдений;

– выбор метрики оказывает существенное влияние на результат кластеризации;

– не существует однозначно наилучшего критерия качества кластеризации, целевая функция выбирается эвристически в зависимости от поставленных целей;

– оценка качества кластеризации неоднозначна и может по-разному трактоваться.

Выбор данных параметров осуществляется субъективно и во многом зависит от эксперта.

Кластеризация помогает решить такие задачи, как:

– улучшение анализа данных за счет выявления структурных групп – использование различных методов анализа к различным кластерам позволяет улучшить и облегчить дальнейшую обработку данных;

– хранение данных в компактном виде;

– нахождение нетипичных объектов, которые не попали ни в один из кластеров.

## 1.6 Методы кластеризации

Можно выделить ряд подходов в классификации методов кластеризации:

1) вероятностный подход: дискриминантный анализ, $k$ - means [ 28, 29], $k$ - medians [30, 31], $EM$ - алгоритм [32, 33];

2) теоретико-графовый подход;

3) иерархический подход;

4) логический подход;

5) подход на основе систем искусственного интеллекта: генетические алгоритмы, метод нечеткой кластеризации $c$ - средних ($c$ - means) [8, 34], нейронные сети Кохонена [35, 36] и т.п.;

6) другие методы: ансамбли кластеризаторов, статистические алгоритмы кластеризации [37] и т.п.

В данной работе рассмотрены методы кластеризации, основанные на прототипах (prototype - based methods) [41, 42]. Они базируются на вычислении группы наиболее типичных наблюдений (центроидов, прототипов) из существующей выборки данных. В дальнейшем происходит разбиение наблюдений по кластерам согласно степени близости к выделенным прототипам.

С математической точки зрения данная задача может быть сформулирована в виде минимизации расстояний между наблюдениями и прототипами кластеров в некоторой метрике.

Нахождение таких прототипов $c=(c_1,c_2,...,c_i), \forall i \in [1,r]$, для которых сумма отклонений наблюдений выборки будет минимальной, идентично статистическому методу суммы наименьших квадратов.

Таким образом, целевая функция методов кластеризации, основанных на прототипах, может быть сформулирована следующим образом:

$$E\left(u_i(k),c_i\right)=\sum_{k=1}^{N}\sum_{i=1}^{r}u_i(k)d^2\left(x(k),\mathrm{c}_i\right), \tag{1.9}$$

где $c_i$ – множество центроидов кластеров, каждый из которых имеет ту же размерность, что и отдельное наблюдение $x(k)$;

$d\left(x(k),\mathrm{c}_i\right)$ – расстояние между прототипом $i$-го кластера и $k$-м наблюдением в некоторой метрике;

$u_i(k)$ – уровень принадлежности к кластеру: $u_i(k)=1$, если наблюдение $x(k)$ отнесено к $i$-му кластеру, и $u_i(k)=0$ в противном случае;

$N$ – объем выборки данных;

$r$ – количество кластеров;

$k=1,2,...,N$ – номер наблюдения.

Используя квадрат евклидова расстояния (1.5) при минимизации целевой функции, получаем известный алгоритм кластеризации – метод $k$-средних [43].

В различных задачах анализа данных часто возникает ситуация, когда обрабатываемый вектор признаков с различными уровнями вероятности (возможности, принадлежности, достоверности и т.п.) может принадлежать сразу нескольким классам. Данная ситуация является предметом рассмотрения нечеткого (fuzzy) кластерного анализа, который к настоящему времени достаточно широко используется в различных приложениях [8, 44 - 46].

Ситуация усложняется, когда данные заданы в нечисловых шкалах или в количественной и качественной шкалах одновременно [47]. В [48] для кластеризации порядковых данных предложено использовать гауссовские многомерные функции распределения (копулы), а в [49, 50] использовались

гауссовские латентные переменные. Однако, данные подходы плохо работают с данными, которые не подчиняются нормальному распределению. В [51] введена модель кластеризации порядковых данных на основе стохастического алгоритма двоичного поиска. В дальнейшем, используя латентные переменные и AECM (Alternating Expectation Conditional Maximization) алгоритм [51], данный подход был применен для обработки данных со смешанными характеристиками.

## 1.7 Классификация данных

Многие задачи Data Mining направлены на достижение цели, которая заключается в сопоставлении обрабатываемых наблюдений со специальными значениями характеристик в выборке данных. Эти значения характеристик определяются как метки классов, на которые необходимо разделить входные данные. Исходя из наличия или отсутствия прецедентной информации можно выделить задачи распознавания «с учителем» и «без учителя».

В случае наличия множества прецедентов задача распознавания относится к задачам обучения «с учителем» и называется классификацией. Данные, на основе которых устанавливается существующая между наблюдениями закономерность, являются обучающей выборкой. Полученная вследствие обучения модель классификации (классификатор, решающее правило), может быть использована в дальнейшем для определения классов в данных, где они отсутствуют.

В случае, когда обучающая выборка представлена малым количеством наблюдений, эффективность классификации может быть низкой. В таких случаях классификатор может описывать специфичные случайные характеристики обучающей выборки и не выявлять общую закономерность, присутствующую в структуре остальных данных. Иными словами, такой классификатор достаточно точно классифицирует наблюдения, которые были использованы для его построения, и показывает низкую точность классификации на вновь поступивших данных.

Формальная постановка задачи классификации звучит следующим образом. Пусть задана выборка данных $X$, состоящая из $N$ $r$ - мерных векторов признаков

$X = \{x(1), x(2), ..., x(N)\} \subset R^n$, $x(k) \in X$, где $k = 1, 2, ..., N$ – это номер наблюдения или текущее дискретное время в задачах последовательной обработки. Пускай существует некоторая целевая зависимость – отображение $y^* : X -> Cl$ на множестве меток кластеров $Cl = \{cl_1, cl_2, ..., cl_r\}, i = 1, ..., r.$ Значения $y^*$ известны только на объектах конечной обучающей выборке $X^r = \{(x_1, cl_1), (x_2, cl_2), ..., (x_r, cl_r)\}$. Необходимо построить классификатор $\hat{g}(x) : X -> Cl$, с помощью которого можно классифицировать произвольный объект $x(k) \in X$.

Следует отметить отличие задач кластеризации и классификации. В первом случае необходимо разбить данные на $r$ групп (кластеров, классов) на основе их схожести, используя определенную меру сходства. Во втором случае также необходимо разбить данные на те же $r$ групп, однако, эта поставленная цель достигается с помощью модели классификации $\hat{g}(x)$, полученной на обучающей выборке $X^r$.

Существует множество моделей для классификации данных [37 - 40]:

1) деревья решений;
2) байесовская классификация;
3) продукционные модели;
4) машинное обучение;
5) статистический подход, в частности, линейная регрессия;
6) метод опорных векторов;
7) искусственные нейронные сети.

## 1.8 Искусственные нейронные сети

Исследование работы человеческого мозга стало толчком к развитию такой области искусственного интеллекта, как нейронные сети.

Человеческий мозг – это очень сложная система способная к мышлению, обучению и накоплению данных. Основным его строительным материалом являются нервные клетки или нейроны. Это простые элементы обработки сигналов,

связанные между собой с помощью входов (дендритов). Нейроны способны переходить в возбуждённое состояние, генерируя сигнал на выходе - аксоне, который связан с дендритами множества других нейронов. В результате электрохимических процессов в нейроне каждый сигнал, проходя через синаптическое соединение, или ускоряется, или замедляется. Данное изменение синаптических связей отвечает за процесс запоминания информации и обучения.

Искусственные нейроны моделируют структуру и функции биологических нейронов. Архитектура и особенности искусственных нейронных сетей, образованных нейронами, зависят от конкретных задач, решаемых с их помощью. Структура искусственного нейрона представлена на рисунке 1.2.

Таким образом, можно сказать, что искусственная нейронная сеть (ИНС) – это параллельно распределённая система обработки информации, образованная тесно связанными простыми вычислительными узлами, которая способна накапливать и обобщать данные, делая их доступными для пользователя в форме удобной для интерпретации и принятия решений.

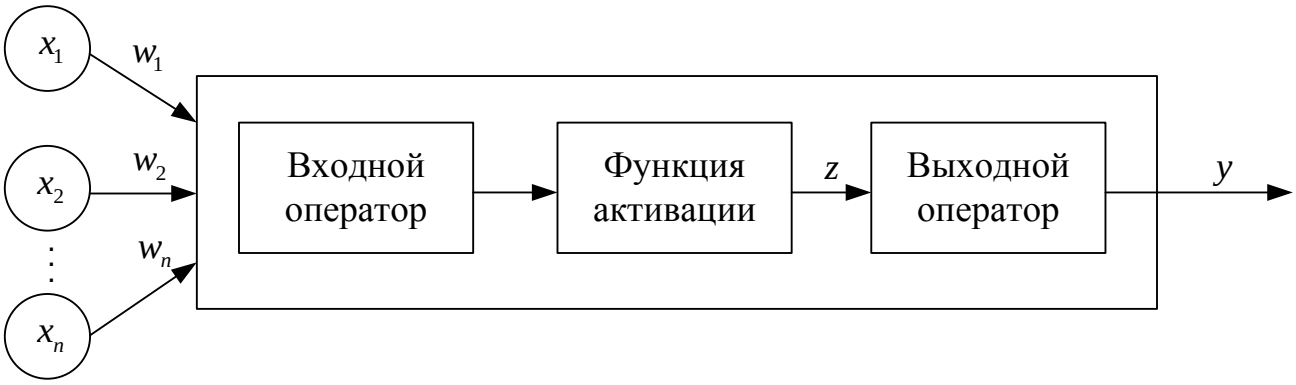



Рисунок 1.2 – Структура искусственного нейрона

Функционирование искусственной нейронной сети отображает работу мозга в следующих аспектах:

– знания накапливаются из окружающей среды в процессе обучения;

– обучение выполняется путем определенного изменения синаптических весов, отражающих связь между нейронами, или архитектуры сети.

На текущий момент нейронные сети принято классифицировать по следующим признакам [52 - 54]:

- по типу образующих сеть нейронов - узлов;
- по способу обучения;
- по архитектуре сети;
- по функциям, реализуемым сетью.

## 1.9 Обучение искусственных нейронных сетей

Одним из главных свойств нейронной сети является ее обучаемость, которая заключается в выработке правильной реакции на предъявляемые ей входные сигналы. Существуют следующие возможности обучения ИНС:

- изменение элементов матрицы связи (весов);
- изменение характеристик нейронов;
- изменение конфигурации сети путем образования новых или исключения существующих связей меду нейронами.

Наиболее популярным стал подход, при котором сеть обучается путем настройки матрицы весовых коэффициентов, при этот сама топология сети задается априорно.

Различают обучение «с учителем» и «без учителя» [54]. Парадигма обучения «с учителем» схематически представлена на рисунке 1.3.

Под «учителем» в данном случае понимается либо сама обучающая выборка, либо тот, кто указал на заданных объектах правильные ответы. В данной схеме «учителю» известна информация о внешней среде, представленная в виде последовательности входных векторов $x$, а также «правильный ответ» на эти векторы в виде обучающего сигнала $\tilde{d}$. Реакция необученной сети $y$ будет отличаться от «правильной» реакции учителя, что приведет к возникновению ошибки $e = \tilde{d} - y$.

Целью обучения нейронной сети является такая настройка весовых коэффициентов, чтобы некоторая скалярная функция от ошибки $E(e)$ была

минимальна.

Другой парадигмой обучения является парадигма обучения «без учителя», или самообучение, схематически представленная на рисунке 1.4. В этом случае правильная реакция на сигналы внешней среды неизвестна.

В основе самообучения в искусственной нейронной сети лежит правило, сформулированное Хеббом и декларирующее идею самоорганизации нейронной сети. Для ограничения роста весов связей в него вводятся механизм забывания, т.е. если нейроны активизируются не одновременно и не часто, то синаптический вес локальной связи этих нейронов уменьшается.

Сети, в основе которых лежит парадигма самообучения, обычно применяются для анализа внутренней структуры входной информации.

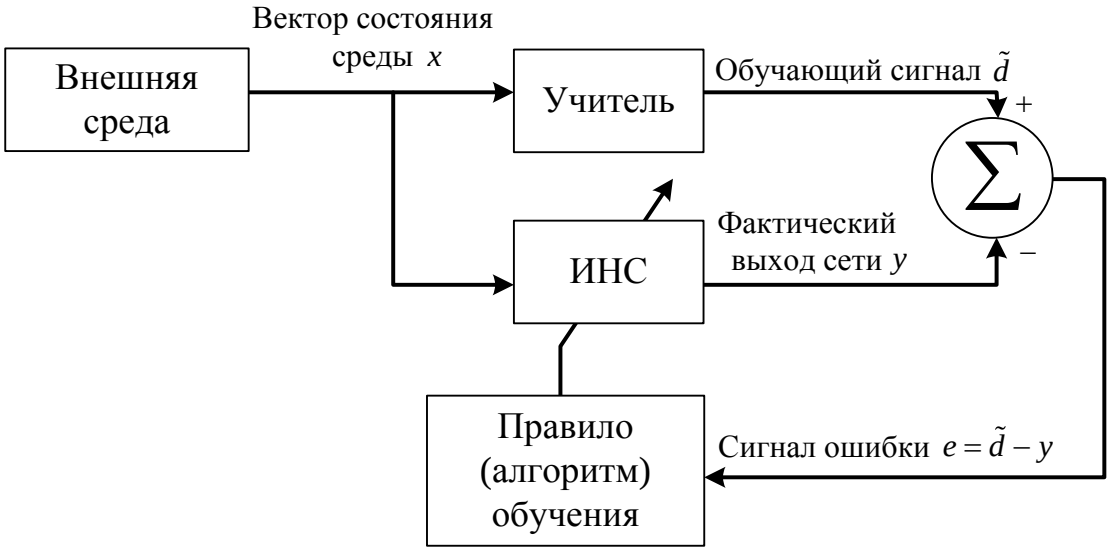



Рисунок 1.3 – Схема обучения с учителем

Известны такие типы задач обучения без учителя, как кластеризация, поиск правил ассоциации, заполнение пропущенных значений, сокращение размерности, факторный анализ и т.п.

С. Хайкин [54] сформулировал пять главных правил обучения, лежащих в основе алгоритмов: обучение на основе коррекции по ошибке, обучение по Больцману, обучение по Хеббу, обучение памяти и конкурентное обучение.

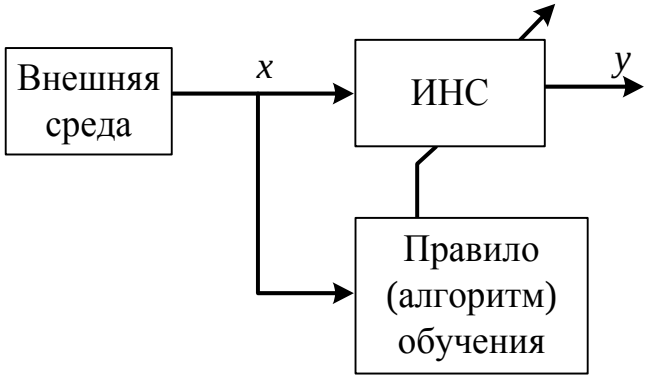



Рисунок 1.4 – Схема обучения «без учителя»

Выбор конкретного алгоритма обучения в целом зависит от типа входных данных и задач, решаемых нейронной сетью.

## 1.10 Нейро-фаззи системы

Развитие теории искусственных нейронных сетей привело к появлению новых методов и процедур обучения в нечетких системах. Процедура обучения нейронных сетей методом обратного распространения ошибки (error backpropagation) для настройки параметрических систем, в том числе и систем нечеткого вывода, позволяет не только строить системы на основе лингвистической информации, но также более точно настраивать их параметры для достижения поставленной задачи. Объединение нечеткой логики и нейронных сетей позволяет преодолеть трудности, связанные с каждой из этих технологий [55 - 57]. Например, решить проблему не интерпретируемости результатов, получаемых при помощи нейронных сетей, а, следовательно, создать более эффективные приложения.

На базе архитектуры Такаги-Сугено были предложены адаптивные сетевые системы нечеткого вывода (Adaptive-Network-Based Fuzzy Inference Systems, ANFIS) [58], реализующие адаптивный метод настройки параметров антецедента.

В [59] предложено настраивать заранее заданные правила в процессе обучения, что привело к получению более точных результатов. В [60] приведен ряд нейро-фаззи модификаций в конкурентных нейронных сетях.

Эквивалентность радиально-базисных нейронных сетей и ANFIS доказана в [61]. В NEFPROX [62] использованы хорошо интерпретируемые правила. Однако, данный подход привел к снижению точности аппроксимации.

## 1.11 Постановка задачи исследования

Повсеместная автоматизация различных областей человеческой деятельности, привела к необходимости быстро и точно обрабатывать большие массивы данных. В таких областях, как медицина, социология, образование данные могут быть представлены в нечисловых шкалах. Использование традиционных методов, основанных на замене лингвистических переменных их рангами, приводит к потере части информации о данных, поскольку не учитывает порядок следования атрибутов и расстояние между ними, что негативным образом сказывается на точности кластеризации данных.

При обработке больших массивов категориальных данных резко возрастает размерность пространства признаков, что приводит к увеличению вычислительной сложности и существенно усложняет решение задачи из-за возникновения эффектов «проклятья размерности», а в нечетком случае – «концентрации норм».

Несмотря на большое количество научных работ в данной области, все еще существует проблема работы с данными, заданными в нечисловых шкалах, вызванная потребностью в методах классификации и кластеризации данных, работающих в условиях перекрывающихся кластеров, а также таких методов, которые могут работать в on-line режиме.

Применение теории искусственных нейронных сетей и аппарата нечеткой логики дает возможность повысить точность обработки данных в нечисловых шкалах, а также расширить круг решаемых задач.

Задача исследования состоит в разработке методов классификации и кластеризации данных, заданных в нечисловых шкалах, в условиях перекрывающихся кластеров, в том числе, в последовательном режиме. Для достижения поставленной цели необходимо рассмотреть следующие вопросы.

1. Анализ известных нейро - фаззи архитектур и методов их обучения, а также методов кластеризации данных, заданных в нечисловых шкалах.

2. Разработка методов адаптивной нечеткой кластеризации данных, заданных в порядковой шкале, для обработки информации, поступающей в последовательном режиме.

3. Разработка методов робастной нечеткой кластеризации данных, заданных в порядковой шкале, сочетающих парадигмы возможностного и вероятностного подхода к нечёткой логике для обработки порядковых данных, которые содержат выбросы.

4. Разработка методов нечеткой кластеризации данных, заданных в категориальной шкале.

5. Разработка гибридной системы вычислительного интеллекта для обработки порядковых данных.

6. Проведение имитационного моделирования разработанных методов и архитектур и решение с их помощью практических задач.

Выводы по разделу 1

1. Проанализировано состояние проблемы классификации и кластеризации данных в нечисловых шкалах, рассмотрены существующие подходы к ее решению.

2. Рассмотрены основные принципы нечёткой логики и систем нечёткого вывода.

3. Рассмотрены существующие методы классификации и кластеризации данных. Выявлены и проанализированы их недостатки.

4. Проведён анализ существующих архитектур искусственных нейронных сетей и методов их обучения и самообучения, используемых для решения задач кластеризации данных.

5. Рассмотрены гибридные системы вычислительного интеллекта на стыке искусственных нейронных сетей и нечеткой логики – нейро-фаззи системы.

Проанализированы существующие архитектуры и методы их обучения в задачах интеллектуальной обработки данных.

6. Сформулирована задача исследования.

# РАЗДЕЛ 2
# МЕТОДЫ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ПОРЯДКОВЫХ ДАННЫХ

Часто в процессе кластеризации исследователь сталкивается с проблемами, связанными с тем, что некоторые характеристики данных неизвестны заранее.

Ситуация, когда исходные данные заданы не в числовой, а в ранговой (порядковой) шкале встречается в социологии, медицине, образовании и т.п. В одномерном случае такая информация задается в виде последовательности упорядоченных лингвистических переменных $x^1, x^2, ..., x^l, ..., x^m$, $1 < ... < l-1 < l < l+1 < ... < m$, где $x^l$ – собственно лингвистическая переменная, $l$ – соответствующий ранг.

Характерным примером является традиционная система оценок в образовании типа «плохо», «удовлетворительно», «хорошо», «отлично». Заметим, что в своей повседневной деятельности человек гораздо чаще пользуется порядковой шкалой, нежели числовой.

Для решения задач кластеризации данных на порядковой шкале наиболее простым представляется подход, основанный на замене лингвистических переменных их рангами, однако в большинстве случаев этот прием оказывается некорректным, поскольку предполагает равенство расстояний между соседними числовыми рангами. Интуитивно ясно, что при оценке знаний учащихся расстояние между «плохо» и «удовлетворительно» гораздо больше, чем расстояние между «удовлетворительно» и «хорошо».

Более естественным представляется подход, основанный на фаззификации исходных данных и дальнейшем использовании методов нечеткой кластеризации [47, 63]. При этом исходный набор лингвистических переменных $x^1, x^2, ..., x^l, ..., x^m$ заменяется множеством функций принадлежности $\mu_1(x), \mu_2(x), ..., \mu_m(x)$, заданных на интервале [0, 1]. Такой прием был использован в [64], где с помощью метода нечетких $c$ - средних (FCM) [23] производилась кластеризация не исходных данных, а параметров, описывающих соответствующие им функции принадлежности, хотя сам способ определения этих параметров не указан.

В данном разделе рассмотрен ряд подходов и основанных на них методов, направленных на обработку данных, представленных в порядковой шкале. Обобщение и универсализация параметров, использующихся для получения этих методов, позволяют создавать модели с минимальным числом настраиваемых вручную параметров, способные гибко приспосабливаться к изменениям характеристик исследуемых данных в широких пределах.

Отдельное внимание уделено схожести схем работы основных методов, что позволяет впоследствии рассматривать ансамбли подобных моделей, направленных на достижение поставленных задач.

## 2.1 Целевая функция нечёткой кластеризации

Нечёткая логика, вкратце описанная в разделе 1.3, почти сразу же стала применяться к задачам классификации и кластеризации.

Простейшая целевая функция нечёткой кластеризации может быть представлена следующим образом:

$$E(u_i, c_i) = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) d^2\left(x(k), \mathrm{c}_i\right), \tag{2.1}$$

удовлетворяющая ограничениям

$$\begin{cases} u_i(k) \geq 0, \forall i = 1,2,...,r; \forall k = 1,2,...,N, \\ \sum_{i=1}^{r} u_i(k) = 1, \forall k = 1,2,...,N, \\ \sum_{k=1}^{N} u_i(k) > 0, \forall i = 1,2,...,r, \end{cases} \tag{2.2}$$

где $N$ – объем выборки данных;

$r$ – количество кластеров;

$k = 1, 2, ..., N$ – номер наблюдения;

$c_i$ – прототип (центроид) $i$ - го кластера;

$u_i(k) \in [0,1]$ – уровень принадлежности вектора $x(k)$ к $i$ - му кластеру;

$d(x(k), c_i)$ – расстояние между прототипом $i$ - го кластера и $k$ - м наблюдением в некоторой метрике;

$\beta$ – неотрицательный параметр фаззификации (фаззификатор), определяющий размытость границ между кластерами.

Результатом кластеризации является матрица $U = \{u_i(k)\}$ размерности $N \times r$, называемая матрицей нечёткого разбиения, которая описывает степень принадлежности каждого наблюдения к каждому из кластеров.

## 2.2 Оптимизация целевой функции

Используя в качестве метрики $d\left(x(k), c_i\right)$ квадрат евклидового расстояния (1.5), можно записать функцию Лагранжа для целевой функции (2.1)

$$L\left(u_i(k), c_i, \lambda(k)\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k)\left\|x(k) - c_i\right\|^2 + \sum_{k=1}^{N}\lambda(k)\left(\sum_{i=1}^{r} u_i(k) - 1\right), \tag{2.3}$$

где $\lambda(k)$ – неопределённый множитель Лагранжа.

Решая систему уравнений Каруша-Куна-Таккера, приходим к результату

$$\begin{cases} \dfrac{\partial L\left(u_i(k), c_i, \lambda(k)\right)}{\partial u_i(k)} = 0, \\ \nabla_{c_i} L\left(u_i(k), c_i, \lambda(k)\right) = 0, \\ \dfrac{\partial L\left(u_i(k), c_i, \lambda(k)\right)}{\partial \lambda(k)} = 0, \end{cases} \tag{2.4}$$

где искомые переменные могут быть получены следующим образом:

$$\begin{cases} u_i(k) = \dfrac{\left(d^2\left(x(k), c_i\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{t=1}^{r}\left(d^2\left(x(k), c_t\right)\right)^{\frac{1}{1-\beta}}}, \\ c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}, \\ \lambda(k) = -\left(\sum\limits_{i=1}^{r}\left(\beta d^2\left(x(k), \mathrm{c}_i\right)\right)^{\frac{1}{1-\beta}}\right)^{1-\beta}. \end{cases} \quad (2.5)$$

При $\beta = 2$ данное решение совпадает с популярным алгоритмом нечётких $c$ - средних Бездека (fuzzy $c$ - means, FCM) [23]

$$\begin{cases} u_i(k) = \dfrac{d^{-2}\left(x(k), c_i\right)}{\sum\limits_{t=1}^{r} d^{-2}\left(x(k), c_t\right)}, \\ c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{2}(k) x(k)}{\sum\limits_{k=1}^{N} u_i^{2}(k)}, \end{cases} \quad (2.6)$$

а при $\beta = 1$ приближается к результатам, полученным при помощи чёткой кластеризации $k$ - средних (hard $k$ - means, HKM).

## 2.3 Вычисление центроидов кластеров на основе частотных прототипов

Предполагая, что выборка наблюдений имеет нормальное распределение, рассмотрим процесс фаззификации [65, 66] последовательности ранговых лингвистических переменных на примере одномерной выборки $x(1), x(2), ..., x(N)$,

где каждому из наблюдений $x(k)$ может быть приписан один из рангов $l$, $l = 1,2,...m$, например, рост человека: «очень низкий», «низкий», «средний», «высокий», «очень высокий».

Пусть значение $x(k)$, соответствующее $l$ - му рангу, встречается в выборке $N_l$ раз. Тогда можно ввести в рассмотрение относительные частоты появления $l$ - го ранга

$$f_l = \frac{N_l}{N} \tag{2.7}$$

и накопленные частоты

$$F_1 = \frac{f_1}{2}, F_l = \frac{f_l}{2} + \sum_{s=1}^{l-1} f_s, l = 2,3,...,m, \tag{2.8}$$

при этом естественно выполняется условие

$$\sum_{l=1}^{m} f_l = 1.$$

Хотя при нормальном распределении значения лежат в диапазоне от $-\infty$ до $+\infty$, $99,7\%$ кривой функции Гаусса лежит между $\mu - 3\sigma$ и $\mu + 3\sigma$, где $\mu$ – это математическое ожидание, а $\sigma$ – среднеквадратическое отклонение. Установим $\mu - 3\sigma$ в $0$, а $\mu + 3\sigma$ в $1$. В этом случае $\mu = 0.5$, а $\sigma = 0.16666$. Используя обратное нормальное распределение, получаем искомые центроиды, на основе которых строятся функции принадлежности.

Для фаззификации исходных данных берутся кусочно-линейные функции принадлежности. Трапецеидальные функции принадлежности используются на концах интервала, а треугольные – в середине.

Крайняя левая функция принадлежности описывается уравнением

$$\begin{cases} y = 1, \ \ p_0 \le x \le p_1, \\ \\ y = \dfrac{p_2 - x}{p_2 - p_1}, \ \ p_1 \le x \le p_2 \end{cases}$$

с центроидом

$$c = \frac{\dfrac{p_1^2 - p_0^2}{2} + \dfrac{p_2(p_1 + p_2)}{6} - \dfrac{p_1^2}{3}}{\dfrac{p_2 - p_1}{2} + p_1 - p_0}.$$

Крайняя правая функция принадлежности описывается уравнением

$$\begin{cases} y = \dfrac{x - p_0}{p_1 - p_0}, \ \ p_0 \le x \le p_1, \\ \\ \mathrm{y} = 1, \ \ p_1 \le x \le p_2 \end{cases}$$

с центроидом

$$c = \frac{\dfrac{p_2^2 - p_1^2}{2} + \dfrac{p_1^2}{3} - \dfrac{p_0(p_0 + p_1)}{6}}{\dfrac{p_1 - p_0}{2} + p_2 - p_1}.$$

Треугольная функция принадлежности задается уравнением

$$\begin{cases} y = \dfrac{x - p_0}{p_1 - p_0}, \ \ p_0 \le x \le p_1, \\ \\ y = \dfrac{p_2 - x}{p_2 - p_1}, \ \ p_1 \le x \le p_2 \end{cases}$$

с центроидом

$$c = \frac{p_0 + p_1 + p_2}{3}.$$

Имея нечеткое покрытие из $n$ нечетких множеств, мы получаем $n+2$ параметра $a_0, a_1, ..., a_{n+1}$ и $n$ уравнений для вычисления центроидов

$$c_0 = \frac{\frac{a_1^2 - a_0^2}{2} + \frac{a_2(a_1 + a_2)}{6} - \frac{a_1^2}{3}}{\frac{a_2 - a_1}{2} + a_1 - a_0}, \tag{2.9}$$

$$c_j = \frac{a_j + a_{j+1} + a_{j+2}}{3},\ j = 1, ..., n-2, \tag{2.10}$$

$$c_{n-1} = \frac{\frac{a_{n+1}^2 - a_n^2}{2} + \frac{a_n^2}{3} - \frac{a_{n-1}(a_{n-1} + a_n)}{6}}{\frac{a_n - a_{n-1}}{2} + a_{n+1} - a_n}, \tag{2.11}$$

где $a_0 = 0$, а $a_{n+1} = 1$.

Ограничением вышеописанного подхода является предположение о гауссовом распределении исходных данных, что во многих приложениях не выполняется.

## 2.4 Фаззификация данных, заданных в порядковой шкале

Используя формулы (2.7) и (2.8), вычисляются частоты встречаемости $l$ - го ранга $f_l$ и накопленные частоты $F_l$.

На основе накопленных частот формируются центры набора функций принадлежности $\mu_l(x)$ так, как это показано на рисунке 2.1, при этом для вычисления центроидов удобно воспользоваться рекуррентным соотношением

$$c_1 = 0{,}5F_1,\ c_l = c_{l-1} + 0{,}5\left(\mathrm{F}_{l-1} + F_l\right),\ l = 2,3,...m\,, \tag{2.12}$$

а сами функции принадлежности задать в форме

$$\begin{cases} \mu_1(x) = 1,\ \ x \in [0, c_1], \\ \mu_l(x) = \begin{cases} \dfrac{x - c_{l-1}}{c_l - c_{l-1}},\ \ x \in [c_{l-1}, c_l], \\ \dfrac{c_{l+1} - x}{c_{l+1} - c_l},\ \ x \in [c_l, c_{l+1}], \\ 0,\ \ x \notin [c_{l-1}, c_{l+1}], \end{cases} \\ \mu_m(x) = 1,\ \ x \in [c_m, 1]. \end{cases} \tag{2.13}$$

Такой способ задания функций принадлежности автоматически обеспечивает разбиение Руспини, т.е. выполнение условия

$$\sum_{l=1}^{m} \mu_l(x) = 1,$$

хотя, конечно, возможно использование функций иного вида с конечным носителем

$$\operatorname{supp} \mu_l(x) = [c_{l-1}, c_{l+1}].$$

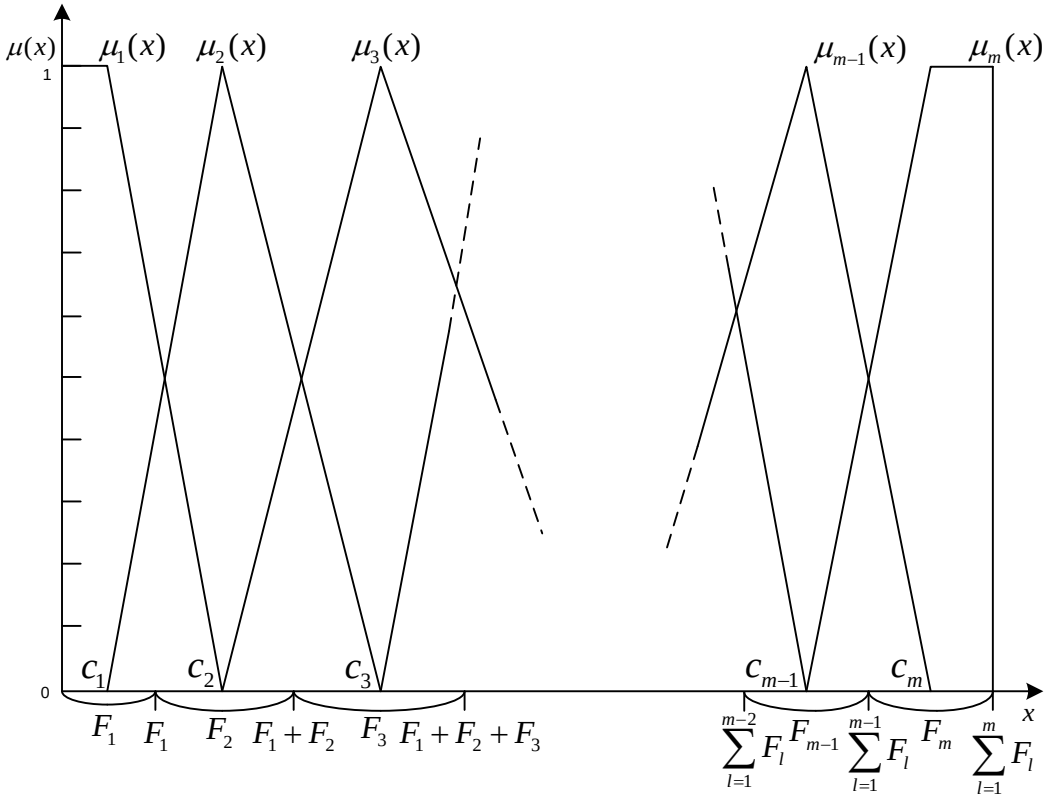



Рисунок 2.1 – Функции принадлежности для выборки ранговых переменных

Рассмотрим далее две соседние функции принадлежности $\mu_l(x)$ и $\mu_{l+1}(x)$ (рис.2.2). Используя понятие $\alpha$ - разреза в виде

$$A_\alpha = \{x \in X : \mu(x) \geq \alpha\},$$

можно ввести области влияния двух соседних рангов (на рисунке 2.2 заштрихованы) в форме

$$\begin{cases} A_l^R = \{x \in [c_l, c_l + 0{,}5\, f_l] : \mu_l(x) \geq \alpha_l^R = 1 - 0{,}5 \dfrac{f_l}{c_{l+1} - c_l}, \\ A_{l+1}^L = \{x \in [c_{l+1} - 0{,}5\, f_{l+1}, c_{l+1}] : \mu_{l+1}(x) \geq \alpha_{l+1}^L = 1 - 0{,}5 \dfrac{f_{l+1}}{c_{l+1} - c_l}, \end{cases} \quad (2.14)$$

где $R$ и $L$ обозначают правую или левую стороны соседних функций принадлежности.

При попадании некоторого наблюдения в область влияния конкретного ранга, можно говорить о «четкой» принадлежности этого наблюдения к данному рангу.

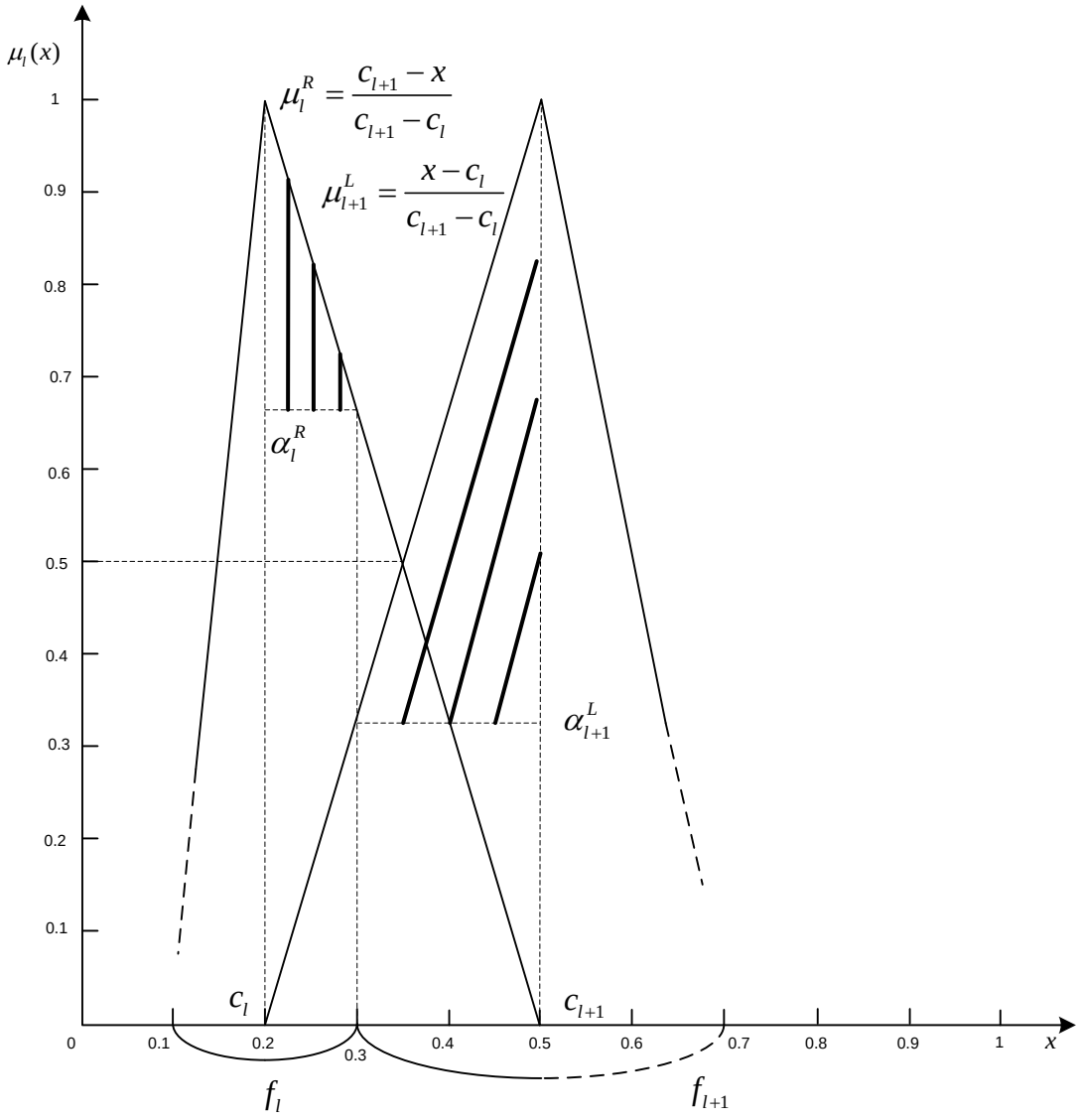



Рисунок 2.2 – Области влияния соседних рангов

## 2.5 Метод нечеткой кластеризации порядковых данных

Поскольку кластеризации подлежит выборка многомерных наблюдений-векторов, аналогично предыдущему необходимо провести фаззификацию по каждой из координат $n$-мерного пространства признаков. При этом формируется $nm$ функций принадлежности с центрами $c_{jl}$ так, как это показано на двумерном

примере (рис. 2.3). Здесь же приведен объект $x(k)$, подлежащий кластеризации, с координатами

$$x(k) = \begin{pmatrix} x_1^2(k) \equiv a \\ x_2^4(k) \equiv e \end{pmatrix},$$

при этом срабатывают функции принадлежности $\mu_{12}(x_1), \mu_{13}(x_1), \mu_{14}(x_1)$, $\mu_{22}(x_2), \mu_{23}(x_2), \mu_{24}(x_2)$ так, что немедленно принять решение о принадлежности $x(k)$ к одному из классов «плохо», «удовлетворительно», «хорошо», «отлично», весьма затруднительно (рис. 2.3).

Процесс нечеткой кластеризации порядковых переменных проведем на этом же примере (рис. 2.4). После формирования $mn$ функций принадлежности (в примере $2 \times 4 = 8$), в рассмотрение вводятся $n$ - мерные векторы - центроиды кластеров $c_i = (c_{1l}, c_{2l}, ..., c_{nl})^T, i = 1,2,...m; l = 1,2,...,m$ (в примере $c_1 = (c_{11}, c_{12})^T$, $c_2 = (c_{21}, c_{22})^T, c_3 = (c_{31}, c_{32})^T, c_4 = (c_{41}, c_{42})^T$) со своими областями влияния, описываемыми соотношениями (2.14). При попадании в область влияния, можно говорить о четкой принадлежности объекта $x(k)$ конкретному кластеру. Здесь же приведен классифицируемый объект $x(k) = (e, a)^T$, который после фаззификации представлен в числовой форме с координатами $c_{12}$ и $c_{24}$.

Далее вычисляются расстояния между $x(k)$ и всеми центроидами $c_i$ $d\left(x(k), c_i\right) = \| x(k) - c_i \|$, после чего уровни принадлежности $u_i(k)$ вектора $x(k)$ $i$ - му кластеру можно определить согласно FCM алгоритму [23] в виде

$$u_i(k) = \frac{\| x(k) - c_i \|^{-2}}{\sum_{t=1}^{m} \| x(k) - c_t \|^{-2}} = \frac{d^{-2}\left(x(k), c_i\right)}{\sum_{t=1}^{m} d^{-2}\left(x(k), c_t\right)}. \qquad (2.15)$$

Недостатком оценки (2.15) является то, что в результате (кроме случаев, когда $x(k)$ попадает в область влияния одного из центроидов) объект

«размазывается» по всем существующим кластерам, что в порядковой шкале ведет к потере физического смысла. Так рассматриваемый объект $x(k)=(e,a)^T$ с ненулевым уровнем принадлежности может относиться как к кластеру «отлично», так и к кластеру «плохо», что, конечно же, бессмысленно.

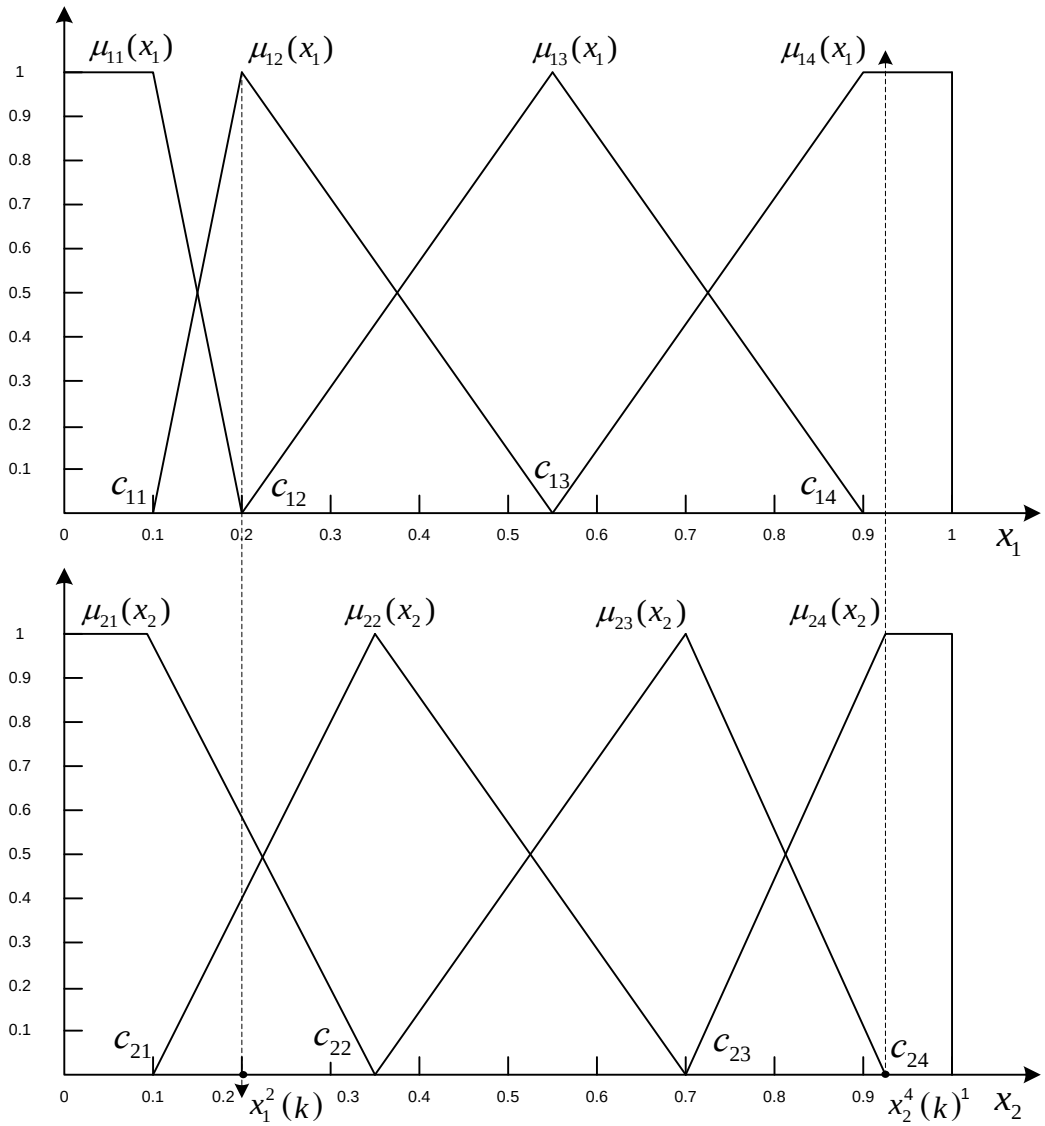



Рисунок 2.3 – Функции принадлежности двумерного пространства признаков

В связи с этим представляется целесообразным после вычисления всех расстояний $d\left(x(k),c_i\right)$ провести их ранжирование по возрастанию и выбрать наименьшее $d_{\min\min}\left(x(k),c_i\right)$ и следующее за ним $d_{\min}\left(x(k),c_i\right)$. Далее можно воспользоваться формулой (2.15) с той разницей, что в расчет принимаются только

два наименьших расстояния. В результате $x(k)$ будет принадлежать двум соседним кластерам с центроидами $c_i$ и $c_{i+1}$ (или $c_{i-1}$) с некоторыми уровнями принадлежности $u_i(k)$ и $u_{i+1}(k)$ (или $u_{i-1}(k)$).

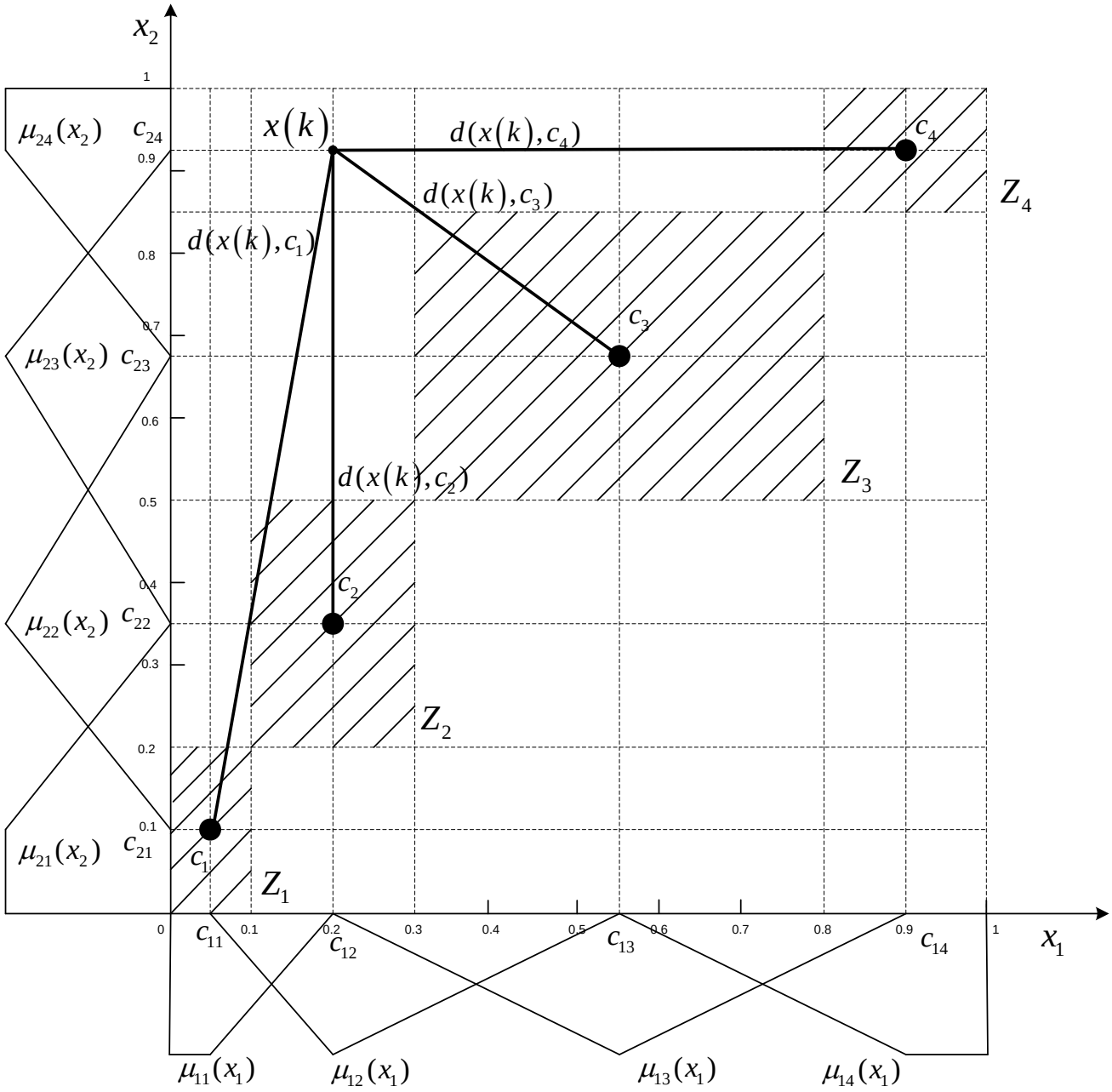



Рисунок 2.4 – Нечеткая кластеризация ранговых переменных

Таким образом, метод нечеткой кластеризации многомерных наблюдений, заданных в порядковой шкале, реализуется в виде последовательности следующих шагов.

Шаг 1. Вычисление относительных $f_l$ и накопленных $F_l$ частот по выборке $x(1), x(2), ..., x(k), ..., x(N)$.

Шаг 2. Фаззификация исходной выборки лингвистических переменных построением $mn$ функций принадлежности $\mu_{jl}(x_j), l = 1,2,...m;\ j = 1,2,...n$ и $m$ векторов – центроидов $c_i = (c_{1l}, c_{2l}, ..., c_{nm})^T$ формируемых кластеров.

Шаг 3. Построение областей влияния $Z_i$ центроидов $c_i$ в виде ортотопа с ребрами $c_{jl} \pm 0,5 f_{jl}$.

Шаг 4. Проверка возможности четкой кластеризации в виде: если $x(k) \in Z_i$, то данное наблюдение однозначно классифицируется, т.е. $u_i(k) = 1$ и $u_i(k) = 0$.

Шаг 5. Если предыдущее условие не выполняется, производится расчет всех расстояний $d\left(x(k), c_i\right) = \| x(k) - c_i \|$.

Шаг 6. Выделение двух наименьших расстояний $d_{\min\min}\left(x(k), c_i\right)$ и $d_{\min}\left(x(k), c_j\right)$, где $j$ может принимать значение или $i-1$, или $i+1$.

Шаг 7. Расчет уровней принадлежности $x(k)$ к двум соседним кластерам в виде:

$$u_i(k) = \frac{d^{-2}_{\min\min}\left(x(k), c_i\right)}{d^{-2}_{\min\min}\left(x(k), c_i\right) + d^{-2}_{\min}\left(x(k), c_i\right)},$$

$$u_j(k) = \frac{d^{-2}_{\min}\left(x(k), c_j\right)}{d^{-2}_{\min\min}\left(x(k), c_j\right) + d^{-2}_{\min}\left(x(k), c_j\right)}.$$

## 2.6 Отображение порядковых переменных в числовую шкалу

В основе классификации объектов по прототипам лежит сравнение их схожести (similarity). Схожесть между двумя объектами одного типа представляет собой агрегацию схожести между параметрами этих двух объектов. В случае если данные представлены в числовой шкале, схожесть между двумя наблюдениями может быть представлена как расстояние между ними. Схожесть между объектами

номинальной шкалы может быть выражена как 0 или 1, в зависимости от совпадения или несовпадения сравниваемых объектов. Данные, представленные в порядковой шкале, имеют ряд особенностей (таких как порядок следования атрибутов), не позволяющих вычислять схожесть вышеизложенными способами.

Исходной информацией для решения задачи является выборка наблюдений, сформированная из $N$ $n$ - мерных векторов признаков $X = \{x(1), x(2), ..., x(k), ..., x(N)\}$, где $k = 1, ..., N$.

Предположим, что каждый объект в $X$ имеет одинаковый тип свойств, и одно из этих свойств представлено в порядковой шкале. Пусть $L = \{l_1, ..., l_m\}$ – это множество возможных значений данной порядковой характеристики, удовлетворяющее свойству $l_1 < l_2 < ... < l_m$.

Для каждого значения $l_s$ пусть существует подмножество объектов $X_s \subseteq X$, включающее в себя $l_s$. Подобие *(*similarity*)* между двумя значениями $l_s$ и $l_t$, отражающее матрицу $U$, определяется как среднее подобий между объектами в $X_s$ и $X_t$ соответственно [67]

$$sim(U, l_s, l_t) = sim(U, X_s, X_t), \forall s = 1...m; t = 1...m; s \neq t,$$

где подобие между двумя непересекающимися подмножествами $X$ определяется следующим образом:

$$\forall A, B \subseteq X \ э \ A \cap B \neq \varnothing,$$

$$sim(U, A, B) = \frac{\sum_{x \in A; y \in B} sim(U, x, y)}{|A||B|},$$

где $sim(U, x, y)$ – это подобие между $x \in X$ и $y \in X$.

Подобие между двумя наблюдениями $x$ и $y$ в $X$, отражающее $U$, определено контекстно - ориентированной близостью (context - based proximity) между $x$ и $y$

$$sim(U,x,y) = prox(U,x,y),$$

а контекстно-ориентированная близость между $x(j)$ и $x(k)$ в отношении $U$, которая используется в (2.24), определяется выражением

$$prox\big(U,x(j),x(k)\big) = \sum_{i=1}^{r} \min\big(u_i(j),u_i(k)\big).$$

Рассмотрим три последовательных значения $l_{s-1}$, $l_s$ и $l_{s+1}$. Когда $sim(U,l_{s-1},l_s) > sim(U,l_s,l_{s+1})$, мы можем в общем сказать, что $l_s$ ближе к $l_{s-1}$, чем к $l_{s+1}$ для всех $s = 2,...,m-1$ в данном множестве объектов $X$. Введем для удобства два дополнительных значения характеристики $l_0$ и $l_{m+1}$ такие, что $l_0 < l_1$, а $l_m < l_{m+1}$. Теперь порядково-цифровое отображение $g$ для порядковых значений характеристики в данном множестве объектов $X$ можно описать выражением

$$\begin{cases} g(l_0) = 0, \\ g(l_l) = \sum_{s=1}^{l} \dfrac{1 - sim(U,l_{s-1},l_s)}{\sum_{t=1}^{m+1}\big(1 - sim(U,l_{t-1},l_t)\big)}, \forall l = 1...m, \\ g(l_{m+1}) = 1, \end{cases} \qquad (2.16)$$

где

$$\begin{cases} sim(U, l_0, l_1) = sim(U, X_1, X - X_1), \\ sim(U, l_{l-1}, l_l) = sim(U, X_s, X_t), \forall l = 2...m, \\ sim(U, l_m, l_{m+1}) = sim(U, X_m, X - X_m). \end{cases}$$

## 2.7 Адаптивный метод нечеткой кластеризации на основе порядково-цифрового отображения

В ряде задач, таких как обработка речи, Web Mining, медицинская диагностика, обработка сигналов датчиков в робототехнике и т.п. часто необходима обработка данных в реальном времени. В связи с этим целесообразным является использование рекуррентных процедур кластеризации.

Предлагаемый метод имеет достаточно близкую алгоритмическую структуру к алгоритму «fuzzy $c$ - means» (FCM) [23]. Задача кластеризации для количественных характеристик решается путем минимизации целевой функции (2.1) при ограничениях

$$\begin{cases} u_i(k) \geq 0, \forall i = 1,..., r; \forall k = 1,..., N, \\ \sum_{i=1}^{r} u_i(k) = 1, \forall k = 1,..., N, \\ \sum_{k=1}^{N} u_i(k) > 0, \forall i = 1,..., r. \end{cases}$$

Анализируя уравнение (2.5), можно заметить, что задача поиска седловой точки Лагранжа может быть сведена к решению последовательности задач поиска седловой точки локальных модификаций функции Лагранжа. В связи с этим мы будем использовать для расчета уровня принадлежности локальную модификацию лагранжана

$$L_S\left(u_i(k), c_i, \lambda(k)\right) = \sum_{i=1}^{r} u_i^{\beta}(k) d^2\left(x(k), c_i\right) + \lambda(k) \sum_{i=1}^{r} \left(u_i(k) - 1\right),$$

где $c_i$ – прототип (центроид) $i$ - го кластера, рассчитанного для выборки данных из $k$ наблюдений;

$d\left(x(k), \mathrm{c}_i\right)$ – расстояние между прототипом $i$ - го кластера и $k$ - м наблюдением в некоторой метрике;

$u_i(k) \in [0,1]$ – уровень принадлежности вектора $x(k)$ к $i$ - му кластеру;

$\beta$ – неотрицательный параметр фаззификации (фаззификатор), определяющий размытость границ между кластерами;

$\lambda(k)$ – неопределённый множитель Лагранжа.

Используя процедуру оптимизации Эрроу - Гурвица - Удзавы, получаем алгоритм вида

$$\begin{cases} u_i(k) = \dfrac{\left(d^2\left(x(k), c_i(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{t=1}^{r}\left(d^2\left(x(k), c_t(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ c_i(k+1) = c_i(k) - \eta(k) \nabla_{c_i} L(k)\left(u_i(k), c_i(k), \lambda(k)\right) = \\ = c_i(k) - \eta(k) u_i^{\beta}(k) d\left(x(k+1), c_i(k)\right) \nabla_{c_i} d\left(x(k+1), c_i(k)\right), \end{cases} \tag{2.17}$$

где $\eta(k)$ – параметр шага обучения.

Процедура (2.17) по структуре близка к алгоритму нечеткого конкурентного обучения Чанга-Ли [68], а в случае, когда параметр фаззификации $\beta = 2$ – к градиентному алгоритму нечеткой кластеризации Парка - Дэггера [69]

$$\begin{cases} u_i(k) = \dfrac{\| x(k) - c_i(k) \|^{-2}}{\sum\limits_{t=1}^{r} \| x(k) - c_t(k) \|^{-2}}, & (2.18) \\ c_i(k+1) = c_i(k) + \eta(k) u_i^2(k) \left( x(k+1) - c_i(k) \right). & (2.19) \end{cases}$$

Используя порядково-цифровое отображение с помощью рекуррентной нечеткой кластеризации, получаем метод кластеризации порядковых данных, состоящий из последовательности шагов.

Инициализация.

Шаг 1. Инициализируем все порядково-цифровые отображения равномерными интервалами.

Шаг 2. Трансформируем данное множество объектов $X = \{x(1),...,x(N)\}$ в $X^*$, заменяя все порядковые характеристики порядково-цифровыми отображениями.

Шаг 3. Инициализируем центроиды кластеров $c_i, \forall i = 1...r$ случайными значениями.

Повтор.

Шаг 1. Вычисляются функции принадлежности $u_i(k), \forall i = 1...r; k = 1...N$ с помощью формулы (2.18).

Шаг 2. Вычисляются центроиды кластеров $c_i, \forall i = 1...r$ с помощью формулы (2.19).

Шаг 3. Для каждого порядкового значения вычисляется порядково-цифровое отображение с помощью формулы (2.16).

Шаг 4. Трансформируем данное множество объектов $X$ в $X^*$ с новыми порядково - цифровыми отображениями.

Рассмотренный алгоритм позволяет работать с данными в порядковой шкале в on-line режиме и прост в численной реализации.

## 2.8 Правдоподобие и вероятность

Существует несколько основных подходов к кластеризации данных – иерархический, метрический, итерационный и т.п. [9]. Итерационная кластеризация применяется во многих областях, при этом алгоритм в цикле находит лучшие кластеры, к которым могут принадлежать наблюдения.

Кластеризация данных в порядковой шкале сталкивается с определенными трудностями, поскольку большинство классических алгоритмов направлено на работу с данными в численном виде.

При работе с порядковыми данными чаще всего используется подход, основанный на замене лингвистических переменных их рангами. Однако в большинстве случаев этот прием оказывается некорректным, поскольку предполагает равенство расстояний между соседними числовыми рангами, что не всегда соответствует действительности.

Более естественным представляется подход, развиваемый в [70] и основанный на максимизации функции правдоподобия. Ограничением этого подхода является предположение о гауссовом распределении исходных данных, что во многих приложениях не выполняется, а также способ вычисления правдоподобия для порядковых переменных.

Рассмотрим простейший пример, в котором каждое наблюдение имеет четыре атрибута $x_1, x_2, x_3, x_4$. Предполагая, что они являются взаимно независимыми, задача итерационной кластеризации сводится к задаче нахождения кластера $cl$ путем максимизации правдоподобия $P(cl \mid x_1 x_2 x_3 x_4)$ для каждого наблюдения с характеристиками $x_1, x_2, x_3, x_4$. По формуле Байеса это правдоподобие может быть вычислено следующим образом:

$$P(cl \mid x_1 x_2 x_3 x_4) = \frac{P(x_1 x_2 x_3 x_4 \mid cl) P(cl)}{P(x_1 x_2 x_3 x_4)},$$

то есть, нахождение кластера $cl$ путем максимизации правдоподобия $P(cl \mid x_1 x_2 x_3 x_4)$ эквивалентно решению этой задачи путем максимизации условной вероятности $P(x_1 x_2 x_3 x_4 \mid cl)$. Более того, предположение о том, что характеристики взаимно независимы, позволяет записать очевидное соотношение

$$P(x_1 x_2 x_3 x_4 \mid cl) = P(x_1 \mid cl) P(x_2 \mid cl) P(x_3 \mid cl) P(x_4 \mid cl). \quad (2.20)$$

Следовательно, проблема нахождения кластера $cl$ представляет собой проблему максимизации правой части уравнения (2.20).

Таким образом, можно говорить о том, что проблема нахождения кластеров решается путем максимизации произведения индивидуальных условных вероятностей характеристик наблюдения. Заметим, что вероятность $P(x(k) \mid cl)$ выражает как часто наблюдение $x(k)$ появляется в выборке со всеми одинаковыми значениями характеристик в кластере $cl$. То есть, $P(x(k) \mid cl)$ выражает определенный вид частоты встречаемости $x(k)$ с одинаковыми значениями параметров в кластере $cl$.

## 2.9 Метод нечеткой кластеризации порядковых данных на основе совместного использования функций принадлежности и функции правдоподобия

Анализируя выражение (2.5) видим, что при вычислении уровня принадлежности $u_i(k)$ конкретного наблюдения $x(k)$ к кластеру $cl$ используется расстояние между наблюдением и соответствующими центроидом кластера $c_i$. Далее пересчитывается $c_i$ на основе уровней принадлежности к кластерам $u_i(k)$. Вычисления производятся итерационно, пока не будет выполнено условие остановки алгоритма.

Идея предлагаемого алгоритма состоит в том, чтобы использовать правдоподобия наблюдений для определения кластеров вместо расстояний в алгоритме «fuzzy $c$ - means» (FCM). Таким образом, задача решается путем максимизации целевой функции

$$Q = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) L_i(k), \tag{2.21}$$

или соответственно минимизации

$$Q = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) U_i(k), \tag{2.22}$$

при ограничениях

$$\begin{cases} u_i(k) \geq 0, \forall i = 1,...,r; \forall k = 1,...,N, \\ \sum_{i=1}^{r} u_i(k) = 1, \forall k = 1,...,N, \\ \sum_{k=1}^{N} u_i(k) > 0, \forall i = 1,...,r, \end{cases}$$

где $L_i(k)$ – правдоподобие принадлежности $k$ - го наблюдения к $i$ - му кластеру;

$U_i(k)$ – логарифм несходства $k$ - го наблюдения с $i$ - м кластером;

$u_i(k)$ – уровень принадлежности вектора $x(k)$ к $i$ - му кластеру.

Правдоподобие $L_i(k)$ в (2.21) вычисляется согласно формуле

$$L_i(k) = \prod_{j=1}^{n} p_{ij}(k), \tag{2.23}$$

где $p_{ij}(k)$ – условная вероятность появления определенного значения $j$ - ой характеристики $k$ - го наблюдения в $i$ - ом кластере.

При этом $p_{ij}(k)$ вычисляется по формуле

$$p_{ij}(k) = P\left(x_j(k) \mid cl_i\right). \tag{2.24}$$

Логарифм несходства в (2.22) определяется выражением

$$U_i(k) = -\ln L_i(k), \tag{2.25}$$

а целевую функцию (2.22) можно переписать в виде

$$Q = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) U_i(k) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k)\left(-\ln\prod_{j=1}^{n} p_{ij}(k)\right) = -\sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k)\sum_{j=1}^{n}\ln p_{ij}(k).$$

Для вычисления $u_i(k)$ применяется выражение [70]

$$u_{t,j} = \frac{1}{\sum_{i=1}^{r}\left(\frac{U_t(k)}{U_i(k)}\right)^{\frac{1}{\beta-1}}}, \forall t = 1,...,r; \forall k = 1,...,N, \tag{2.26}$$

а для подсчета условных вероятностей $p_{ij}(k)$ используются функции принадлежности, описанные ниже.

Недостатком данного подхода является то, что рассматриваемый объект «размывается» по всем существующим кластерам, что в ранговой шкале ведет к потере физического смысла. В связи с этим представляется целесообразным после вычисления центроидов, пересчитать все расстояния $d\left(x(k), \mathrm{c}_i\right)$, провести их ранжирование по возрастанию и выбрать наименьшее $d_{\min\min}\left(x(k), \mathrm{c}_i\right)$ и следующее за ним $d_{\min}\left(x(k), \mathrm{c}_i\right)$. Принимая в расчет два наименьших расстояния, можно воспользоваться приведенными ниже формулами:

$$u_i(k) = \frac{d^{-2}_{\min\min}\left(x(k), \mathrm{c}_i\right)}{d^{-2}_{\min\min}\left(x(k), \mathrm{c}_i\right) + d^{-2}_{\min}\left(x(k), \mathrm{c}_l\right)}, \tag{2.27}$$

$$u_i(l) = \frac{d^{-2}_{\min}\left(x(k), \mathrm{c}_l\right)}{d^{-2}_{\min\min}\left(x(k), \mathrm{c}_i\right) + d^{-2}_{\min}\left(x(k), \mathrm{c}_l\right)}. \tag{2.28}$$

Таким образом, метод состоит из последовательности шагов.

Шаг 1. Инициализация $p_{ij}(k), \forall i = 1,...,r; \forall k = 1,...,N; \forall j = 1,...,n$ случайными значениями.

Шаг 2. Подсчет $u_i(k), \forall i = 1,...,r; \forall k = 1,...,N$ с помощью формулы (2.26).

Шаг 3. Подсчет $p_{ij}(k), \forall i = 1,...,r; \forall k = 1,...,N; \forall j = 1,...,n$ с помощью формулы (2.32).

Шаг 4. Шаги 2 и 3 повторяется итерационно до выполнения условия

$$\varepsilon \le \max_i(k)\left\{\left|old\_\mu_i(k) - new\_\mu_i(k)\right|\right\}.$$

Шаг 5. Расчет всех расстояний $d\left(x(k), c_i\right) = \| x(k) - c_i \|$ и выделение двух наименьших расстояний $d_{\min\min}\left(x(k), \mathrm{c}_i\right)$ и $d_{\min}\left(x(k), \mathrm{c}_l\right)$, где $l$ может принимать значение или $i-1$, или $i+1$.

Шаг 6. Расчет уровней принадлежности $x(k)$ к двум соседним кластерам по формулам (2.27) и (2.28).

2.9.1 Вычисление условной вероятности $p_{i,j}(k)$ и фаззификация исходных данных

Процесс фаззификации последовательности порядковых лингвистических переменных рассмотрим на примере одномерной выборки $x(1),...,x(N)$, где каждому из наблюдений может быть приписан один из рангов $l$, $l = 1,...,m.$

Пусть значение $x(k)$, соответствующее $l$ - му рангу, встречается в выборке $N_l$ раз. Тогда в рассмотрение вводятся относительные частоты появления $l$ - го ранга, вычисляемые по формуле (2.7), при этом, естественно, выполняется условие

$$\sum_{l=1}^{m} f_l = 1.$$

На основе относительных частот формируются усредненные частоты встречаемости наблюдений, при этом для их вычисления удобно воспользоваться рекуррентным соотношением (2.12)

$$F_1 = 0.5 f_1,\ F_l = F_{l-1} + 0.5\left( f_{l-1} + f_l \right), \forall l = 2,...,m.$$

Далее все порядковые данные заменяются соответствующими усредненными частотами встречаемости наблюдений. Этап фаззификации представлен в виде гистограммы на рисунке 2.5.

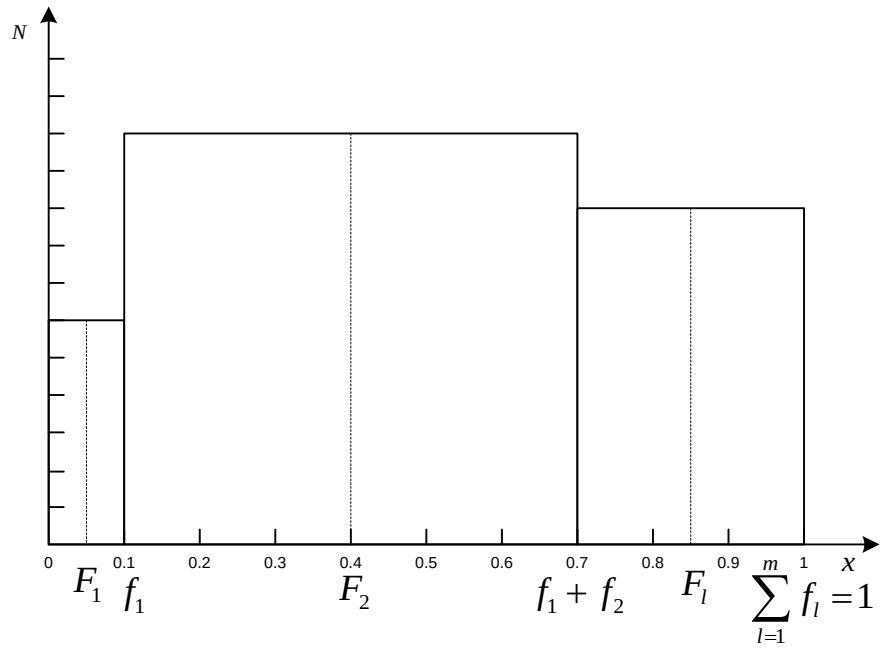



Рисунок 2.5 – Гистограмма распределения порядковых переменных по частоте встречаемости в выборке

Предполагая, что уровень принадлежности наблюдений к кластерам $u_i(k), \forall i = 1,...,r; \forall k = 1,...,N$ известен, вычисляется мода для каждой характеристики по каждому из кластеров $x_{ij}^*, \forall i = 1,...,r; \forall j = 1,...,n$.

Далее, учитывая полученные моды, строятся ассиметричные функции принадлежности.

1. Если $x_{ij}^* > 0.5$, то функция принадлежности имеет вид, представленный на рисунке 2.6 и описываемый формулой

$$\mu_{ij}(k) = \begin{cases} \dfrac{x(k)}{x_{ij}^*}, x \in [0, x_{ij}^*], \\ \dfrac{2x_{ij}^* - x(k)}{x_{ij}^*}, x \notin [0, x_{ij}^*]. \end{cases} \quad (2.29)$$

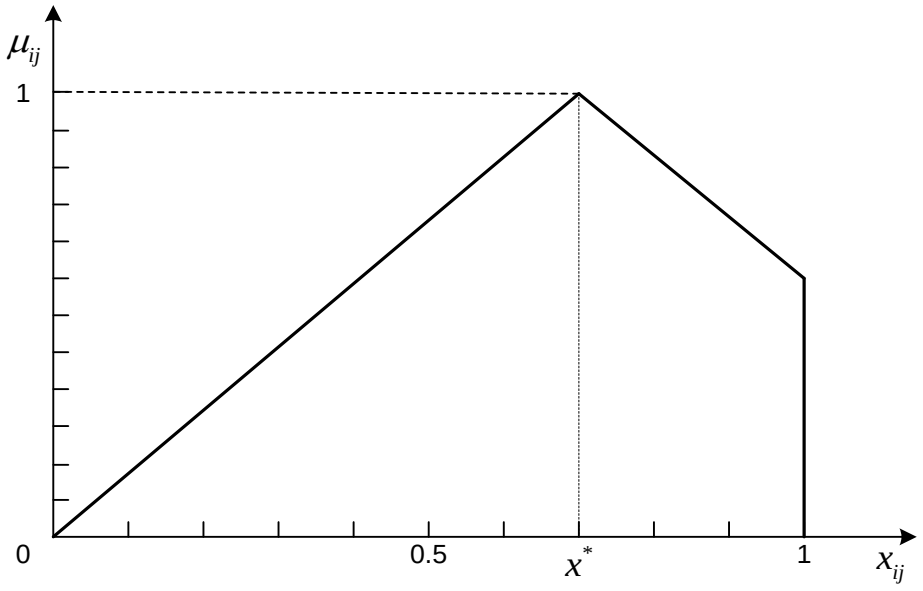



Рисунок 2.6 – Функция принадлежности для выборки ранговых переменных, когда $x_{ij}^* > 0.5$

2. Если $x_{ij}^* < 0.5$, то функция принадлежности имеет вид, представленный на рисунке 2.7 и описываемый формулой

$$\mu_{ij}(k)=\begin{cases}\dfrac{1-x(k)}{1-x_{ij}^{*}},\ x\in[x_{ij}^{*},1],\\[2ex]\dfrac{x(k)-2x_{ij}^{*}+1}{1-x_{ij}^{*}},\ x\notin[x_{ij}^{*},1].\end{cases} \tag{2.30}$$

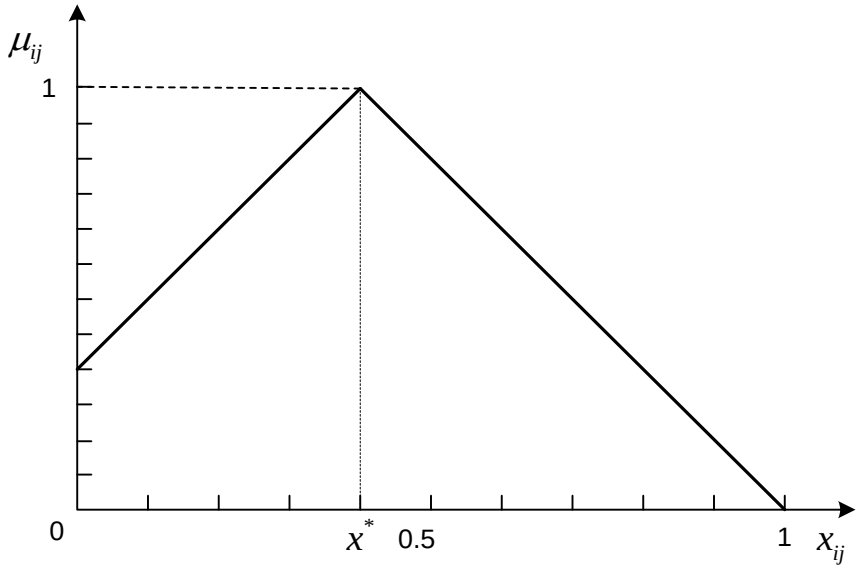



Рисунок 2.7 – Функция принадлежности для выборки ранговых переменных, когда $x_{ij}^{*}<0.5$

3. Если $x_{ij}^{*}=0.5$, то функция принадлежности имеет вид, представленный на рисунке 2.8 и описываемый формулой

$$\mu_{ij}(k)=\begin{cases}\dfrac{x(k)}{x_{ij}^{*}},\ x\in[0,x_{ij}^{*}],\\[2ex]\dfrac{1-x(k)}{1-x_{ij}^{*}},\ x\in[x_{ij}^{*},1].\end{cases} \tag{2.31}$$

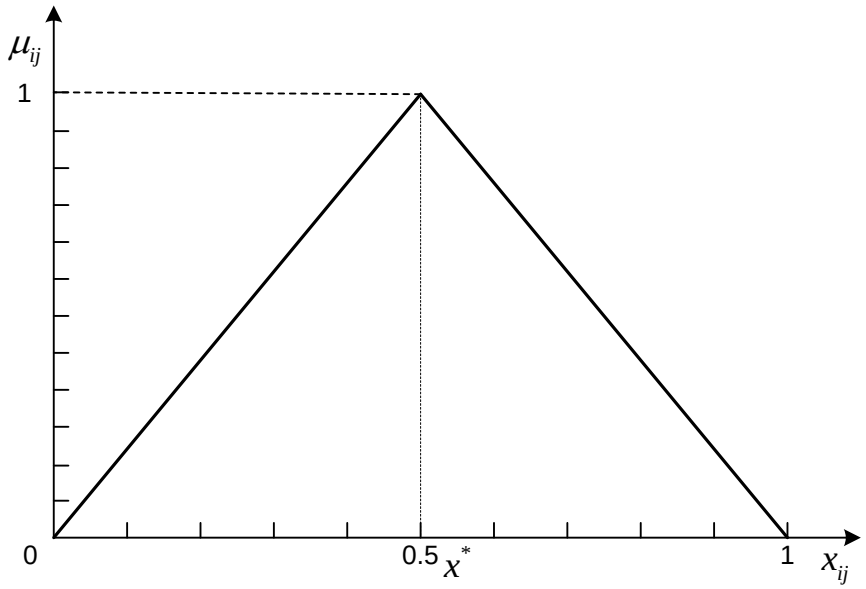



Рисунок 2.8 – Функция принадлежности для выборки ранговых переменных, когда $x_{ij}^* = 0.5$

Поскольку условная вероятность $p_{ij}(k)$ напрямую зависит от частоты встречаемости конкретного значения характеристики в выборке, а данные идут в четко заданном порядке от самого малого к самому большому, то можно сказать, что

$$p_{ij}(k) = \mu_{ij}(k). \tag{2.32}$$

Одним из преимуществ данного подхода является его устойчивость к выбросам благодаря использованию порядка следования переменных при построении функций принадлежности.

## 2.10 Нечеткая робастная кластеризация данных на основе меры схожести

В реальных задачах при анализе и обработке данных часто возникает проблема, суть которой заключается в зашумленности экспериментальных данных, т.е. наличии выбросов. Использование классических алгоритмов кластеризации при работе с такими данными показало значительное смещение прототипов и радиусов кластеров. Возникла необходимость в развитии методов,

нечувствительных к наличию выбросов в данных. Решением этой проблемы стали «робастные» процедуры обработки данных [71-76].

Робастность предполагает нечувствительность модели к отклонениям от априорных предположений о характере экспериментальных данных.

Для обработки данных с большим разбросом были разработаны адаптивные методы кластеризации, представленные в работах [77–80]. В них было предложено вместо метрики в целевой функции (2.1) использовать критерий близости, поскольку он убывает медленнее, чем квадрат евклидового расстояния. Наиболее распространенными критериями близости являются функция Грина, модуль евклидового расстояния, функция Коши [71], сравнение которых приведено на рисунке 2.9.

С содержательной точки зрения более удобным представляется использование вместо целевых функций так называемых «мер схожести» (SM) [81], к которым предъявляются более мягкие, чем для метрик, условия (отсутствует неравенство треугольника)

$$\begin{cases} S\left(\tilde{x}(k),\tilde{x}(p)\right)\geq 0, \\ S\left(\tilde{x}(k),\tilde{x}(p)\right)=S\left(\tilde{x}(p),\tilde{x}(k)\right), \\ S\left(\tilde{x}(k),\tilde{x}(k)\right)=1\geq S\left(\tilde{x}(k),\tilde{x}(p)\right), \end{cases}$$

а задача кластеризации может рассматриваться как максимизация какой-либо из этих мер.

Рисунок 2.10 иллюстрирует использование в качестве меры схожести традиционного гауссиана с разными параметрами ширины $\sigma^2<1$.

Введем в рассмотрение функцию меры схожести

$$S\left(\tilde{x}(k),c_i\right)=e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}}=e^{-\frac{d^2\left(\tilde{x}(k),c_i\right)}{2\sigma^2}}, \qquad (2.33)$$

где $c_i$ – вектор координат центроида (прототипа) $i$ - го кластера.

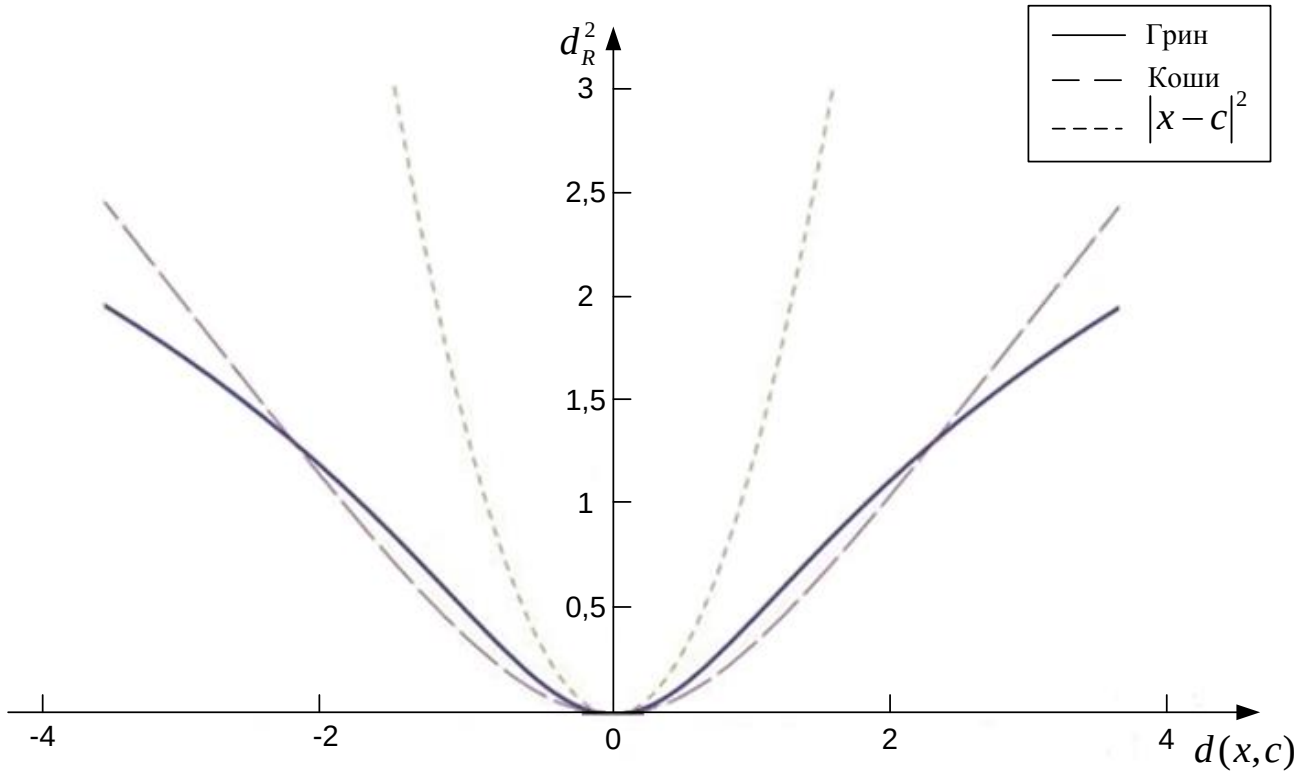



Рисунок 2.9 – Сравнение функции Грина, функции Коши и модуля евклидового расстояния

Подбирая параметр ширины $\sigma^2$ функции (2.33), можно подавить влияние далеко отстоящих от центроида наблюдений, что в принципе невозможно сделать с помощью традиционной евклидовой метрики

$$d^2\left(\tilde{x}(k), c_i\right) = \left\|\tilde{x}(k) - c_i\right\|^2.$$

Введем в рассмотрение целевую функцию, основанную на мере схожести (2.33)

$$E_s\left(u_i(k), c_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) S\left(\tilde{x}(k), c_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}},$$

где $\beta > 0$ – фаззификатор, используемый в теории нечеткой кластеризации [9,10].

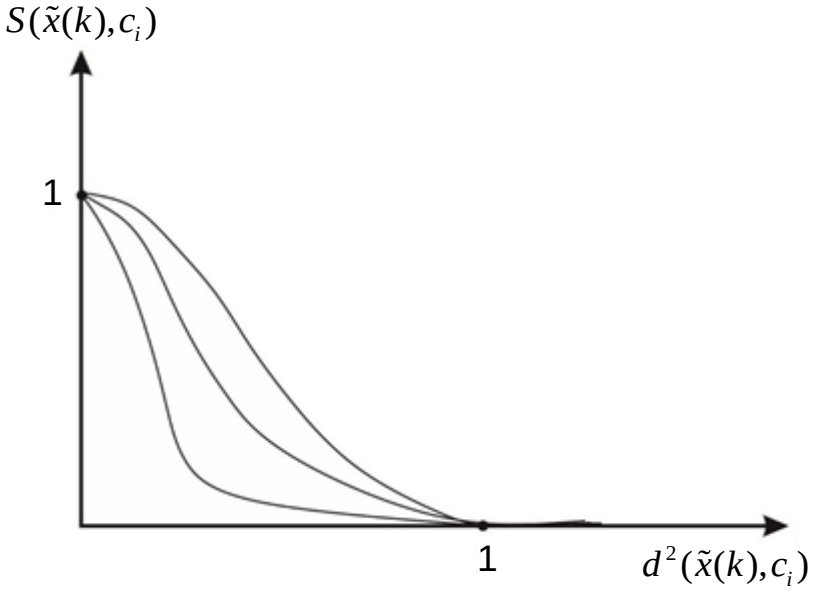



Рисунок 2.10 – Гауссиан в качестве меры схожести

При стандартных вероятностных ограничениях получаем функцию Лагранжа

$$L_S\left(u_i(k),c_i,\lambda_i\right)=\sum_{k=1}^{N}\sum_{i=1}^{r}u_i^{\beta}(k)e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}}+\sum_{k=1}^{N}\lambda(k)\left(\sum_{i=1}^{r}u_i(k)-1\right), \tag{2.34}$$

где $\lambda(k)$ – неопределенные множители Лагранжа.

Решая систему уравнений Каруша - Куна - Таккера, приходим к решению

$$\begin{cases} u_i(k)=\dfrac{S\left(\tilde{x}(k),c_i\right)^{\frac{1}{\beta-1}}}{\sum\limits_{t=1}^{r}S\left(\tilde{x}(k),c_t\right)^{\frac{1}{\beta-1}}}, \\ \lambda(k)=-\left(\sum\limits_{t=1}^{r}\beta S\left(\tilde{x}(k),c_t\right)^{\frac{1}{\beta-1}}\right)^{\beta-1}, \\ \nabla_{c_i}L_S\left(u_i(k),c_i,\lambda(k)\right)=\sum\limits_{k=1}^{N}w_i^{\beta}(k)e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}}\cdot\dfrac{\tilde{x}(k)-c_i}{\sigma^2}=\vec{0}. \end{cases} \tag{2.35}$$

Первые два уравнения системы (2.35) имеют аналитическое решение, тогда как третье такого решения явно не имеет, а потому для нахождения седловой точки Лагранжа (2.34) можно воспользоваться процедурой Эрроу - Гурвица - Удзавы, в результате применения которой приходим к алгоритму

$$\begin{cases} u_i(k+1) = \dfrac{S\left(\tilde{x}(k+1), c_i(k)\right)^{\frac{1}{\beta-1}}}{\sum\limits_{t=1}^{r} S\left(\tilde{x}(k+1), c_t(k)\right)^{\frac{1}{\beta-1}}}, \\ c_i(k+1) = c_i(k) + \eta(k+1)\nabla_{c_i} L_S\left(u_i(k+1), c_i(k)\right) = \\ = c_i(k) + \eta(k+1) u_i^{\beta}(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}, \end{cases}$$

где $\eta(k+1)$ – параметр шага обучения.

Полагая значение фаззификатора $\beta = 2$, получаем робастный вариант нечетких $c$ - средних, основанному на мере схожести

$$\begin{cases} u_i(k+1) = \dfrac{S\left(\tilde{x}(k+1), c_i(k)\right)}{\sum\limits_{t=1}^{r} S\left(\tilde{x}(k+1), c_t(k)\right)}, \\ c_i(k+1) = c_i(k) + \eta(k+1) u_i^2(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}. \end{cases}$$

Используя далее концепцию ускоренного машинного времени, можно ввести адаптивную робастную процедуру нечеткой кластеризации вида

$$\begin{cases} w_i^{(\tau+1)}(k+1) = \dfrac{S\left(\tilde{x}(k), c_i^{(\tau)}(k)\right)^{\frac{1}{\beta-1}}}{\sum\limits_{t=1}^{r} S\left(\tilde{x}(k), c_t^{(\tau)}(k)\right)^{\frac{1}{\beta-1}}}, \\ c_i^{(Q)}(k) = c_i^{(0)}(k+1), \\ c_i^{(\tau+1)}(k+1) = c_i^{(\tau)}(k+1) + \eta(k+1)\left(u_i^{(Q)}(k)\right)^{\beta} e^{-\frac{\left\|\tilde{x}(k+1)-c_i^{(\tau)}(k+1)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1)-c_i^{(\tau)}(k+1)}{\sigma^2}, \end{cases}$$

при этом $\tau = 0,1,\ldots,Q$ – ускоренное машинное время такое, что между поступлениями двух соседних наблюдений $\tilde{x}(k)$ и $\tilde{x}(k+1)$ производится $Q$ итераций машинного времени.

Решение о принадлежности каждого $\tilde{x}(k)$ к конкретному кластеру $c_i$ принимается по максимальному значению меры схожести.

## 2.11 Вариант возможностной нечеткой робастной кластеризации порядковых данных

Аналогичным образом может быть синтезирован алгоритм робастной возможностной [82, 91] нечеткой кластеризации по критерию

$$E_s\left(u_i(k), c_i, \mu_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) S\left(\tilde{x}(k), c_i\right) + \sum_{i=1}^{r} \mu_i \left(1-u_i(k)\right)^{\beta}, \tag{2.36}$$

где параметр $\mu_i \geq 0$ определяет расстояние, на котором уровень принадлежности принимает значение 0.5, т.е. если

$$\left\|\tilde{x}(k) - c_i\right\|^2 = \mu_i,$$

то

$$u_i(k) = 0.5.$$

Решая задачу максимизации (2.36), получаем

$$\begin{cases} u_i(k+1) = \left(1 + \left(\dfrac{S^{-1}\left(\tilde{x}(k+1), c_i(k)\right)}{\mu_i(k)}\right)\right)^{-1}, \\ c_i(k+1) = c_i(k) + \eta(k+1) u_i^{\beta}(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}, \\ \mu_i(k+1) = \dfrac{\sum\limits_{t=1}^{k+1} u_i^{\beta}(t) S^{-1}\left(c_i(k+1), \tilde{x}(t)\right)}{\sum\limits_{t=1}^{k+1} u_i^{\beta}(t)}, \end{cases}$$

а при $\beta = 2$

$$\begin{cases} u_i(k+1) = \dfrac{1}{1 + \dfrac{S^{-1}\left(\tilde{x}(k+1), c_i(k)\right)}{\mu_i(k)}}, \\ c_i(k+1) = c_i(k) + \eta(k+1) u_i^{2}(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}, \\ \mu_i(k+1) = \dfrac{\sum\limits_{t=1}^{k+1} u_i^{2}(t) S^{-1}\left(\tilde{x}(t), c_i(k+1)\right)}{\sum\limits_{t=1}^{k+1} u_i^{2}(t)}. \end{cases}$$

Вводя ускоренное время, получаем процедуру

$$
\begin{cases}
u_i^{(\tau+1)}(k) = \dfrac{1}{1+\left(\dfrac{S^{-1}\left(\tilde{x}(k), c_i^{(\tau)}(k)\right)}{\mu_i^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \\
c_i^{(Q)}(k) = c_i^{(0)}(k+1), \\
c_i^{(\tau+1)}(k+1) = c_i^{(\tau)}(k+1) + \eta(k+1)\left(u_i^{(Q)}(k)\right)^{\beta} e^{-\frac{\left\|\tilde{x}(k+1)-c_i^{(\tau)}(k+1)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1)-c_i^{(\tau)}(k+1)}{\sigma^2}, \\
\mu_i^{(\tau+1)}(k) = \dfrac{\sum\limits_{t=1}^{k}\left(u_i^{(\tau+1)}(t)\right)^{\beta} S^{-1}\left(\tilde{x}(t), c_i^{(\tau+1)}(k)\right)}{\sum\limits_{t=1}^{k}\left(u_i^{(\tau+1)}(t)\right)^{\beta}}.
\end{cases}
$$

Введенная группа робастных алгоритмов нечеткой кластеризации, основанных на мере схожести, предназначена для обработки многомерных наблюдений, заданных в порядковой шкале. В основе подхода лежит отображение лингвистических переменных в числовую шкалу и модификация известного метода нечетких $c$ - средних, подавляющая аномальные выбросы. Устойчивость метода достигается благодаря использованию так называемых «мер схожести». Регулирование параметр ширины $\sigma^2$ функции меры схожести, позволяет подавить влияние далеко отстоящих от центроида наблюдений.

Применение возможностного подхода позволило анализировать порядковые данные, содержащие выбросы, в условиях перекрывающихся кластеров в on-line режиме.

Предложенные методы позволяют преодолеть ряд недостатков, присущих классическим алгоритма. Они просты в численной реализации, являясь по сути градиентными процедурами оптимизации целевых функций специального вида.

Выводы по разделу 2

1. Предложен метод нечеткой кластеризации данных, заданных в порядковой шкале, с фаззификацией исходных данных на основе частоты встречаемости характеристик в выборке. Данный подход позволяет повысить точность кластеризации и обрабатывать данные, не подчиняющиеся нормальному закону распределения.

2. Предложен адаптивный метод рекуррентной нечеткой кластеризации данных, заданных в порядковой шкале, на основе их отображения в числовую шкалу, что позволило обрабатывать порядковые данные в on-line режиме. Данный метод может быть использован для решения задач интеллектуального анализа данных, представленных как порядковыми характеристиками, так и смешанными.

3. Получил развитие метод нечеткой кластеризации порядковых данных путем совместного использования функций принадлежности и функции правдоподобия, что позволило обрабатывать данные, не связанные с гипотезой нормальности распределения. Метод фаззификации порядковых данных и способ определения условной вероятности появления конкретных наблюдений в каждом кластере позволяют быстро и точно кластеризовать выборку.

4. Усовершенствованы методы робастной нечеткой кластеризации порядковых данных путём введения критерия специального вида, подавляющего выбросы. Это позволило улучшить качество кластеризации порядковых данных, содержащих выбросы, в on-line режиме.

# РАЗДЕЛ 3
# МЕТОДЫ КЛАСТЕРИЗАЦИИ КАТЕГОРИАЛЬНЫХ ДАННЫХ

Во многих практических задачах, возникающих в Web Mining, Text Mining, Medical Data Mining и т.п., достаточно часто возникает ситуация, когда признаки $x_j(k)$ заданы не в числовой, а в категориальной (номинальной) шкале, при этом каждый такой признак может принимать конечное значение «имен» $x_j^l(k)$, где $j = 1,2,...,n; l = 1,2,...,\mathrm{m}_j; k = 1,2,...,N.$

Примером таких данных могут служить покупки в супермаркете, где каждое наблюдение отражает покупки отдельно взятого клиента. В таком случае данные будут иметь вид: «молоко», «хлеб», «масло», «вино» и т.п. Поскольку, номенклатура товаров в супермаркете очень разнообразна, то и выборки подобных данных отличаются большими размерами.

Понятно, что в этой ситуации традиционные методы не работают в силу отсутствия самого понятия «расстояние» в категориальной шкале.

В принципе, данные, описываемые в номинальной шкале, без особых проблем могут быть трансформированы в бинарную шкалу, однако, при этом резко возрастает размерность пространства признаков, что существенно усложняет решение задачи из-за возникновения эффектов «проклятья размерности», а в нечетком случае – «концентрации норм». Существует необходимость синтеза адаптивных алгоритмов обработки данных, заданных в категориальной шкале.

## 3.1 Метод робастной кластеризации категориальных данных (ROCK)

Рассматриваемый метод ROCK (Robust Clustering using Links) [83] является наиболее распространенным иерархическим методом.

В кластеризации при определении степени соседства объектов существенную роль играет используемая функция расстояния. Часто для определения меры близости между наблюдениями используется функция евклидова расстояния (1.4). Однако эта метрика имеет ряд недостатков при работе

с категориальными данными. Главным из них является неправильный учет атрибутов, отсутствующих у одного объекта и имеющихся у другого.

Более простой мерой близости, по сравнению с евклидовым расстояние, является коэффициент Жаккарда [84]. Схожесть двух объектов по этому коэффициенту вычисляется с помощью преобразования всех их атрибутов в точки двух множеств. То есть, коэффициент Жаккарда – это отношение пересечения подобных множеств к их объединению. Однако, в случае плохо разделяемых категориальных данных данный коэффициент не работает.

В методе ROCK мерой близости выступает новый параметр, который описывает количество общих соседей (ссылок) у каждой пары объектов.

В случае, если объекты в достаточной степени близки, они называются соседями рассматриваемого объекта

$$sim\left(x_q, x_t\right) \geq \theta, \tag{3.1}$$

то есть две точки $x_q$ и $x_t$ считаются соседями, если значение их сходства превышает заданный порог $\theta$.

Связей между двумя объектами столько, сколько общих соседей у этих объектов. Функция связи $link\left(x_q, x_t\right)$ между двумя точками $x_q$ и $x_t$ вычисляется согласно количеству общих соседей этих точек.

Две точки принадлежат одному кластеру, если у них большое значение функции связи, и разным в противном случае.

При кластеризации целевая функция будет иметь следующий вид:

$$E = \sum_{i=1}^{r} N_i \sum_{x_q, x_t \in cl_i} \frac{link\left(x_q, x_t\right)}{N_i^{1+2f(\theta)}}, \tag{3.2}$$

где $cl_i - i$ - тый кластер;

$N_i$ – его размер.

Чтобы распределить по разным кластерам те точки, которые имеют мало связей между собой, необходимо действительную сумму связей в кластере разделить на ожидаемую сумму связей – $N_i^{1+2f(\theta)}$.

Обычно в качестве функции $f$ используется $\frac{1-\theta}{1+\theta}$. Данный подход не позволяет относить точки с низким значением связи к одному кластеру. Значение связи между кластерами вычисляется по формуле

$$link\left[cl_i, cl_j\right] = \sum_{x_q \in cl_i, x_t \in cl_j} link\left(x_q, x_t\right). \tag{3.3}$$

Целевая функция качества, используемая при выборе кластеров для объединения, имеет вид

$$g\left(cl_i, cl_j\right) = \frac{link\left[cl_i, cl_j\right]}{\left(N_i + N_j\right)^{1+2f(\theta)} - N_i^{1+2f(\theta)} - N_j^{1+2f(\theta)}}. \tag{3.4}$$

Максимальное значение данной функции для двух кластеров говорит о том, что они наиболее всего подходят для объединения.

Чтобы предотвратить притягивание крупными кластерами, которые, обычно, имеют большое количество связей, более мелких, значение связи между кластерами в этой функции делится на ожидаемое значение связи

$$\left(N_i + N_j\right)^{1+2f(\theta)} - N_i^{1+2f(\theta)} - N_j^{1+2f(\theta)}.$$

Данный метод нечувствителен к выбросам и не требует четкого разделения объектов на кластеры.

Метод предназначен для кластеризации данных с большим количеством числовых и номинальных атрибутов.

Основным недостатком метода ROCK является увеличение вычислительной сложности, так как процесс подсчета ссылок – самый долгий во всей кластеризации. Также, данный метод неприменим в ситуации, когда наблюдение с разной степенью принадлежности относится сразу к нескольким кластерам.

## 3.2 Кластеризация категориальных данных методом $k$ - modes

Для решения проблемы, создаваемой «проклятьем размерности», при замене категориальных данных бинарными числами, было предложено несколько методов. В [85-87] предлагается вместо традиционного евклидова расстояния, лежащего в основе классического метода $k$ - средних, использовать «несходство» (dissimilarity) между векторами-образами, а вместо стандартных средних – моды отдельных признаков.

При этом несходство между двумя векторами $x(k)$ и $x(q)$ может быть описано с помощью выражения

$$d\left(x(k),x(q)\right)=\sum_{j=1}^{n}\delta\left(x_j(k),x_j(q)\right), \tag{3.5}$$

где

$$\delta\left(x_j(k),x_j(q)\right)=\begin{cases}0, if\ x_j(k)=x_j(q),\\ 1, if\ x_j(k)\neq x_j(q),\end{cases}$$

при этом, если $x(k)=x(q)$, то $d\left(x(k),x(q)\right)=0$, а при полном несовпадении компонент этих векторов $d\left(x(k),x(q)\right)=n$, т.е. $0\le d\left(x(k),x(q)\right)\le n$.

В этом случае в качестве прототипов центроидов кластеров используются наиболее часто встречающиеся в данном кластере значения компонент наблюдений – моды.

И хотя метод $k$ - мод ($k$ - modes), являясь «ближайшим родственником» $k$ - средних, нагляден и прост в численной реализации, его использование ограничивается тем фактом, что мода каждого из кластеров не единственна, что не позволяет получить устойчивое решение.

### 3.3 Метод $k$ - средних для кластеризации категориальных данных

Для преодоления отмеченного недостатка в [87] в качестве прототипов кластеров категориальных данных предложено использовать не обычные моды, а так называемые «представители» (representatives), учитывающие значения частот появления отдельных значений признаков.

Пусть $i$ - й кластер содержит $N_i$ наблюдений $x(k)$ так, что $cl_i = \{x(1), x(2), ..., x(N_i)\} \subset R^n$, $\sum_{i=1}^{r} N_i = N$. При этом вектор-прототип этого кластера может быть представлен в виде $c_i = (c_{i1}, c_{i2}, ..., c_{in})^T$, а для каждой компоненты $c_{ij}$ может быть рассчитана частота появления соответствующего значения признака в кластере в виде

$$f_{ij} = \frac{N_{ij}}{N_i}, \tag{3.6}$$

где $N_{ij}$ – число появлений признака $x_j$ в $cl_i$.

В связи с тем, что каждый признак $x_j$ может принимать только конечное число значений $x_j^l, l = 1, 2, ..., m_j$, выражение (3.6) можно также переписать в виде

$$f_{ij}^{l} = \frac{N_{ij}^{l}}{N_i}.$$

Тогда в качестве меры несходства между прототипом $c_i$ и наблюдением $x(k)$ вместо (3.5) используется оценка

$$d\left(x(k),c_{ij}\right)=\sum_{j=1}^{n}\sum_{l_j=1}^{m_j} f_{ij}^{l}\delta\left(x_j(k),c_{ij}\right). \tag{3.7}$$

Понятно, что (3.7) также лежит в интервале $0\le d\left(x(k),c_{ij}\right)\le n$.

Авторами [87] показано, что использование меры несходства (3.7) позволяет максимально приблизить задачу кластеризации категориальных данных к нахождению стандартных $k$ - средних путем минимизации целевой функции

$$E\left(u_i(k),c_i\right)=\sum_{k=1}^{N}\sum_{i=1}^{r}u_i(k)d^2\left(x(k),c_i\right), \tag{3.8}$$

где

$$\begin{cases}\sum_{i=1}^{r}u_i(k)=1,\\ u_i(k)\in\{0,1\}.\end{cases}$$

При этом если $x(k)$ принадлежит $cl_i$, то $u_i(k)=1$ и $u_i(k)=0$ в противоположном случае.

Собственно, процесс кластеризации реализуется в форме последовательности шагов.

Шаг 1. Некоторым достаточно произвольным образом задаются $r$ начальных прототипов $c_i$, $i=1,2,...,r.$

Шаг 2. Приписать наблюдение $x(k)$ к $cl_i$, если $d\left(x(k),c_i\right)<d\left(x(t),c_t\right)$, $\forall t=1,2,...,r;t\ne i$.

Шаг 3. Рассчитать моды-прототипы для всех кластеров $cl_i$ и соответствующие им частоты $f_{ij}^{l}$.

Шаг 4. Рассчитать $N_r$ оценок несходства новых прототипов со всеми $x(k)$.

Шаг 5. Продолжать до тех пор, пока не стабилизируются прототипы.

Дополнительно к мере несходства (3.7) можно ввести в рассмотрение оценку «сходства» (similarity) в виде

$$0 \le sim\left(x(k), c_i\right) = 1 - \frac{d\left(x(k), c_i\right)}{n} \le 1. \tag{3.9}$$

Это значение может служить простейшей оценкой уровня нечеткой принадлежности в случае возможного перекрытия формируемых кластеров, т.е. $sim\left(x(k), c_i\right) = u_i(k)$.

## 3.4 Метод нечеткой кластеризации категориальных данных

В теории и практике нечеткой кластеризации количественных переменных наибольшее распространение получил метод нечетких $c$ - средних (FCM) Дж. Бездека [8], основанный на минимизации целевой функции (2.1).

Минимизация (2.1) при ограничениях (2.2) с помощью стандартной техники нелинейного программирования приводит к известному результату

$$\begin{cases} c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}, \\ u_i(k) = \dfrac{\left(\left\| x(k) - c_i \right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{t=1}^{r} \left(\left\| x(k) - c_t \right\|^2\right)^{\frac{1}{1-\beta}}}. \end{cases} \tag{3.10}$$

В [88-90] были введены модификации стандартного FCM, позволяющие обрабатывать векторы наблюдений, образованные категориальными

переменными. Так, в [90] показано, что использование меры несходства (3.7) приводит к оценке уровня принадлежности наблюдений $x(k)$ кластеру $cl_i$ вида

$$u_i(k) = \frac{d^{\frac{1}{1-\beta}}\left(x(k), c_i\right)}{\sum_{t=1}^{r} d^{\frac{1}{1-\beta}}\left(x(k), c_t\right)}, \tag{3.11}$$

по сути совпадающей со вторым соотношением (3.10). Для расчета же мод - прототипов вектор $x(k)$ приписывается к кластеру $cl_i$, для которого

$$u_i(k) > u_t(k), \forall t = 1,2,...,c; t \neq i. \tag{3.12}$$

Таким образом, процесс нечеткой кластеризации реализуется аналогично предыдущему в форме последовательности шагов.

Шаг 1. Некоторым достаточно произвольным образом задаются $r$ начальных прототипов $c_i, i = 1,2,...,r$.

Шаг 2. Рассчитать $N_r$ оценок несходства (3.7) для каждого $cl_i$ и каждого $x(k)$.

Шаг 3. Рассчитать уровни принадлежности каждого $x(k)$ каждому $cl_i$ согласно выражению (3.11).

Шаг 4. Приписать наблюдение $x(k)$ к $cl_i$ в соответствии с условием (3.12).

Шаг 5. Рассчитать моды-прототипы для всех кластеров $cl_i$ и соответствующие им частоты $f_{ij}^{l}$.

Шаг 6. Рассчитать $N_r$ оценок несходства новых прототипов со всеми $x(k)$.

Шаг 7. Продолжать до тех пор, пока не стабилизируются прототипы.

Как видно, данный подход принципиально не отличается от стандартного FCM, в связи с чем логично его распространить и на случай, когда объем обрабатываемой выборки $N$ заранее не фиксируется, а растет с течением времени [73, 91].

## 3.5 Возможностная нечеткая кластеризация массивов категориальных данных с использованием частотных прототипов и мер несходства

Несмотря на эффективность и широкое распространение FCM, ему присущ и существенный недостаток, который можно пояснить простым примером.

Предположим, что сформировано два кластера с прототипами $c_1$ и $c_2$ и пусть на обработку поступило наблюдение $x(k)$, не принадлежащее ни одному из кластеров, однако в смысле несходства (3.7) равноотстоящее от обоих прототипов. Тогда, в силу первого ограничения (2.2) это наблюдение с равными уровнями принадлежности будет приписано обоим классам в соответствии с оценкой (3.11).

Этого недостатка лишен метод возможностных $c$ - средних (PCM) [82, 92], порождаемый минимизацией целевой функции

$$E\left(u_i(k),\mathrm{c}_i\right)=\sum_{k=1}^{N}\sum_{i=1}^{r}u_i^{\beta}(k)\left\|x(k)-c_i\right\|^2+\sum_{i=1}^{r}\mu_i\sum_{k=1}^{N}\left(1-u_i(k)\right)^{\beta}, \tag{3.13}$$

где $\mu_i>0$ – расстояние между $x(k)$ и $c_i$, на котором уровень принадлежности $u_i(k)$ принимает значение $0.5$.

Минимизация (3.13) по $c_i$, $u_i(k)$ и $\mu_i$ ведет к результату

$$\begin{cases} c_i=\dfrac{\sum\limits_{k=1}^{N}u_i^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N}u_i^{\beta}(k)}, \\ u_i(k)=\dfrac{1}{1+\left(\dfrac{\left\|x(k)-c_i\right\|^2}{\mu_i}\right)^{\frac{1}{\beta-1}}}, \\ \mu_i=\left(\sum\limits_{k=1}^{N}u_i^{\beta}(k)\right)^{-1}\left(\sum\limits_{k=1}^{N}u_i^{\beta}(k)\left\|x(k)-c_i\right\|^2\right), \end{cases} \tag{3.14}$$

который в случае номинальных переменных приобретает вид

$$\begin{cases} u_i(k) = \left(1 + \dfrac{d\left(x(k), \mathrm{c}_i\right)}{\mu_i}\right)^{\frac{1}{1-\beta}}, \\ \mu_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k) d\left(x(k), \mathrm{c}_i\right)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}. \end{cases} \tag{3.15}$$

Сам же процесс возможностной нечеткой кластеризации реализуется в форме последовательности шагов.

Шаг 1. Некоторым достаточно произвольным образом задаются $r$ начальных прототипов $c_i, i = 1, 2, ..., r.$

Шаг 2. Рассчитать $N_r$ оценок несходства (3.7) для каждого $cl_i$ и каждого $x(k).$

Шаг 3. Рассчитать уровни принадлежности каждого $x(k)$ каждому $cl_i$ согласно выражению (3.15).

Шаг 4. Приписать наблюдение $x(k)$ к $cl_i$ в соответствии с условием (3.12).

Шаг 5. Рассчитать моды-прототипы для всех кластеров $cl_i$ и соответствующие им частоты $f_{ij}^{l}$.

Шаг 6. Рассчитать $N_r$ оценок несходства новых прототипов со всеми $x(k).$

Шаг 7. Продолжать до тех пор, пока не стабилизируются прототипы.

Оценка (3.15) с вычислительной точки зрения несколько сложнее (3.11), однако имеет меньше недостатков, присущих FCM.

Выводы по разделу 3

1. Рассмотрен метод робастной нечеткой кластеризации данных, заданных в категориальной шкале. Основным недостатком метода является его большая

вычислительная сложность, а также неприменимость в ситуации, когда наблюдение с разной степенью принадлежности относится сразу к нескольким кластерам.

2. Проведен сравнительный анализ методов $k$-мод и $k$-средних для кластеризации категориальных данных. Выявлены и проанализированы существующие недостатки рассмотренных методов.

3. Рассмотрен метод вероятностной нечеткой кластеризации массивов категориальных данных. Преимуществом данного метода является возможность обработки данных, представленных в категориальной шкале, в условиях перекрывающихся кластеров. Вместе с тем, метод неприменим для обработки категориальных данных в on-line режиме.

4. Предложен метод возможностной нечеткой кластеризации массивов категориальных данных путем использования частотных прототипов и мер несходства, что позволило преодолеть недостатки классических методов такие, как «проклятье размерности» и «концентрация норм» и повысить точность кластеризации данных.

# РАЗДЕЛ 4
# НЕЙРО-ФАЗЗИ СИСТЕМЫ ДЛЯ КЛАССИФИКАЦИИ ДАННЫХ, ЗАДАННЫХ В ПОРЯДКОВОЙ ШКАЛЕ

## 4.1 Радиально-базисная нейронная сеть

Радиально-базисная нейронная сеть (Radial Basis Function Neural Network – RBFN) была предложена в 1988г. Это сеть, построенная на искусственных нейронах, обладающих выраженными локальными характеристиками [93]. Радиально-базисные нейронные сети основываются на оценках Парзена [94, 95], методе потенциальных функций [96], ядерной [54] и непараметрической [97-100, 101] регрессиях и являются универсальными аппроксиматорами [102 – 105].

В простейшем варианте сеть, представленная на рисунке 4.1, состоит из входного, скрытого и выходного слоев. Нейроны $\Phi$, образующие скрытый слой сети, выполняют нелинейное преобразование входного пространства $R^n$ в скрытое пространство $R^h$, при этом размерность скрытого слоя зачастую превышает размерность входного пространства.

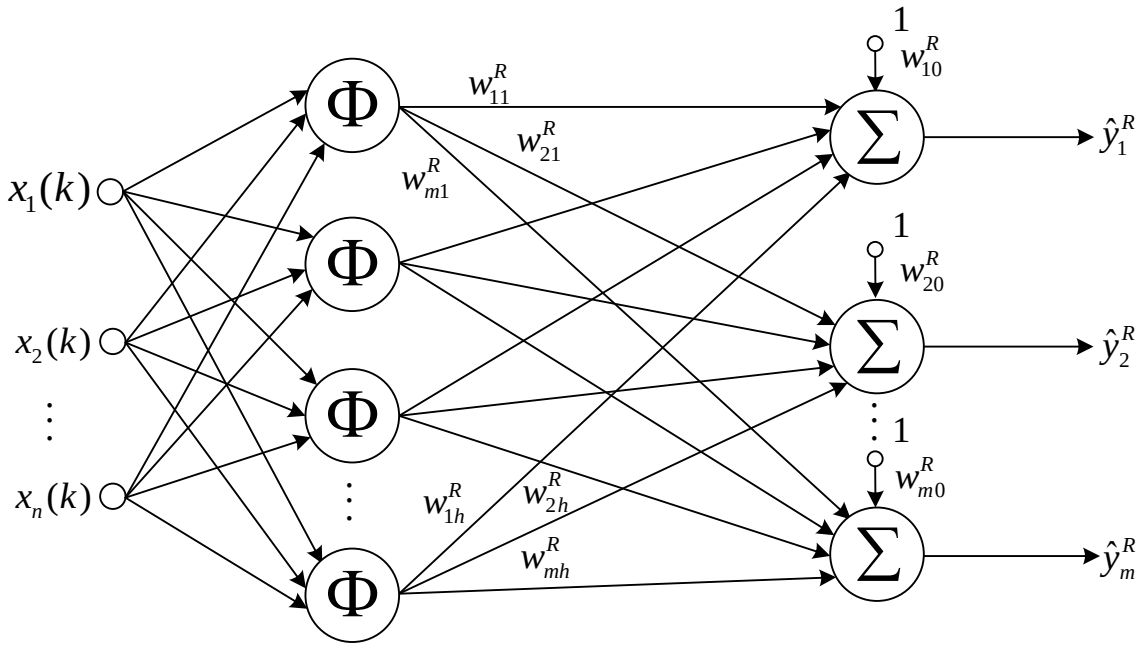



Рисунок 4.1 – Радиально-базисная нейронная сеть

Выходной слой выполняет линейное преобразование выхода скрытого слоя,

формируя отклик сети $\hat{y}^R = (\hat{y}_1^R, \hat{y}_2^R, \ldots, \hat{y}_m^R)^T$ на входной сигнал $x = (x_1, x_2, \ldots, x_n)^T$ в виде

$$\hat{y}^R = F_i(x) = w_0^R + \sum_{i=1}^{h} w_i^R \varphi_i^R(x),\ i = 1,2 \ldots h,$$

где $\varphi_i^R(x)$ – радиально - базисные функции;

$w_i^R$ – синаптические веса;

$h$ – размерность скрытого слоя;

$w_0^R$ – порог нейрона, являющийся фиксированным значением.

Скрытый слой сети представлен радиально-базисными функциями $\varphi_i^R(x)$, которые преобразуют $n$ - мерное пространство входов в $m$ - мерное пространство выходов $R^n \to R^m$

$$\varphi_i^R(x) = \Phi\left(\|x - c_i\|, \sigma_i\right) = \Phi(d, \sigma_i),$$

где $d$ – расстояние между входным вектором $x$ и центром $c_i$ в принятой метрике;

$\sigma_i$ – параметр ширины, определяющий локальную область входного пространства, на которую «откликается» данная функция.

Чаще всего используются гауссовские функции (4.1), представленные на рисунке 4.2

$$\varphi_i^R\left(x\right) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right), i = 1,2,\ldots,h^R. \qquad (4.1)$$

Обычно при решении различных задач центры узлов $c_i$ и параметры ширины $\sigma_i$ остаются неизменными, а настраиваются только синаптические веса $w_i^R$. Однако, в случае решения более сложных задач могут настраиваться все три

вышеперечисленных параметра. Это приводит к тому, что число базисных функций экспоненциально растет с размерностью входного пространства и возникает, так называемое, «проклятье размерности».

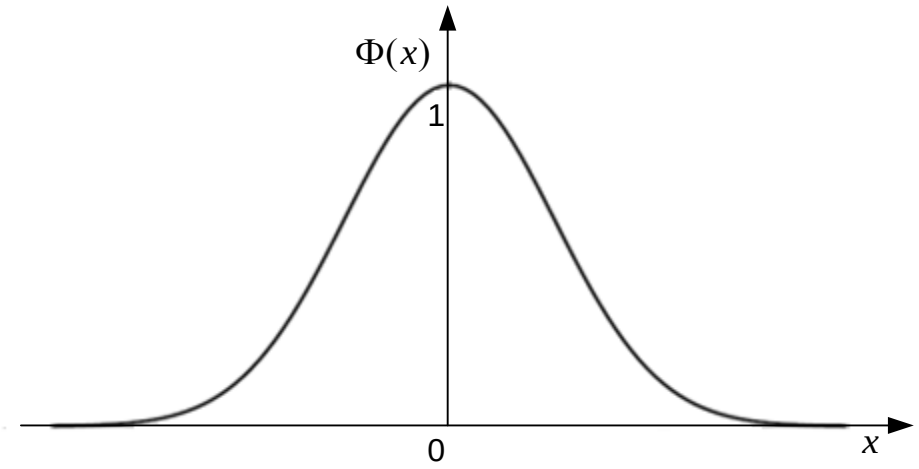



Рисунок 4.2 – Одномерная гауссова радиально-базисная функция

Следует так же отметить сложность в нахождении центров активационных функций, для чего приходится применять дополнительные алгоритмы кластеризации или процедуры самообучения.

## 4.2 Нейро-фаззи сеть Ванга-Менделя

Для борьбы с недостатками радиально-базисных нейронных сетей и, прежде всего, «проклятья размерности» была предложена нечеткая нейронная сеть Ванга-Менделя, реализующая нелинейную зависимость [106-109]

$$\hat{y} = \frac{\sum_{i=1}^{M} w_i \prod_{j=1}^{n} \mu_{ij}\left(x_j\right)}{\sum_{i=1}^{M} \prod_{j=1}^{n} \mu_{ij}\left(x_j\right)}, \tag{4.2}$$

где $w_i$ – весовые синаптические коэффициенты;

$\mu_{ij}$ – функции принадлежности нечетких множеств, обычно имеющие вид одномерных гауссовых функций (1.3);

$n$ – размерность вектора $x$;

$M$ – число продукционных правил.

Классическая нейро-фаззи сеть Ванга-Менделя состоит из четырех слоев. Ее архитектура представлена на рисунке 4.3.

Первый слой осуществляет фаззификацию входных наблюдений. Настраиваемые параметры данного слоя – это параметры используемых функций принадлежности.

Второй слой данной сети осуществляет вычисление результирующих функций принадлежности предпосылок нечетких правил. Этот слой не имеет настраиваемых параметров.

Третий слой состоит из двух нейронов. Он осуществляет суммирование и взвешенное суммирование выходных сигналов второго слоя. Настраиваемыми параметрами этого слоя являются весовые коэффициенты $w_i$.

Четвертый слой представлен одним нейроном. Он осуществляет нормализацию, формируя выходной сигнал $\hat{y}(x)$ и не содержит настраиваемых параметров.

В общем случае процесс обучения нечеткой сети представляет собой вычисление ряда параметров:

- значений весов нейронов выходного слоя $w_i$;
- центроидов $c_{ij}$ и отклонений $\sigma_{ij}$ радиальных базисных функций;
- числа продукционных правил $M$.

Настройка параметров сети $c_{ij}$, $\sigma_{ij}$ и $w_i$ может осуществляться методом обратного распространения ошибки, например, градиентным методом. Однако, настройку весов нейронов выходного слоя $w_i$ можно осуществить и с помощью формул для определения коэффициентов линейной регрессии по методу наименьших квадратов [107].

Для вычисления центроидов $c_{ij}$ возможно использовать следующие методы:

– размещение центров радиальных функций $c_{ij}$ в обучающих точках;

– размещение центров радиальных функций $c_{ij}$ в центрах кластеров обучающих данных;

– размещение центров радиальных функций $c_{ij}$ в узлах равномерной сетки или случайных точках.

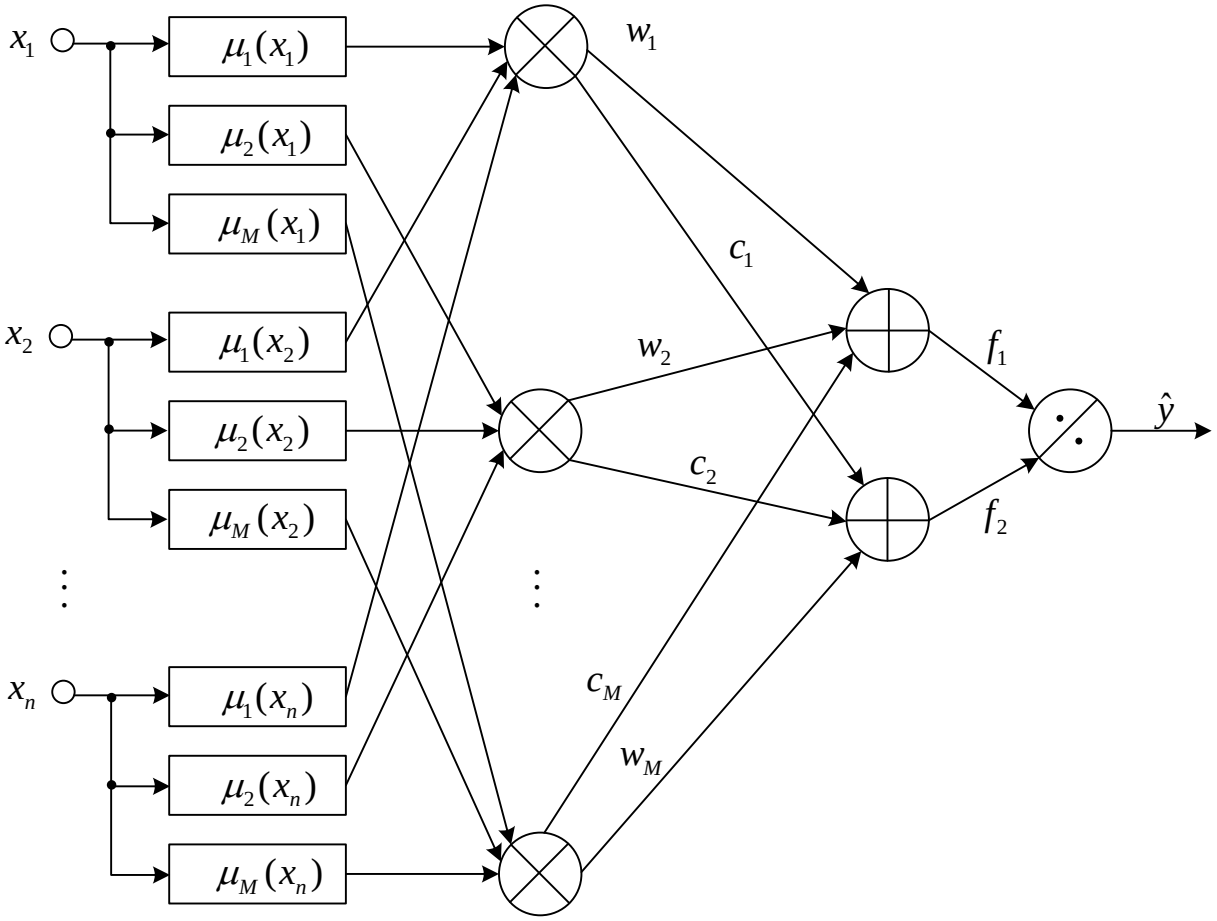



Рисунок 4.3 – Структура нейро-фаззи сети Ванга - Менделя

Достоинством представленной системы вычислительного интеллекта является возможность обучения в on-line режиме, а также прозрачность и интерпретируемость получаемых результатов. К основным недостаткам можно отнести громоздкость архитектуры, связанную с «проклятьем размерности», а также возможность возникновения дыр в фаззифицированном пространстве входов при рассеянном разбиении исходного пространства входов.

Использование решетчатого разбиения автоматически ведет к «проклятью размерности». В связи с этим целесообразным представляется использование нетрадиционных нейро-фаззи архитектур, лишенных отмеченных недостатков.

## 4.3 Структура нео-фаззи нейрона

Нео-фаззи нейрон (NFN), введенный Т. Ямакавой с коллегами [110-112], является нелинейной обучаемой системой, схема которой приведена на рисунке 4.4. Данная конструкция отличается высокой скоростью обучения, возможностью нахождения глобального минимума критерия обучения в реальном времени и вычислительной простотой. Нелинейные синапсы $NS_j$ нео-фаззи нейрона реализуют элементарные правила нечеткого вывода вида

$$if\ x_j\ is\ X_j\ then\ f_j(x_j)=\sum_{l=1}^{m_j}\mu_{jl}(x_j)w_{jl},\ j=1...n,$$

где $X_j$ – лингвистическое значение (нечеткое множество) на $j$ - м входе;

$m_j$ – количество функций принадлежности в каждом нелинейном синапсе;

$\mu_{jl}(x_j)$ – функции принадлежности;

$w_{jl}$ – весовые синаптические коэффициенты.

При подаче на вход нейрона векторного сигнала $x(k)$ на его выходе появляется скалярное значение

$$y(k)=\sum_{j=1}^{n}f_j\left(x_j(k)\right)=\sum_{j=1}^{n}\sum_{l=1}^{m_j}\mu_{jl}\left(x_j(k)\right)w_{jl}(k), \qquad (4.3)$$

зависящее от настраиваемых синаптических весов $w_{jl}(k-1)$ и функций принадлежности $\mu_{jl}$.

Если аргументы активационных функций нео-фаззи нейрона – скаляры, то он имеет близкую структуру с радиально-базисными нейронными сетям (RBFN) [113, 114]. Также, он в определенной мере близок к системе нечеткого вывода (FIS) М. Сугено [115].

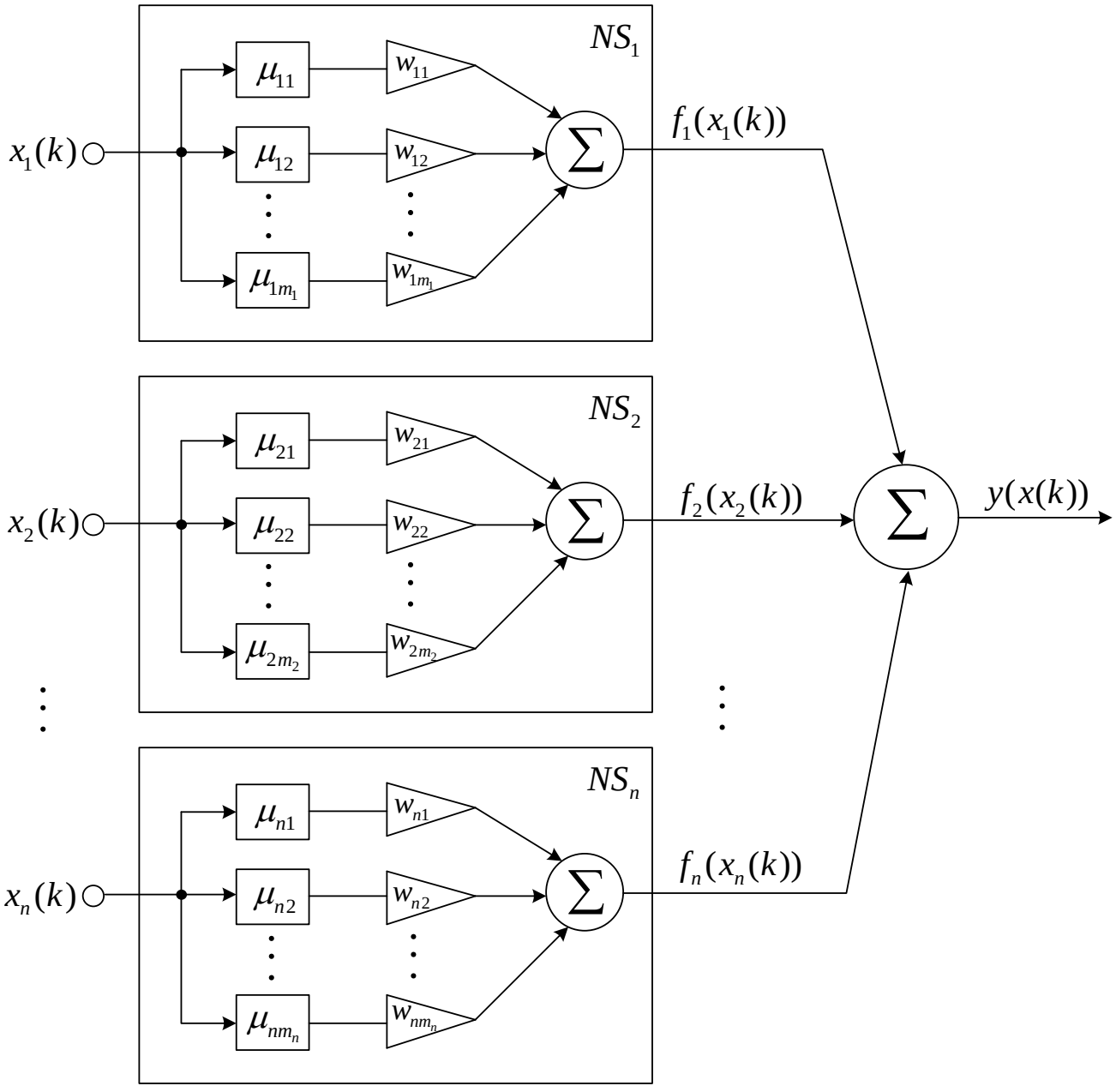



Рисунок 4.4 – Структура нео-фаззи нейрона

В качестве функций принадлежности в нео-фаззи нейроне обычно используются треугольные конструкции, представленные на рисунке 4.5 и описываемые следующим образом:

$$\mu_{j1}(x_j)=\begin{cases}\dfrac{c_{j1}-x_j}{c_{j2}},\ x_j\in[0,c_{j2}],\\[2ex] 0,\ x_j\notin[0,c_{j2}],\end{cases}$$

$$\mu_{jl}(x_j)=\begin{cases}\dfrac{x_j-c_{j,l-1}}{c_{jl}-c_{j,l-1}},\ x_j\in[c_{j,l-1},c_{jl}],\\[2ex] \dfrac{c_{j,l+1}-x_j}{c_{j,l+1}-c_{jl}},\ x_j\in[c_{jl},c_{j,l+1}],\ l=2,...,m_j-1,\\[2ex] 0 \text{ в противном случае},\end{cases}$$

$$\mu_{jm_j}(x_j)=\begin{cases}\dfrac{x_j-c_{j,m_j-1}}{1-c_{j,m_j-1}},\ x_j\in[c_{j,m_j-1},1],\\[2ex] 0,\ x_j\notin[c_{j,m_j-1},1],\end{cases}$$

$$c_{j1}=0, c_{j2}=\frac{1}{m_j-1},...,c_{jl}=\frac{l-1}{m_j-1},...,c_{jm_j}=1.$$

Отметим, что такая конструкция функций принадлежности автоматически обеспечивает разбиение Руспини

$$\sum_{l=1}^{m_j}\mu_{jl}\left(x_j(k)\right)=1, l=1,2,...,m_j; j=1,2,...,n, \tag{4.4}$$

которое делает ненужным введение скрытого слоя нормализации, обычно присутствующего в нейро-фаззи системах.

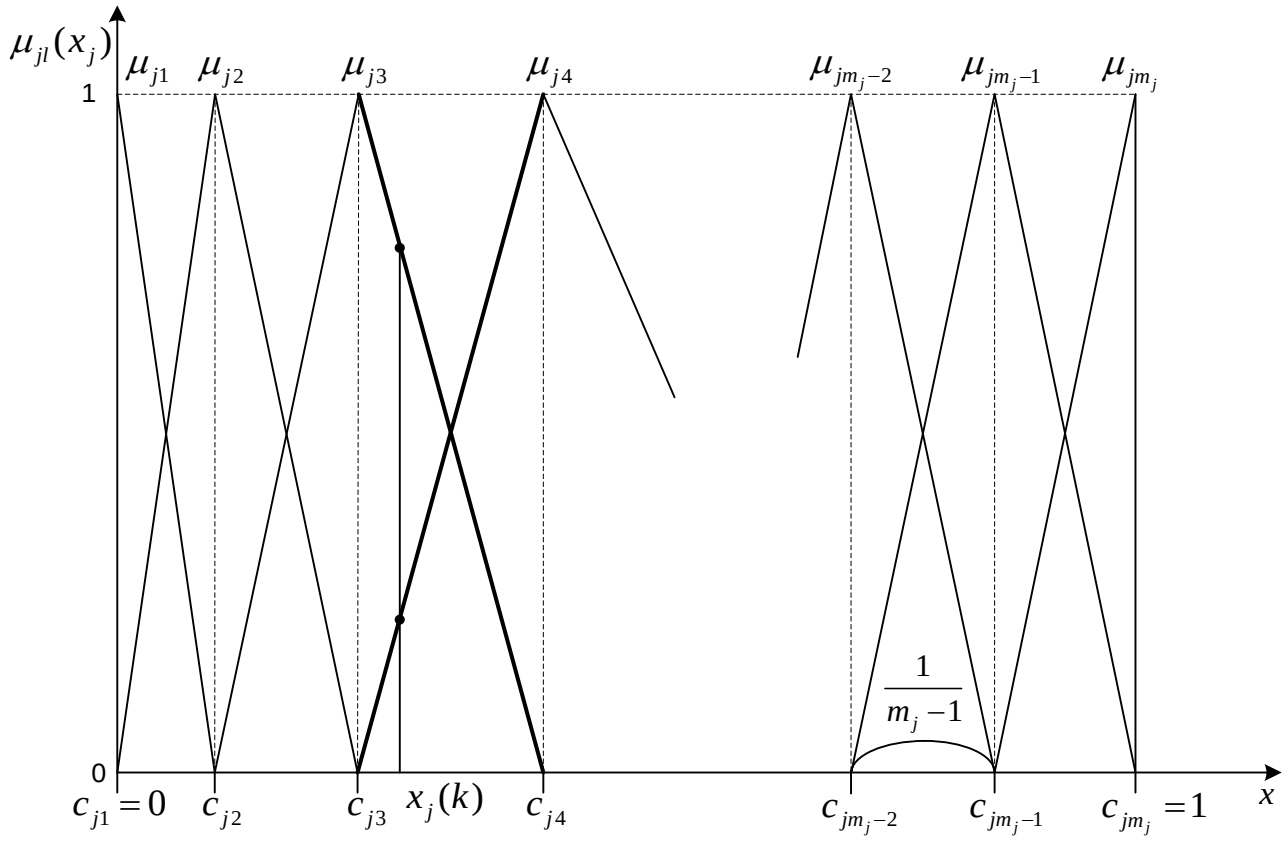



Рисунок 4.5 – Треугольные функции принадлежности нео-фаззи нейрона

В случае, когда активен нечеткий интервал $p$, выход нелинейного синапса можно выразить следующим образом:

$$f_j\left(x_j(k)\right)=\sum_{l=1}^{m_j}\mu_{jl}\left(x_j(k)\right)w_{jl}(k)=\mu_{jp}\left(x_j(k)\right)w_{jp}(k)+\mu_{j,p+1}\left(x_j(k)\right)w_{j,p+1}(k)=$$

$$=\frac{c_{j,p+1}-x_j(k)}{c_{j,p+1}-c_{jp}}w_{jp}(k)+\frac{x_j(k)-c_{jp}}{c_{j,p+1}-c_{jp}}w_{j,p+1}(k)=a_j(k)x_j(k)+b_j(k),$$

где

$$a_j(k)=\frac{w_{j,p+1}(k)-w_{jp}(k)}{c_{j,p+1}-c_{jp}},$$

$$b_j(k) = \frac{c_{j,p+1} w_{jp}(k) - c_{jp} w_{j,p+1}(k)}{c_{j,p+1} - c_{jp}}.$$

Фаззификация текущего сигнала $x_j(k)$, выполняемая нелинейным синапсом $NS_j$, представлена на рисунке 4.5. Жирными линиями показаны активные функции принадлежности. Таким образом, каждый нелинейный синапс реализует кусочно-линейную аппроксимацию $f_j(x_j)$ нелинейного сигнала $x_j$.

В качестве критерия обучения нео-фаззи нейрона обычно используется стандартная квадратичная ошибка

$$E(k) = \frac{1}{2}\left(\tilde{d}(k) - y(k)\right)^2 = \frac{1}{2}e^2(k) = \frac{1}{2}\left(\tilde{d}(k) - \sum_{j=1}^{n}\sum_{l=1}^{m_i}\mu_{jl}\left(x_j(k)\right)w_{jl}\right)^2,$$

максимизация которой с помощью градиентной процедуры ведет к алгоритму обучения [112]

$$w_{jl}(k+1) = w_{jl}(k) + \eta e(k)\mu_{jl}\left(x_j(k)\right), \qquad (4.5)$$

где $\tilde{d}(k)$ – внешний обучающий сигнал;

$\eta$ – параметр шага поиска, определяющий скорость сходимости процесса обучения.

Для улучшения аппроксимирующих свойств нео-фаззи нейрона в [116] была предложена конструкция, названная двойной нео-фаззи нейрон.

## 4.4 Двойной нео-фаззи нейрон

Архитектура двойного нео-фаззи нейрона приведена на рисунке 4.6 и содержит два слоя: первый слой, образованный $n$ нелинейными синапсами $NS_j$ с $m_j$ функциями принадлежности и синаптическими весами каждый, и выходной

слой, образованный нелинейным синапсом $NS_0$ с $m_0$ функциями принадлежности $\mu_{l_0 0}$, $l_0 = 1,2,...,m_0$ и синаптическими весами $w_{l_0 0}$.

При подаче на вход двойного нео-фаззи нейрона вектора - образа $x(k)$ на его выходе появляется сигнал

$$y(k) = f_0\left(u(k)\right) = f_0\left(\sum_{j-1}^{n} f_j\left(x_j(k)\right)\right) = \sum_{l_0=1}^{m_0} \mu_{l_0 0}\left(u(k)\right) w_{l_0 0} = \sum_{l_0=1}^{m_0} \mu_{l_0 0}\left(\sum_{j=1}^{n}\sum_{l=1}^{m_j} \mu_{jl}\left(x_j(k)\right) w_{jl}\right) w_{l_0 0},$$

где $\mu_{jl}(x_j)$ – функции принадлежности первого слоя;

$u(k)$ – выходной сигнал первого слоя;

$\mu_{l_0}\left(u(k)\right)$ – функции принадлежности выходного слоя;

$w_{jl}$ – синаптические веса первого слоя;

$w_{l_0 0}$ – синаптические веса выходного слоя.

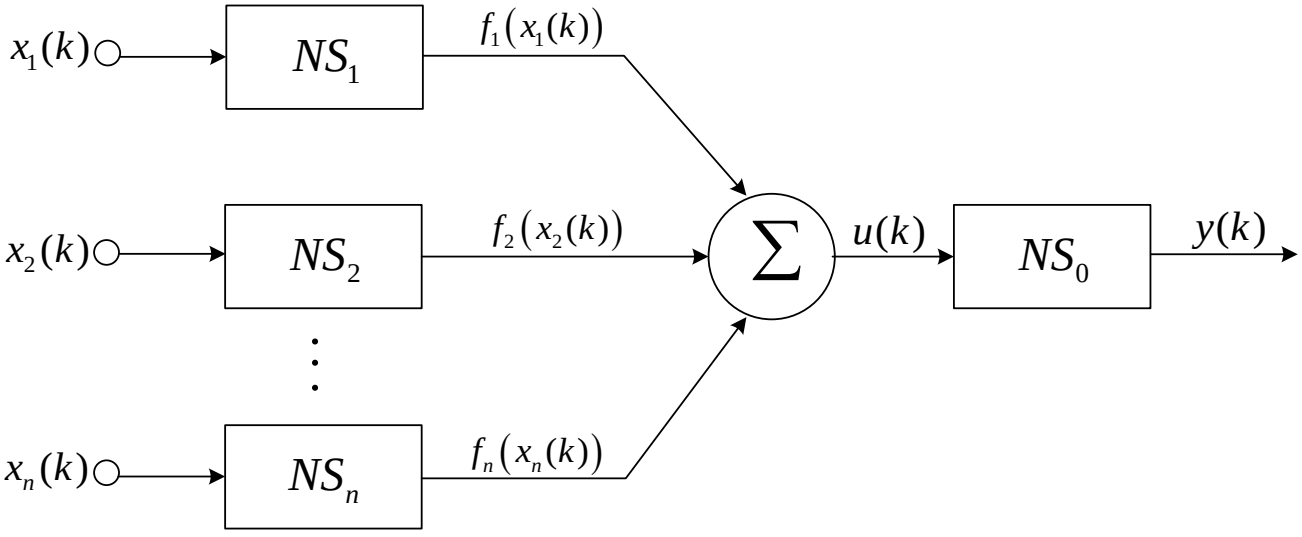



Рисунок 4.6 – Архитектура двойного нео-фаззи нейрона

Можно видеть, что значение выходного сигнала двойного нео-фаззи нейрона определяется как значениями компонент $x_j(k)$ входного образа, так и значениями $\sum_{j=1}^{n} m_j + m_0$ функций принадлежности и соответствующих им синаптических весов.

Равномерно распределенные на отрезке [0,1] треугольные функции принадлежности, удовлетворяющие условию (4.4), могут быть заданы в виде

$$\mu_{j1}(x_j)=\begin{cases}\dfrac{c_{j2}-x_j}{c_{j2}},\ x_j\in[0,c_{j2}],\\ \\ 0,\ x_j\notin[0,c_{j2}],\end{cases} \tag{4.6}$$

$$\mu_{jl}(x_j)=\begin{cases}\dfrac{x_j-c_{j,l-1}}{c_{jl}-c_{j,l-1}},\ x_j\in[c_{j,l-1},c_{jl}],\\ \\ \dfrac{c_{j,l+1}-x_j}{c_{j,l+1}-c_{jl}},\ x_j\in[c_{jl},c_{j,l+1}],\\ \\ l=2,...,m_{j-1},\\ \\ 0\text{ в противном случае,}\end{cases} \tag{4.7}$$

$$\mu_{jm_j}(x_j)=\begin{cases}\dfrac{x_j-c_{j,m_j-1}}{1-c_{j,m_j-1}},\ x_j\in[c_{j,m_j-1},1],\\ \\ 0,\ x_j\notin[c_{j,m_j-1},1],\end{cases} \tag{4.8}$$

$$\mu_{10}(u)=\begin{cases}\dfrac{c_{20}-u}{c_{20}},\ u\in[0,c_{20}],\\ \\ 0,\ u\notin[0,c_{20}],\end{cases} \tag{4.9}$$

$$\mu_{l_0 0}(u) = \begin{cases} \dfrac{u - c_{l_0-1,0}}{c_{l_0 0} - c_{l_0-1,0}}, u \in [c_{l_0-1,0}, c_{l_0 0}], \\ \dfrac{c_{l_0+1,0} - u}{c_{l_0+1,0} - c_{l_0 0}}, u \in [c_{l_0 0}, c_{l_0+1,0}], \\ l_0 = 2,...,m_{0-1}, \\ 0 \text{ в противном случае}, \end{cases} \tag{4.10}$$

$$\mu_{m_0 0}(u) = \begin{cases} \dfrac{u - c_{m_0-1,0}}{1 - c_{m0-1,0}}, u \in [c_{m_0-1,0}, 1], \\ 0, u \notin [c_{m_0-1,0}, 1], \end{cases} \tag{4.11}$$

где $c_{jl}$, $c_{l_0 0}$ – центры соответствующих функций принадлежности.

При этом расстояние между центрами постоянно для каждого нелинейного синапса и составляет

$$c_{jl} - c_{j,l-1} = (m_j - 1)^{-1},$$

$$c_{l_0 0} - c_{l_0-1,0} = (m_0 - 1)^{-1}$$

соответственно.

Использование разбиения Руспини приводит к тому, что на каждом шаге обучения активируются только две соседние функции принадлежности. Обозначая эти функции принадлежности $\mu_{jp}$, $\mu_{j,p+1}$, $j = 0,1,...,n$, можно записать

$$f_j\left(x_j(k)\right)=\sum_{l=1}^{m_j}\mu_{jl}\left(x_j(k)\right)w_{jl}(k)=\mu_{jp}\left(x_j(k)\right)w_{jp}(k)+\mu_{j,p+1}\left(x_j(k)\right)w_{j,p+1}(k)=$$

$$=\frac{c_{j,p+1}-x_j(k)}{c_{j,p+1}-c_{jp}}w_{jp}(k)+\frac{x_j(k)-c_{jp}}{c_{j,p+1}-c_{jp}}w_{j,p+1}(k)=a_j(k)x_j(k)+b_j(k),$$

где

$$a_j(k)=\frac{w_{j,p+1}(k)-w_{jp}(k)}{c_{j,p+1}-c_{jp}},$$

$$b_j(k)=\frac{c_{j,p+1}w_{jp}(k)-c_{jp}w_{j,p+1}(k)}{c_{j,p+1}-c_{jp}}$$

и

$$u(k)=\sum_{j=1}^{n}a_j(k)x_j(k)+b_j(k),$$

$$y(k)=\sum_{l_0=1}^{m_0}\mu_{l_0 0}\left(u(k)\right)w_{l_0 0}(k)=\mu_{p0}\left(u(k)\right)w_{po}(k)+\mu_{p+1,0}\left(u(k)\right)w_{p+1,0}(k)=$$

$$=\frac{c_{p+1,0}-u(k)}{c_{p+1,0}-c_{p0}}w_{p0}(k)+\frac{u(k)-c_{p0}}{c_{p+1,0}-c_{p0}}w_{p+1,0}(k)=a_0(k)u(k)+b_0(k),$$

где

$$a_0(k)=\frac{w_{p+1,0}(k)-w_{p0}(k)}{c_{p+1,0}-c_{p0}},$$

$$b_0(k)=\frac{c_{p+1,0}w_{p0}(k)-c_{p0}w_{p+1,0}(k)}{c_{p+1,0}-c_{p0}}.$$

Таким образом, двойной нео-фаззи нейрон обеспечивает кусочно - линейную аппроксимацию некоторой нелинейной разделяющей функции в виде

$$y(k) = a_0(k)\left( \sum_{j=1}^{n} a_j(k) x_j(k) + b_j(k) \right) + b_0(k),$$

где настраиваемые параметры $a_j(k)$, $b_j(k)$ определяются как значениями соответствующих функций принадлежности, так и обучаемыми синаптическими весами.

4.4.1 Фаззификация порядковых данных и построение функций принадлежности

Двойной нео-фаззи нейрон, впрочем, как и обычный нео-фаззи нейрон, предназначен для обработки информации, заданной в шкале натуральных чисел. Ситуация существенно усложняется, когда исходные данные заданы не в числовой, а в порядковой (ранговой) шкале, что часто встречается в социологии, экономике, медицине, образовании и т.п. [117]. В одномерном случае такая информация задается в виде упорядоченной последовательности лингвистических переменных $x_j^1, x_j^2, ..., x_j^{l_j}, ..., x_j^{m_j}, 1 \le ... \le l_j - 1 \le l_j \le l_{j+1} \le ... \le m_j \le ... \le N$, где $x_j^{l_j}$ – собственно лингвистическая переменная, $l_j$ – соответствующий ранг.

Таким образом, исходной информацией для решения задачи нечеткой классификации является выборка образов, сформированная из $N$ $n$ - мерных векторов признаков $X = \{x(1), x(2), ..., x(k), ..., x(N)\}$ (здесь $x(k) = \{x_j^{l_j}(k)\}$, $j = 1, 2, ..., n$, $l_j = 1, 2, ..., m_j$ – ранг конкретного значения лингвистической переменной по $j$-й координате $n$- мерного пространства для $k$ - го наблюдения) и выборка обучающих сигналов $\tilde{D} = \{\tilde{d}(1), \tilde{d}(2), ..., \tilde{d}(k), ..., \tilde{d}(N)\}$, где $\tilde{d}(k) = \tilde{d}^{l_0}(k)$, $l_0 = 1, 2, ..., m_0$ – ранг значения обучающего сигнала в выборке $\tilde{D}$. В результате

обучения двойного нео-фаззи нейрона по ранжированным данным должно быть обеспечено разбиение исходного массива данных $X$ на $m_0$ возможно пересекающихся классов с вычислением уровней принадлежности $\mu_{l_0 0}$ $k$ - го образа $l_0$ - му классу.

Процесс фаззификации последовательности ранговых лингвистических переменных рассмотрим на примере одномерной выборки $x_j(1), x_j(2),..., x_j(N)$, где каждому из наблюдений $x_j(k)$ может быть приписан один из рангов $l_j$, $l_j = 1,2,...,m_j$.

Используя метод фаззификации, рассмотренный в разделе 2.4, вычисляем относительную частоту появления каждого наблюдения в выборке по формуле (2.7), при этом естественно выполняется условие

$$\sum_{l_j=1}^{m_j} f_{l_j} = 1.$$

На основе полученных частот формируются несимметричные неравномерно расположенные функции принадлежности $\mu_{jl}$, $\mu_{l_0 0}$ с центрами, вычисляемыми с помощью рекуррентных соотношений (2.12).

Сами же функции принадлежности вычисляются с помощью выражений аналогичных соотношениям (4.6) – (4.11) с той лишь разницей, что вместо (4.6) используется

$$\mu_{j1}(x_j) = \begin{cases} 1, x_j \in [0, c_{j1}], \\ \dfrac{c_{j2} - x_j}{c_{j2} - c_{j1}}, x_j \in [c_{j1}, c_{j2}], \\ 0, x_j \notin [0, c_{j2}], \end{cases} \tag{4.12}$$

вместо (4.8) используется

$$\mu_{jm_j}(x_j)=\begin{cases}\dfrac{x_j-c_{j,m_j-1}}{c_{jm_j}-c_{j,m_j-1}},\ x_j\in[c_{j,m_j-1},c_{jm_j}],\\ 1,\ x_j\in[c_{jm_j},1],\\ 0,\ x_j\notin[c_{j,m_j-1},1],\end{cases} \tag{4.13}$$

вместо (4.9) –

$$\mu_{10}(u)=\begin{cases}1,\ u\in[0,c_{10}],\\ \dfrac{c_{20}-u}{c_{20}},u\in[c_{10},c_{20}],\\ 0,\ u\notin[0,c_{20}],\end{cases} \tag{4.14}$$

вместо (4.11) –

$$\mu_{m_0 0}(u)=\begin{cases}\dfrac{u-c_{m_0-1,0}}{c_{m_0 0}-c_{m_0-1,0}},\ u\in[c_{m_0-1,0},c_{m_0 0}],\\ 1,\ u\in[c_{m_0 0},1],\\ 0,\ u\notin[c_{m_0-1,0},1].\end{cases} \tag{4.15}$$

Способ задания функций принадлежности (4.7), (4.10), (4.12), (4.13), (4.14), (4.15) также обеспечивает выполнение условия разбиения Руспини.

### 4.4.2 Процедура обучения двойного нео-фаззи нейрона

Для обучения двойного нео-фаззи нейрона будем использовать градиентную процедуру минимизации с переменным параметром шага поиска $\eta_j(k)$. Тогда для настройки выходного синапса $NS_0$ можно записать алгоритм

$$\begin{cases} w_{l_0 0}(k+1) = w_{l_0 0}(k) + \eta_0(k) e(k) \mu_{l_0 0}\left(u(k)\right), l_0 = p, p+1, \\ w_{l_0 0}(k+1) = w_{l_0 0}(k), \forall l_0 \neq p \neq p+1. \end{cases} \tag{4.16}$$

Таким образом, на каждой итерации настраиваются веса, соответствующие активированным функциям принадлежности $\mu_{p0}$ и $\mu_{p+1,0}$. Для увеличения скорости сходимости и введения дополнительных сглаживающих свойств целесообразно использовать алгоритм вида [118]

$$\begin{cases} w_{l_0 0}(k+1) = w_{l_0 0}(k) + \eta_0^{-1}(k) e(k) \mu_{l_0 0}\left(u(k)\right), l_0 = p, p+1, \\ \eta_0(k+1) = \alpha \eta_0(k) + \mu_{p0}^2\left(u(k+1)\right) + \mu_{p+1,0}^2\left(u(k+1)\right), \\ w_{l_0 0}(k+1) = w_{l_0 0}(k), \forall l_0 \neq p \neq p+1, \\ 0 \leq \alpha \leq 1, \end{cases} \tag{4.17}$$

совпадающий при $\alpha = 0$ с оптимальным по быстродействию методом Качмажа-Уидроу-Хоффа [119, 120], а при $\alpha = 1$ с процедурой стохастической аппроксимации Гудвина-Рэмеджа-Кэйнеса [121, 122].

Для настройки синаптических весов первого слоя запишем критерий обучения в виде

$$E(k)=\frac{1}{2}\left(\tilde{d}(k)-f_0\left(u(k)\right)\right)^2=\frac{1}{2}\left(\tilde{d}(k)-f_0\left(\sum_{j=1}^{n}\sum_{l=1}^{m_j}\mu_{jl}\left(x(k)\right)w_{jl}\right)\right)^2 \qquad (4.18)$$

и введем производную

$$\frac{\partial E(k)}{\partial w_{jl}}=-e(k)\frac{\partial f_0\left(u(k)\right)}{\partial u(k)}\cdot\frac{\partial u(k)}{\partial w_{jl}}=-e(k)a_0(k)\frac{\partial u(k)}{\partial w_{jl}}.$$

Тогда градиентная процедура минимизации (4.18) может быть записана в форме

$$\begin{cases} w_{jl}(k+1)=w_{jl}(k)+\eta_j(k)e(k)a_0(k)\mu_{jl}\left(x_j(k)\right),\ l=p,\ p+1;\ j=1,2,...,n, \\ w_{jl}(k+1)=w_{jl}(k),\ \forall l\neq p\neq p+1, \end{cases}$$

где параметр шага $\eta_j(k)$ подлежит определению.

Вводя обозначение

$$\mu_{jl0}\left(x_j(k)\right)=a_0(k)\mu_{jl}\left(x_j(k)\right)$$

и используя технику оптимизации, предложенную в [123, 124], получаем простой и эффективный алгоритм обучения нелинейных синапсов первого слоя, совпадающий по структуре с процедурой (4.17)

$$\begin{cases} w_{jl}(k+1) = w_{jl}(k) + \eta_j^{-1}(k) e(k) \mu_{jl0}\left(x_j(k)\right), l = p, p+1;\ j = 1,2,...,n, \\ \eta_j(k+1) = \alpha \eta_j(k) + \mu_{jp0}^2\left(x_j(k+1)\right) + \mu_{j,p+10}^2\left(x_j(k+1)\right), \\ w_{jl}(k+1) = w_{jl}(k), \forall l \neq p \neq p+1, \\ 0 \leq \alpha \leq 1. \end{cases}$$

Предложенный подход к нечеткой классификации данных в порядковой шкале на основе двойного нео-фаззи нейрона, обучаемого с помощью быстродействующего алгоритма, обладает дополнительными сглаживающими свойствами. Данная сеть позволяет эффективно обрабатывать информацию, заданную как в числовой, так и в порядковой шкалах.

Выводы по разделу 4

1. Рассмотрены и проанализированы радиально-базисные нейронные сети, при этом использование гауссовских функций при решении рассматриваемой задачи требует настройки ряда параметров таких, как центры узлов $c_i$, параметры ширины $\sigma_i$ и синаптические веса $w_i^R$, что приводит к экспоненциальному росту числа базисных функций и возникновению, так называемого, «проклятья размерности».

Кроме того, нахождение центров активационных функций также является нетривиальной задачей, для решения которой приходится применять дополнительные алгоритмы кластеризации или процедуры самообучения.

2. Проведен анализ нейро-фаззи сети Ванга-Менделя и методов ее обучения для рассматриваемой задачи. К ее достоинствам можно отнести возможность обучения в on-line режиме, а также прозрачность и интерпретируемость получаемых результатов. Основным недостатками являются громоздкость архитектуры, связанная с «проклятьем размерности», а также возможность

возникновения дыр в фаззифицированном пространстве входов при рассеянном разбиении исходного пространства входов.

3. Рассмотрена нейро-фаззи система на основе нео-фаззи нейрона (NFN) и методы ее обучения. Данная конструкция отличается высокой скоростью обучения, возможностью нахождения глобального минимума критерия обучения в реальном времени и вычислительной простотой. Однако, эта система предназначена для работы с данными в числовых шкалах и не может использоваться в случае, когда информация задана в нечисловом виде. В связи с этим предложена нейро-фаззи система на основе двойного нео-фаззи нейрона, обладающая улучшенными аппроксимирующими свойствами для рассматриваемых задач. Введенная процедура обучения двойного нео-фаззи нейрона и способ фаззификации порядковых данных просты в реализации. Представленная архитектура отличается быстродействием, высокой точностью классификации порядковых данных и позволяет устранить недостатки рассмотренных выше нейронных сетей.

# РАЗДЕЛ 5
# ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ И РЕШЕНИЕ ПРАКТИЧЕСКИХ ЗАДАЧ

Имитационное моделирование – метод, позволяющий строить модели, описывающие процессы так, как они происходили бы в действительности. Такую модель можно «проиграть» во времени как для одного испытания, так и заданного их множества. При этом результаты будут определяться случайным характером процессов. По этим данным можно получить достаточно устойчивую статистику.

В данном разделе приведены результаты моделирования разработанных в предыдущих разделах методов кластеризации, обучения и нейро-фаззи архитектур. Моделирование выполнялось на примерах решения стандартных тестовых задач классификации, а также для решения практических задач: автоматический анализ клиентской базы предприятия и автоматическая обработка термограмм для диагностики дефектов электрооборудования. В данном разделе также приведены результаты моделирования классических нейро-фаззи систем и методов кластеризации с целью сравнительной оценки качества решения рассматриваемых задач.

## 5.1 Моделирование методов нечеткой кластеризации порядковых данных

### 5.1.1 Моделирование метода нечеткой кластеризации порядковых данных на основе частотных прототипов

Для проверки работоспособности предложенного метода были взяты данные об успеваемости студентов потока на одном из факультетов Харьковского национального университета радиоэлектроники. Набор данных содержит оценки по шести предметам для 135 человек.

Статистический анализ показал, что для каждой из переменных (дисциплин) гипотеза о том, что оценки имеют нормальный закон распределения, не подтверждается (рис. 5.1).

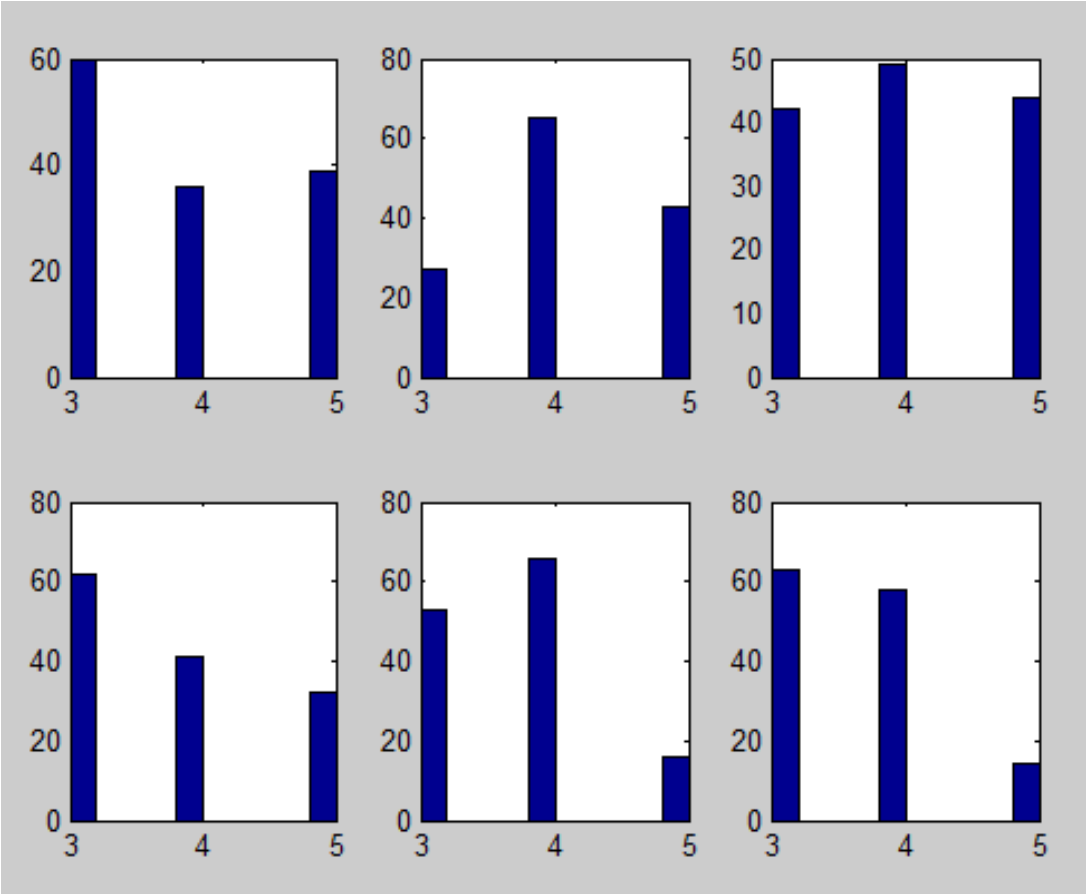



Рисунок 5.1 – Гистограмма распределения оценок по шести предметам

В результате работы метода были определены центроиды для каждого из рангов (оценок) по каждой из переменных и построены функции принадлежности с областями влияния, представленные на рисунке 5.2.

Предложенный алгоритм сравнивался с методом нечётких $c$ - средних при $\beta = 2$ (FCM) и методом кластеризации порядковых данных К. Брауэра (BFCM) [65]. Так как классы в выборке изначально заданы не были, проблематично говорить о точности кластеризации каждого из методов. Однако, анализируя рассчитанные функции принадлежности некоторых наблюдений (табл. 5.1), можно отметить, что предлагаемый метод более корректно производит обработку данных. Интуитивно понятно, что представленные в таблице 5.1 наблюдения с определенной степенью принадлежности относится к кластерам «хорошо» и «отлично» и никоим образом не могут принадлежать кластеру «удовлетворительно», что показывают методы FCM и BFCM (табл. 5.2).

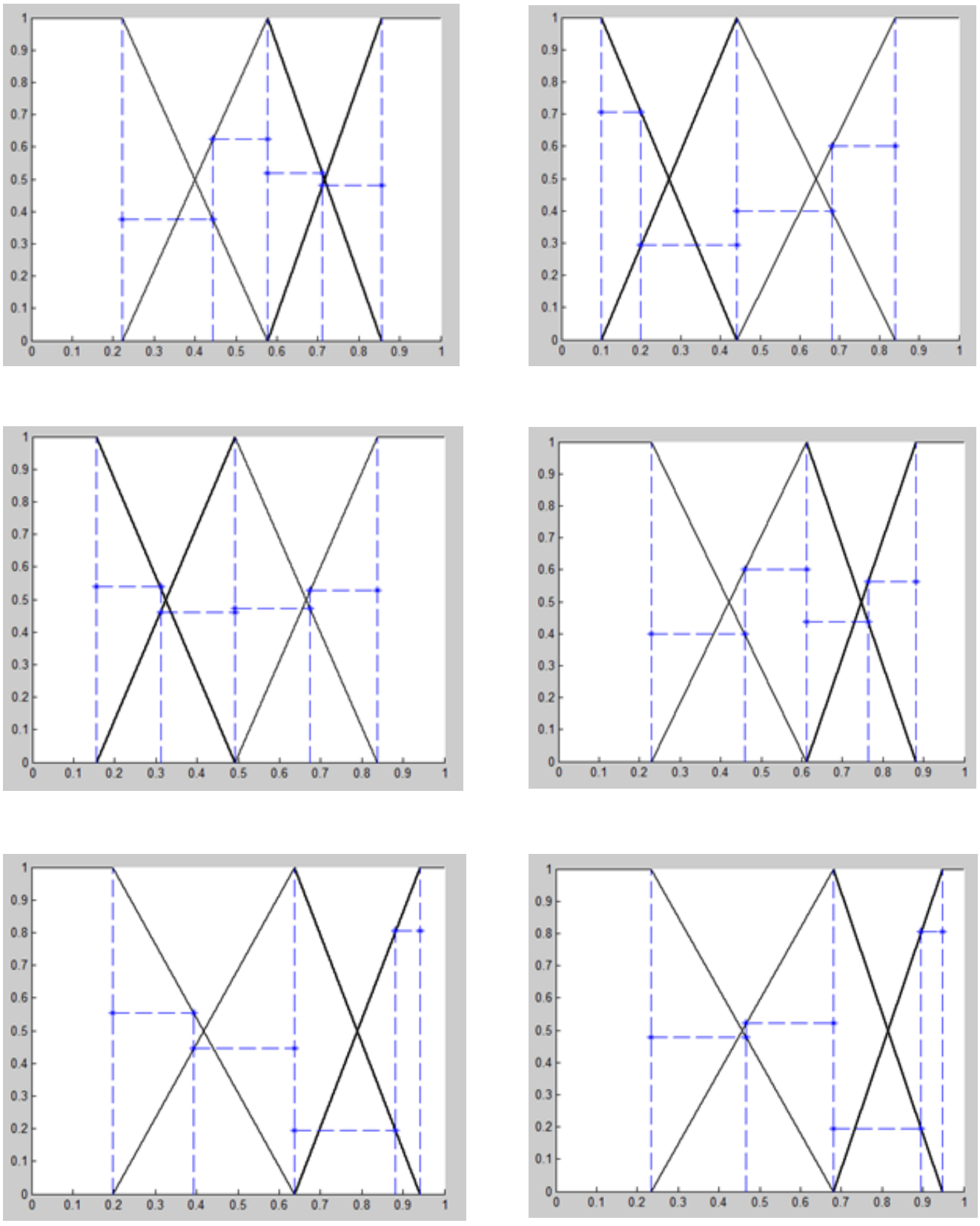


Рисунок 5.2 – Функции принадлежности с $\alpha$ - разрезами для шести предметов

Таблица 5.1 – Наблюдения из выборки данных

| № наб. | Пред. №1 | Пред. №2 | Пред. №3 | Пред.№ 4 | Пред. №5 | Пред. №6 |
|---|---|---|---|---|---|---|
| 1 | удовл. | отлично | отлично | отлично | отлично | отлично |
| 2 | удовл. | отлично | удовл. | хорошо | удовл. | удовл. |

Таблица 5.2 – Степени принадлежности наблюдений к кластерам

| Исследованные модели | № наб. | «удовл.» | «хорошо» | «отлично» |
|---|---|---|---|---|
| FCM | 1 | 0.13 | 0.37 | 0.5 |
| | 2 | 0.35 | 0.48 | 0.17 |
| BFCM | 1 | 0.14 | 0.26 | 0.6 |
| | 2 | 0.31 | 0.49 | 0.2 |
| MBFCM | 1 | 0 | 0.38 | 0.62 |
| | 2 | 0.56 | 0.44 | 0 |

### 5.1.2 Моделирование метода нечеткой кластеризации порядковых данных на основе совместного использования функций принадлежности и функции правдоподобия

Для эксперимента были взяты три размеченные выборки наблюдений из UCI репозитория [125]. Выборка Iris включает 150 наблюдений, разделённых на 3 класса, каждое наблюдение содержит 4 признака. Выборка Wine включает 178 векторов наблюдений, разделённых на 3 класса, каждое наблюдение содержит 13 признаков. Выборка Nursery включает 12958 наблюдений, разделённых на 4 класса, каждое наблюдение содержит 8 признаков, заданных в порядковой шкале. Для проверки предложенного метода выборки Iris и Wine были приведены к

порядковой шакале. Поскольку для каждой выборки существуют метки верной классификации, эффективность кластеризации измерялась в процентах точности относительно эталонного разбиения после дефаззификации.

Сравнивались результаты работы метода нечётких $c$ - средних при $\beta = 2$ (FCM), метода нечетких $c$ - средних для порядковых данных (FCMO) [70] и метода нечеткой кластеризации порядковых данных на основе совместного использования функций принадлежности и функции правдоподобия (LMFCM). В каждой ячейке таблицы 5.3 приведён средний, минимальный и максимальный результат для серии из 50 экспериментов.

Таблица 5.3 – Сравнение точности кластеризации на разных выборках

| Исследованные модели | Iris | | | Wine | | | Nursary | | |
|---|---|---|---|---|---|---|---|---|---|
| | avg | max | min | avg | max | min | avg | max | min |
| FCM ($\beta$ =2) | 70 | 75 | 35 | 69 | 74 | 33 | 62 | 67 | 36 |
| FCMO | 83 | 93 | 58 | 70 | 73 | 45 | 71 | 79 | 43 |
| LMFCM | 85 | 95 | 55 | 71 | 75 | 41 | 74 | 77 | 45 |

Результаты показывают, что эффективность LMFCM выше и стабильнее, чем FCM и FCMO, однако зависят от полноты обучающей выборки.

### 5.1.3 Моделирование адаптивного метода нечеткой кластеризации порядковых данных на основе порядково - цифрового отображения

Ввиду того, что точность и иные меры качества кластеризации адаптивных алгоритмов идентичны их пакетным аналогам, наиболее интересным для экспериментального исследования была признана скорость самообучения системы. В качестве меры модельного времени, как минимальный общий квант пакетных и адаптивных форм методов, принято количество проходов (эпох, итераций) по всей доступной выборке наблюдений. В проведённой серии экспериментов

тестировалось время, за которое система достигает заданной точности кластеризации. Для тестирования использовалась традиционная выборка «Wine» [125]. С каждым из рассмотренных алгоритмов была проведена серия из 50 экспериментов. Каждый эксперимент включал 25 итераций обучения метода. Изначально инициализированный случайно, на каждой итерации метод самообучался на обучающем множестве, включающем 70% выборки, после чего измерялась точность кластеризации на всей выборке. На графиках на рисунке 5.3 приведены средняя точность кластеризации каждого метода в зависимости от количества проходов по выборке.

Проводилось сравнение метода нечётких $c$ - средних (FCM), пакетного метода нечетких $c$ - средних на основе порядково-цифрового отображения (ONMFCM) и адаптивного метода рекуррентной нечеткой кластеризация на основе порядково-цифрового отображения (RONMFCM).

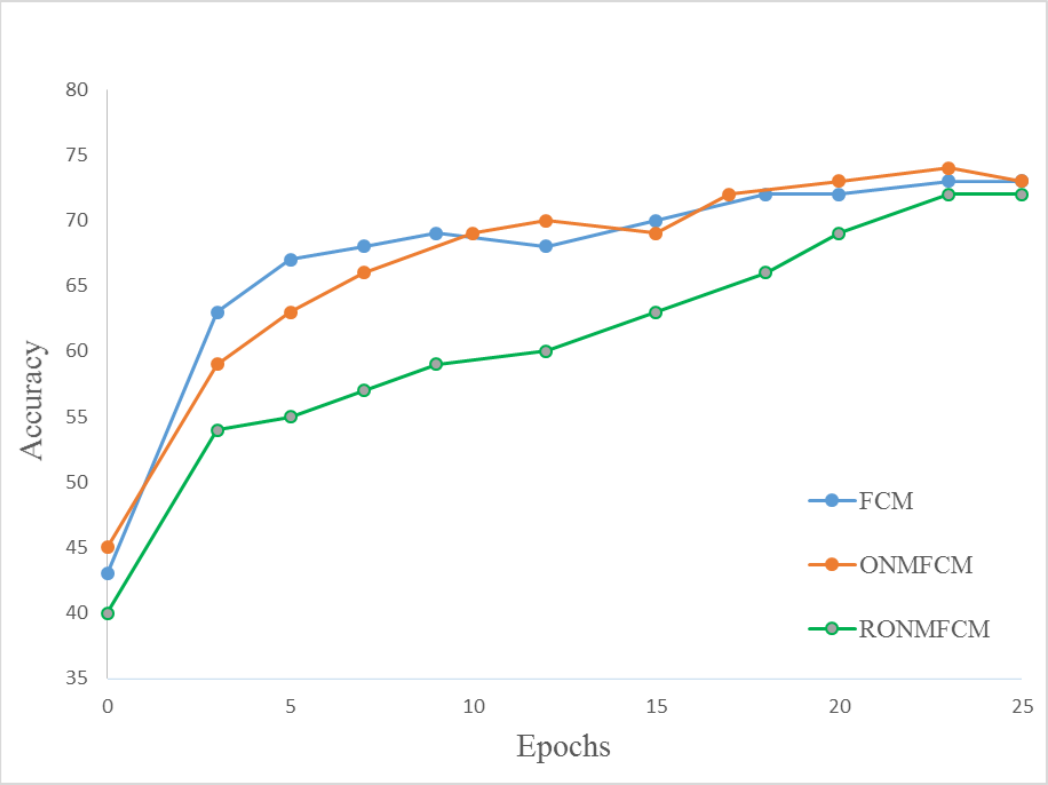



Рисунок 5.3 – Точность кластеризации в зависимости от числа итераций по выборке

Отметим, что из-за большого количества вычисляемых параметров, адаптивный вариант метода, требует большего количества наблюдений для качественной настройки, чем его пакетные формы. Однако, имея намного более гибкие возможности адаптации к входящим наблюдениям, метод рекуррентной нечеткой кластеризация на основе порядково - цифрового отображения сохраняет монотонность роста качества кластеризации с числом принятых наблюдений, что особенно важно при необходимости обработки данных в последовательном режиме.

### 5.1.4 Моделирование адаптивного метода робастной нечеткой кластеризации порядковых данных на основе меры схожести

Отдельная серия экспериментов была проведена для проверки робастности адаптивного метода кластеризации на основе меры схожести, описанного в разделе 2.10. Для этого была создана искусственная выборка из 3 непересекающихся кластеров, к которым было добавлено 20% выбросов. Для получения порядковых атрибутов, сгенерированная выборка была проранжирована на неоднородно разбитых интервалах от 1 до 7. Выборка представлена на рисунке 5.4, отдельно во врезке увеличен центр выборки с отсечёнными выбросами.

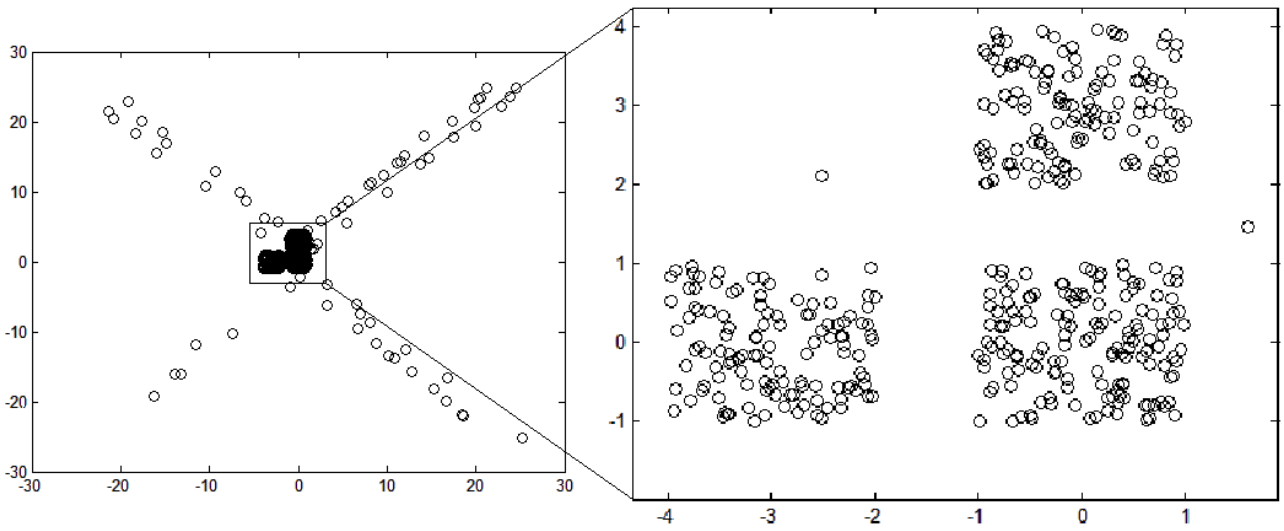
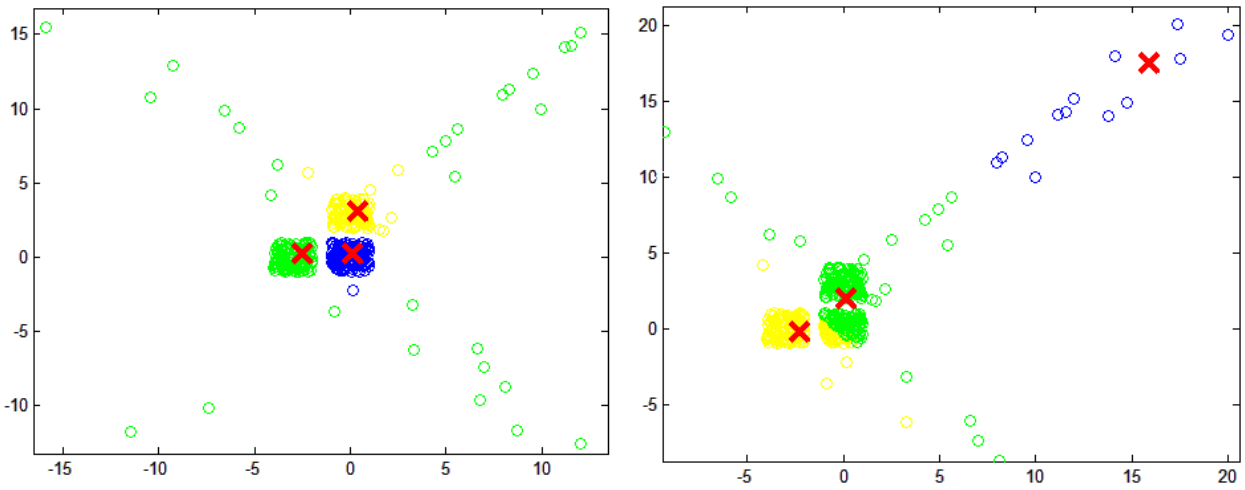



Рисунок 5.4 – Исходная выборка данных

На рисунке 5.5 можно увидеть результат кластеризации методом FCM, известного своей неустойчивостью к выбросам, и разбиение робастным методом кластеризации на основе меры схожести. Из эксперимента видно, что в случае применения робастного метода на результат не влияет наличие выбросов, в то время как классические методы гиперчувствительные к наблюдениям, находящимся далеко от всех прототипов (центроидов).




а б

Рисунок 5.5 – Сравнение робастности методов: а – разбиение методом FCM; б – результат робастной кластеризации на основе меры схожести

## 5.2 Моделирование возможностного метода нечеткой кластеризации категориальных данных с использованием частотных прототипов и мер несходства

Для эксперимента были взяты три размеченные выборки наблюдений, содержащие категориальные характеристики, из UCI репозитория [125]. Выборка Mushroom включает 8124 наблюдений, разделённых на два класса – «съедобный» и «несъедобный», каждое наблюдение содержит 22 атрибута. Атрибуты отображали такие параметры, как форма шляпки гриба («колоколообразная», «конусообразная», «плоская»), цвет шляпки («коричневая», «серая», «розовая»), место произрастания («лес», «луг», «город») и т.п. Выборка SPECT Heart описывает болезнь сердца на основе однофотонной эмиссионной компьютерной томографии, она состроит из 267 наблюдений, каждое из которых содержит 22

категориальных атрибута. Выборка Tic-Tac-Toe Endgame описывает возможные конфигурации на доске в конце игры. Она содержит 958 наблюдений, каждое из которых представлено 9-ю атрибутами. Во всех выборках пропущенных значений нет. Поскольку для выборок существуют метки верной классификации, эффективность кластеризации измерялась в процентах точности относительно эталонного разбиения после дефаззификации.

Сравнивались результаты работы метода ROCK, метода нечётких $c$ - средних на основе мер несходства (FCMK) и возможностного метода нечетких $c$ - средних для категориальных данных с использованием мер несходства (PFCMK). В таблицы 5.4 приведён средний, минимальный и максимальный результат для серии из 50 экспериментов.

Таблица 5.4 – Сравнение точности кластеризации на выборках данных

| Исследованные модели | Mushroom | | | SPECT Heart | | | Tic-Tac-Toe | | |
|---|---|---|---|---|---|---|---|---|---|
| | avg | max | min | avg | max | min | avg | max | min |
| ROCK | 85 | 87 | 81 | 84 | 87 | 73 | 95 | 97 | 92 |
| FCMK | 87 | 90 | 83 | 83 | 86 | 75 | 95 | 96 | 90 |
| PFCMK | 88 | 92 | 82 | 86 | 88 | 76 | 96 | 98 | 88 |

Из таблицы 5.4 видно, что у метода PFCMK выше точность кластеризации выборок. Хотя метод ROCK также имел высокие показатели точности, однако на выборке с большим количеством наблюдений, такой как Mushroom, он дольше обрабатывал данные по сравнению с остальными методами.

## 5.3 Моделирование нейро-фаззи системы на основе двойного нео-фаззинейрона

Чтобы продемонстрировать эффективность нейро-фаззи системы на основе двойного нео-фаззи нейрона и ее метода обучения в качестве тестовой выборки

была выбрана выборка Nursery из UCI репозитория [125], представленная 12958 наблюдениями по 8 атрибутов каждое, заданных в порядковой шкале. Выборка была поделена на обучающую и тестовую в пропорции 70/30.

Сравнение проводилось между следующими архитектурами: нео-фаззи нейрон и двойной нео-фаззи нейрон. С каждой из рассмотренных систем была проведена серия из 50 экспериментов. Измерялась точность кластеризации на обучающей и тестовой выборках, а также скорость обучения системы. Результаты работы нейро-фаззи систем представлены в таблице 5.5.

Таблица 5.5 – Результаты кластеризации выборки данных Nursery

| Исследованные системы | Время обучения | Ошибка обучения | Ошибка тест. |
|---|---|---|---|
| DNFN | 1,45 с | 0,0121 | 0,0127 |
| NFN | 1,30 с | 0,0167 | 0,0171 |

На основании эксперимента можно сделать вывод, что обе системы отличаются вычислительной простотой. Нейро-фаззи система на основе двойного нео-фаззи нейрона требует больше времени для настройки весовых коэффициентов, чем система на основе нео-фаззи нейрона. Однако, точность кластеризации порядковых данных у DNFN выше, чем у NFN. Результаты экспериментов позволяют говорить о высоком качестве кластеризации данных предложенной системой.

## 5.4 Решение задачи анализа клиентской базы данных предприятия

В современной экономической ситуации можно отметить возрастание уровня конкуренции среди украинских компаний и высокой изменчивостью клиентских предпочтений. Поиск новых способов эффективного управления компанией является одной из важнейших задач современного бизнеса. Отметим, что

предприятия с высоким уровнем клиентской лояльности имеют больше шансов на успешную деятельность в условиях кризиса.

Таким образом, внедрение методов управления бизнесом, направленных на понимание потребностей своих клиентов и повышение эффективности работы с ними [126], является актуальным на текущий момент.

Клиентоориентированное ведение бизнеса позволяет компании повысить свой доход за счет оптимизации операционных затрат и увеличения выручки от существующей клиентской базы.

Для анализа данных существует множество статистических пакетов, основное внимание в которых уделяется классическим методикам таким, как, регрессионный, корреляционный, факторный анализ и т.п. Однако, работа с данными пакетами требует специальной подготовки пользователя, кроме того они слишком «тяжеловесные» для повсеместного применения в бизнесе.

Следует отметить, что большинство статистических методов использует усредненные характеристики выборки, что при решении реальных сложных задач часто приводит к искажению результатов анализа. Наиболее мощными и распространенными статистическими пакетами являются STATGRAPHICS, SAS, SPSS, STATISTICA и т.п.

В данном разделе описано решение задачи анализа клиентской базы фирмы ООО «Южэлектропроект», специализирующейся на реализации электротехнического оборудования для предприятий угледобывающей, химической, энергетической, металлургической отраслей промышленности, а также для предприятий транспортного сектора.

Анализ данных основан на методе кластеризации, описанном в разделе 3 данной работы, и направлен на выявление скрытых закономерностей в поведенческой истории клиентов, с целью проведения персонализированной маркетинговой политики среди них.

Исходные данные для анализа были представлены клиентской базой предприятия, содержащей сведения об осуществленных сделках за 2015 г. За этот период предприятием было совершено порядка 6 тыс. сделок, а количество

активных клиентов составило 680. Информация в базе хранится в виде данных о клиентах и сделках, совершаемых клиентами. Данные о клиентах представлены такими характеристиками, как идентификационный номер клиента, название фирмы, вид (юридическое или физическое лицо), сфера деятельности, географический регион расположения. Информация о сделках представлена следующими характеристиками: идентификационный номер клиента, дата совершения сделки, состояние сделки (открыта, отказ, успех), причина отказа в случае неуспешной сделки, источник информации о приобретаемом товаре, купленный товар, сумма сделки, дата оплаты. Так как данные заданы в категориальной шкале, а степень пересечения кластеров неизвестна, то для анализа данных целесообразно применить метод возможностной нечеткой кластеризация категориальных данных с использованием частотных прототипов и мер несходства, описанный в разделе 3.6 данной работы.

В результате проведенного исследования были выявлены следующие кластеры (рис. 5.6):

– кластер 1 (5%). Значимые клиенты компании, переодически совершающие сделки на крупную сумму;

– кластер 2 (52%). Клиенты со средними и низкими суммами чеков, являющиеся постоянными клиентами;

– кластер 3 (34%). Клиенты, совершившие одноразовую сделку на маленькую или среднюю сумму чека;

– кластер 4 (9%). Клиенты, совершившие одноразовую сделку на крупную сумму чека за анализируемый период.

Каждый кластер был проанализирован по ряду признаков таких, как оборот денежных средств в кластере, количество клиентов в кластере, общее количество сделок в кластере, количество сделок на одного клиента в кластере и т.п.

На основе полученной информации была скорректирована ценовая политика компании, введена дифференцированная система бонусов и скидок для клиентов в зависимости от кластера, в котором он находился. Для постоянных клиентов проведена адресная рассылка с перечнем дополнительных услуг. Данные

мероприятия позволили увеличить доход компании на 3% за первый квартал 2016 г. по сравнению с этим же периодом в 2015 г.

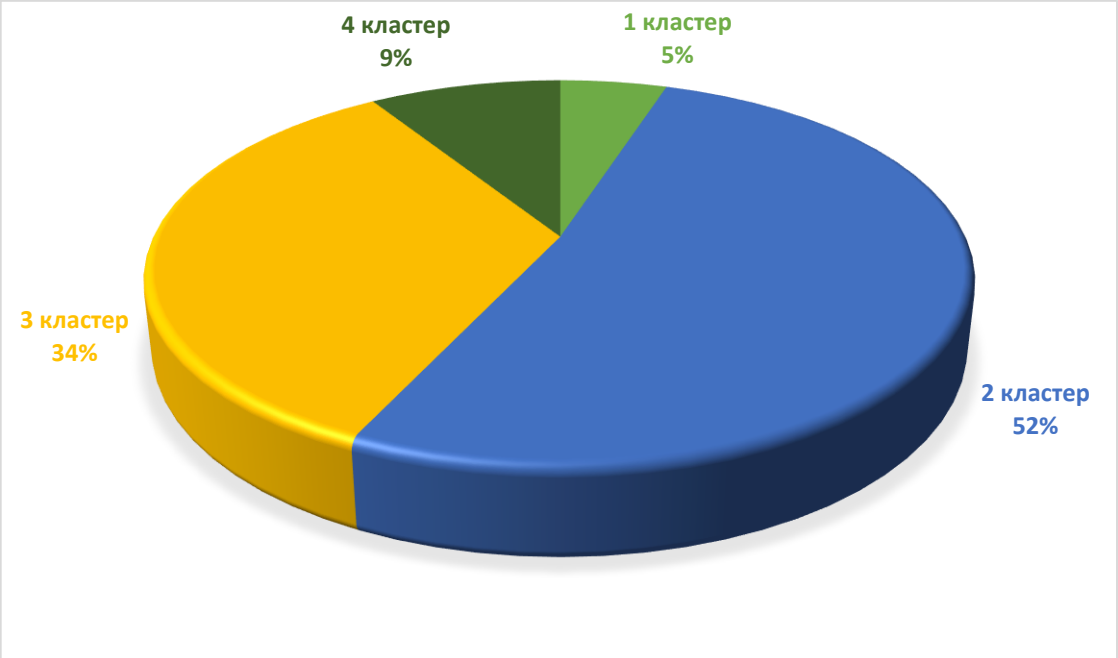



Рисунок 5.6 – Диаграмма кластеризации клиентской базы

## 5.5 Решение задачи автоматизированной обработки термограмм при диагностике электрооборудования

В данном разделе описано решение задачи автоматической обработки термограмм при диагностике электрооборудования для ООО НПФ «Мидиэл», путём их сегментации с последующей кластеризацией. При обработке данных использовался метод, описанный в разделе 2 данной работы.

Диагностика и выявление дефектов электрооборудования является актуальной проблемой на сегодняшнее время. Любое электрооборудование содержит в себе большое число различных элементов, контактов и соединений. Их ослабление, окисление, дисбаланс нагрузки и перегрузки, перегорание изоляции проводки и т.п. приводит не только к неправильной работе оборудования, но также может стать причиной аварии.

Своевременное устранение дефектов электрооборудования позволяет увеличить сроки его эксплуатации и избежать затрат на устранение последствий аварии.

Тепловизионная диагностика, относящаяся к методам теплового неразрушающего контроля [127, 128], дает возможность определять неисправности электрооборудования на ранних стадиях их развития (рис. 5.7).

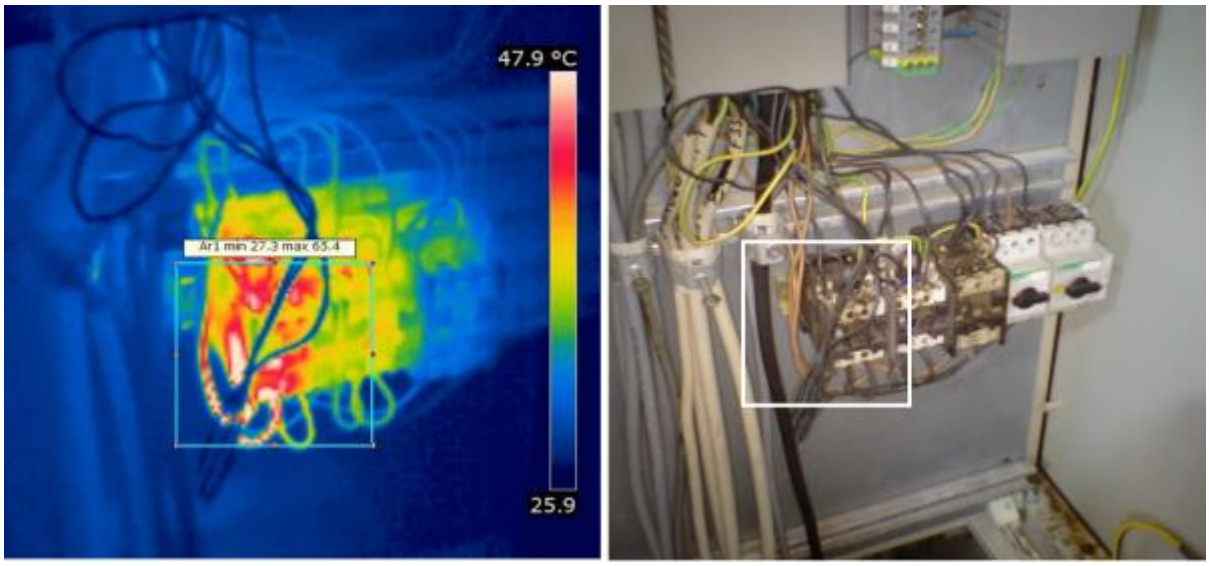
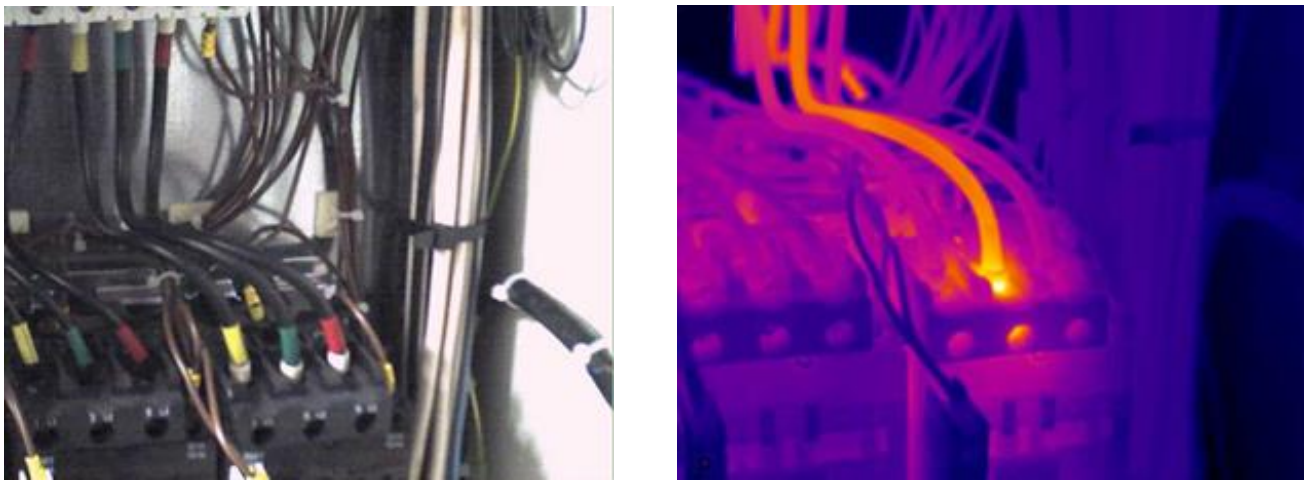



Рисунок 5.7 – Пример тепловизорной диагностики фаз в аппарате защиты и контроля движения шахтной подъемной установки

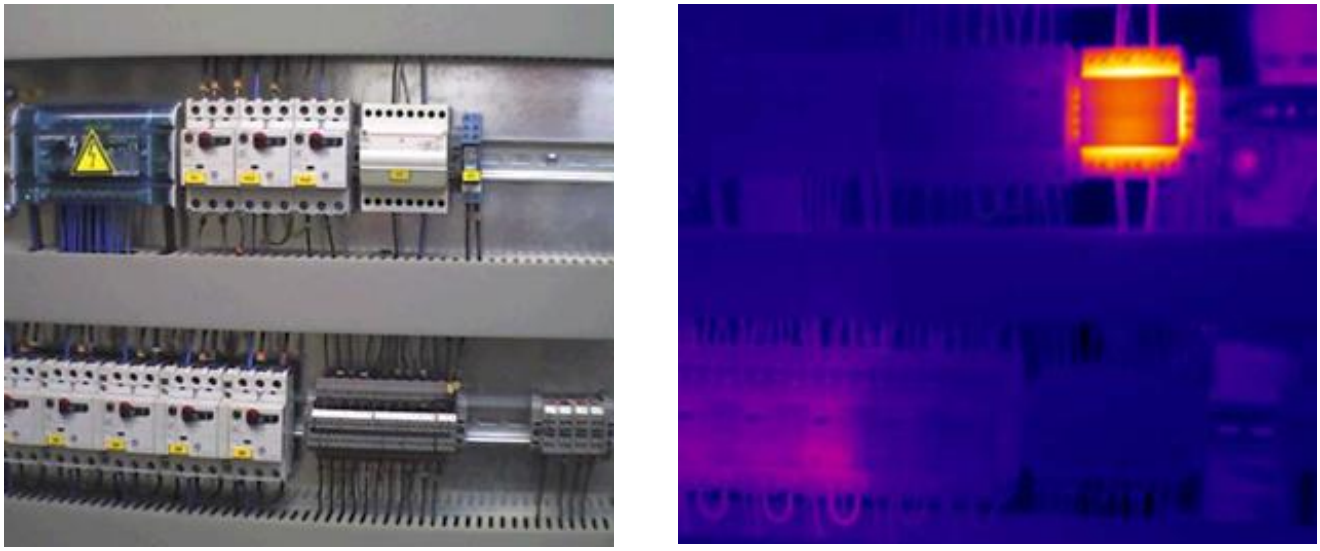
В ее основе лежит анализ температурных полей, получаемых с помощью портативных инфракрасных камер – тепловизоров. Данная методика позволяет проводить диагностику при работающем оборудовании, что дает возможность получать более полную картину существующих и развивающихся дефектов, которые невооруженным глазом заметить проблематично.

С помощью тепловизионной диагностики могут быть определены такие виды дефектов, как: нарушение герметизации элементов электрооборудования, перегрев контактных соединений, ухудшение состояния внутренней изоляции обмоток, пробой секций элементов, дефекты вводов и т.п.

Примеры возможных дефектов, выявляемых при тепловизионной диагностике, приведены на рисунке 5.8.



а



б

Рисунок 5.8 – Примеры дефектов, выявляемых при тепловизионной диагностике: а – превышение допустимой температуры провода одной фазы; б – перегрев клеммной коробки

Преимуществами данного метода являются:

- проведение диагностики электрооборудования в рабочем режиме, т.е. без снятия напряжения;
- обнаружение дефектов на ранних этапах их возникновения и развития;
- прогноз возникновения дефектов;
- небольшие трудозатраты на производство диагностики;
- безопасность рабочих при проведении диагностики;
- отсутствие необходимости организовывать отдельное рабочее место;

– возможность выполнить большой объем работы за сравнительно короткий промежуток времени;

– высокая производительность и информативность диагностики.

Данные для анализа представляют собой набор термограмм порядка 5 тысяч изображений. Примеры изображений представлены на рисунке 5.9.

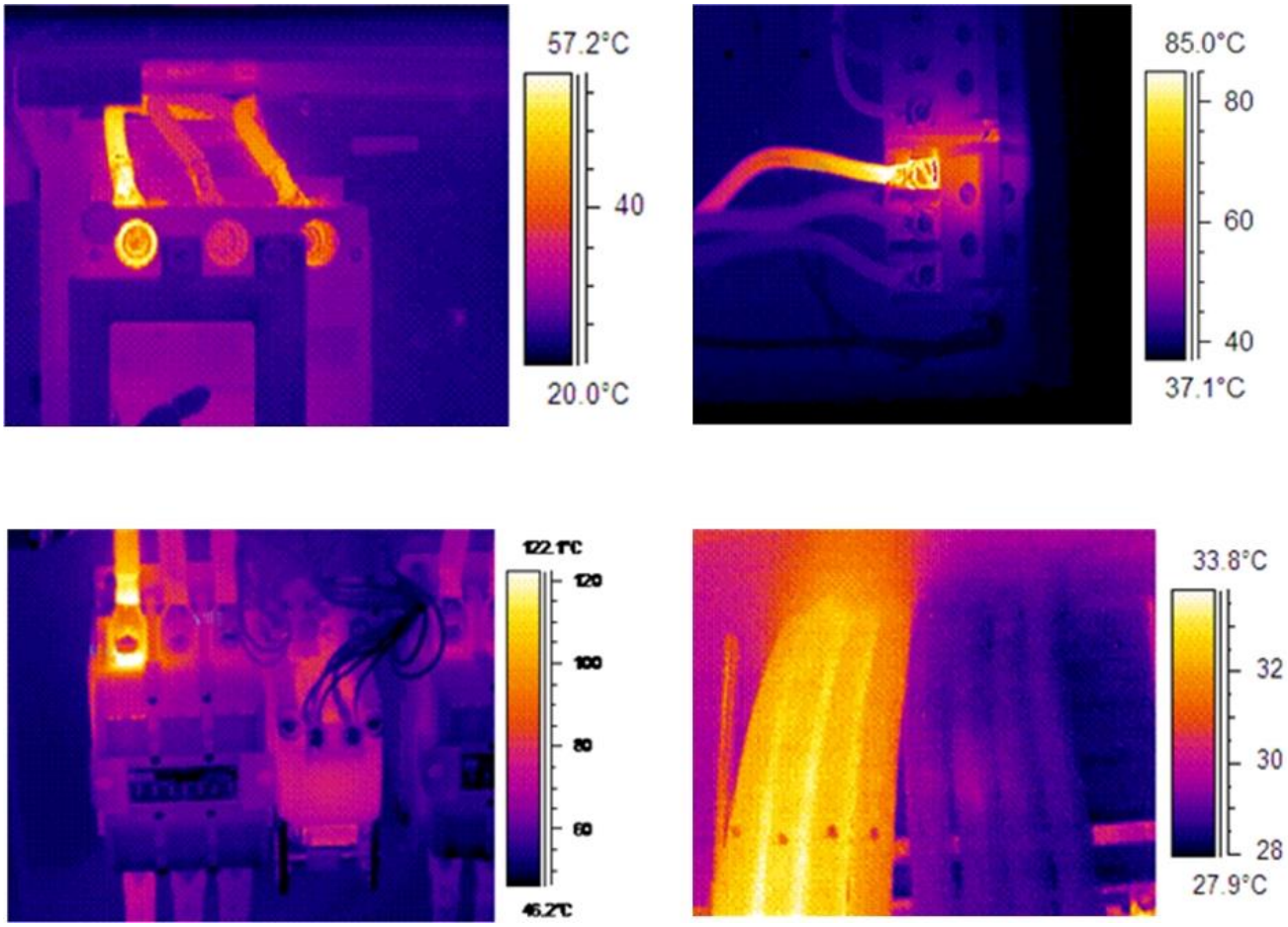
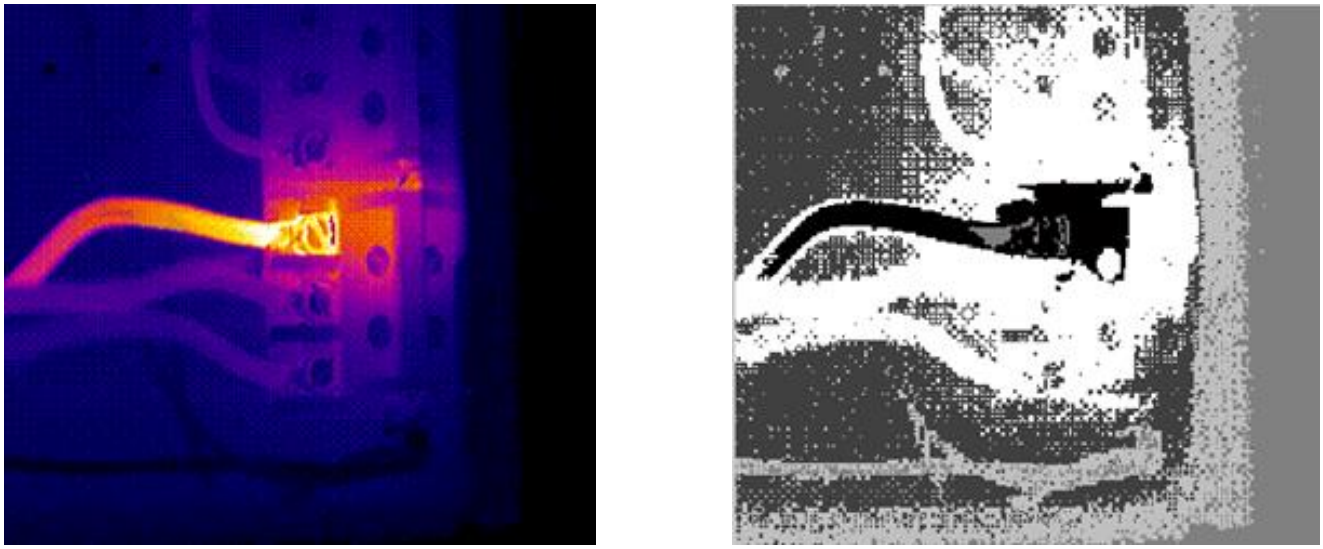



Рисунок 5.9 – Примеры термограмм

Постановка задачи подразумевает разбиение исходных данных на кластеры с целью выявления дефектов электрооборудования и выработки мероприятий по их устранению.

Задача решается с помощью двухэтапной кластеризации каждого изображения. На первом этапе осуществляется сегментация изображения. Для этого изображение представляется в цветовом пространстве $L*a*b*$. В данном пространстве $'L*'$ передает информацию об интенсивности цвета, $'a*'$ отображает цвет пикселя на красно-зеленой оси, а $'b*'$ – на голубо-желтой оси. После

пикселизации изображение представляет собой двумерную выборку, которая в дальнейшем разбивается на кластеры с помощью метода кластеризации нечетких $c$ - средних (FCM) [23]. Пример сегментации термограммы приведен на рисунке 5.10.



а б

Рисунок 5.10 – Пример сегментации термограммы: а – исходный вариант термограммы; б – сегментированная термограмма

Далее происходит ранжирование каждого пикселя сегментированного изображения в соответствии с палитрой цветов термограммы, что позволяет повысить точность и информативность дальнейшей кластеризации. Кроме того, данный подход решает проблему, связанную с возможностью выбора множества различных палитр при тепловизорной съемке.

После ранжирования выявленных сегментов данные принимают порядковый вид. Для дальнейшей обработки применяется метод адаптивной нечеткой кластеризации порядковых данных на основе совместного использования функций принадлежности и функции правдоподобия, описанный в разделе 2.9 данной работы.

Реализация системы позволила выявлять дефекты электрооборудования на ранних этапах их развития, увеличила скорость и продуктивность его диагностики, уменьшила время простоя оборудования.

Выводы по разделу 5

1. Проведено имитационное моделирование метода нечеткой кластеризации порядковых данных на основе частотных прототипов и функций принадлежности.
2. Проведено имитационное моделирование метода нечеткой кластеризации порядковых данных на основе совместного использования функций принадлежности и функции правдоподобия. Предложенный метод формирования центроидов и преобразования порядковых данных в числовую шкалу позволяет более точно кластеризовать выборку порядковых данных.
3. Проведено имитационное моделирование адаптивного метода рекуррентной нечеткой кластеризации порядковых данных на основе порядково-цифрового отображения. Экспериментально доказано, что точность кластеризации порядковых данных в последовательном режиме с помощью предложенного метода сопоставима с его пакетными аналогами, хотя и требует большего времени для обучения.
4. Проведено имитационное моделирование адаптивного метода робастной нечеткой кластеризации порядковых данных на основе меры схожести. Результаты исследований показали устойчивость предложенного метода к наличию в данных выбросов по сравнению с его неробастными аналогами.
5. Проведено имитационное моделирование возможностного метода нечеткой кластеризации категориальных данных с использованием частотных прототипов и мер несходства. Предложенный метод показал высокую точность кластеризации и быстродействие при работе с большими объемами информации.
6. Проведено имитационное моделирование нейро-фаззи системы на основе двойного нео-фаззи нейрона для обработки порядковых данных. Результаты экспериментов показали, что предложенная архитектура и метод ее обучения

позволяют обрабатывать данные в порядковой шкале с высокой точностью, хотя и требуют больше времени для правильной настройки весовых коэффициентов.

7. Решена практическая задача анализа клиентской базы ООО «Южэлектропроект», что помогло в разработке более эффективной маркетинговой политики и повышении дохода предприятия.

8. Решена практическая задача автоматической обработки термограмм при диагностике электрооборудования для ООО НПФ "Мидиэл" с целью увеличения скорости и качества диагностики электрооборудования.

# ВЫВОДЫ

В диссертационной работе представлены результаты, являющиеся в соответствии с поставленной целью решением актуальной научно-технической задачи повышения эффективности нечеткой кластеризации и классификации данных в нечисловых шкалах. Полученные результаты имеют важное практическое значение для создания систем обработки данных в нечисловых шкалах в таких областях, как медицина, образование, социология и т.п. В ходе научных исследований получены такие результаты:

1. Выполнен обзор состояния проблемы интеллектуального анализа данных, которые заданы в нечисловых шкалах. Выявлены недостатки существующих методов: невозможность функционировать в on-line режиме и ограниченность методов трансформации лингвистических характеристик в цифровую шкалу.

2. Впервые предложен метод нечеткой кластеризации данных, заданных в порядковой шкале, с фаззификацией исходных данных на основе частоты встречаемости характеристик в выборке, что позволило повысить точность кластеризации и обрабатывать данные, не подчиняющиеся нормальному распределению.

3. Впервые предложен адаптивный метод рекуррентной нечеткой кластеризации данных, заданных в порядковой шкале, на основе их отображения в числовую шкалу, что позволило обрабатывать порядковые данные в on-line режиме.

4. Получил развитие метод нечеткой кластеризации порядковых данных путем совместного использования функций принадлежности и функции правдоподобия, что позволило обрабатывать данные, не связанные с гипотезой нормальности распределения. Метод фаззификации порядковых данных и способ определения условной вероятности появления конкретных наблюдений в каждом кластере позволяют быстро и точно кластеризовать выборку.

5. Усовершенствованы методы робастной нечеткой кластеризации порядковых данных путём введения критерия специального вида, подавляющего выбросы. Это позволило улучшить качество кластеризации порядковых данных, содержащих выбросы, в on-line режиме.

6. Получил дальнейшее развитие метод возможностной нечеткой кластеризации массивов категориальных данных путем использования частотных прототипов и мер несходства, что позволило преодолеть недостатки классических методов такие, как «проклятье размерности» и «концентрация норм» и повысить точность кластеризации данных.

7. Улучшена нейро-фаззи система на основе нео-фаззи нейрона для классификации порядковых данных путем введения дополнительного выходного слоя, что позволило решить задачу классификации порядковых данных.

8. Проведённое имитационное моделирование предложенных моделей и методов показало их преимущества по сравнению с существующими методами в задачах обработки данных в нечисловых шкалах. Решены актуальные практические задачи интеллектуального анализа данных. Результаты исследований внедрены в ООО «Южэлектропроект» и ООО НПФ «Мидиэл», что подтверждено соответствующими актами.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Jain, A.K. Algorithms for Clustering Data/ A.K. Jain, R.C. Dubes. – Englewood Cliffs, N.J.:Prentice Hall. – 1988. – 318 p.

2. Kaufman, L. Finding Groups in Data: An Introduction to Cluster Analysis / L. Kaufman, P.J. Rousseeuw. – N.Y.: John Wiley & Sons, Inc. – 1990. – 342 p.

3. Han, J. Data Mining: Concepts and Techniques / J. Han, M. Kamber. – San Francisco: Morgan Kaufmann. – 2006. – 800 p.

4. Gan, G. Data Clustering: Theory, Algorithms, and Applications / G. Gan, C. Ma, and J. Wu. – Philadelphia:SIAM. – 2007. – 466 p.

5. Abonyi, J. Cluster analysis for data mining and system identification / J. Abonyi, B. Feil. – Basel: Birkhäuser. – 2007. – 303p.

6. Olson, D.L. Advanced Data Mining Techniques / D.L. Olson, D. Dursun. – Berlin:Springer. – 2008. – 180 p.

7. Aggarwal, C.C. Data clustering: algorithms and applications / C.C. Aggarwal, C.K. Reddy. – Boca Raton:CRC Press. – 2014. – 648 p.

8. Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – N.Y.: Plenum. – 1981. – 272 p.

9. Hoeppner, F. Fuzzy-Clusteranalyse / F. Hoeppner, F. Klawonn, R. Kruse. – Braunschweig: Vieweg, 1997. – 280 S.

10. Hoeppner F. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition / F. Hoeppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester: John Wiley & Sons, 1999. – 300 p.

11. Бодянский, Е.В. Нечеткая кластеризация данных, заданных в порядковой шкале / Е.В. Бодянский, В.А. Опанасенко (В.А. Самитова), А.Н. Слипченко // Системы обработки информации. – 2007. – 4(62). – С. 5 - 9.

12. Бодянский, Е.В. Нечеткая кластеризация данных в порядковой шкале на основе совместного использования функций принадлежности и правдоподобия / Е.В. Бодянский, В.А. Самитова // Сборник научных работ ХУПС. – 2010. – 3(25). – С. 91 - 95.

13. Самитова, В.А. Отображение порядковых характеристик в цифровую шкалу на основе нечеткой кластеризации / В.А. Самитова // Системы обработки информации. – 2015. – 7(132). – С. 107 - 110.

14. Бодянский, Е.В. Нечеткая классификация данных в ранговой шкале на основе двойного нео-фаззи нейрона / Е.В. Бодянский, В.А. Самитова // Восточно-Европейский журнал передовых технологий. – 2008. – 4/2 (34). – С. 4 - 7.

15. Bodyanskiy, Ye. Robust Fuzzy Data Clustering In An Ordinal Scale Based On A Similarity Measure / Ye. Bodyanskiy, O. Tyshchenko, V. Samitova // International Journal of Reseach in Engineering and Science (IJRES). – 2014. – 2(4). – P. 21 - 25.

16. Бодянский, Е.В. Возможностная нечеткая кластеризация массивов категориальных данных с использованием частотных прототипов и мер несходства / Е.В. Бодянский, В.А. Самитова // Бионика интеллекта. – 2016. – 1(82). – С. 72 - 75.

17. Бешелев, С.Д. Математико-статистические методы экспертных оценок / С.Д. Бешелев, Ф.Г. Гурвич. – М.: Статистика. – 1980. – 263 с.

18. Бешелев, С.Д. Экспертные оценки / С.Д. Бешелев, Ф.Г. Гурвич. – М: Наука. – 1973. – 161 с.

19. Zadeh, L. Fuzzy sets / L. Zadeh // Information and Control. – 1965. – 8. – P. 338 – 353.

20. Борисов, А.Н. Обработка нечеткой информации в системах принятия решений / А.Н. Борисов, А.В. Алексеев, Г.В. Меркурьева. – М: Радио и связь. – 1989. – 304 с.

21. Литвак, Б.Г. Экспертная информация. Методы получения и анализа / Б.Г. Литвак. – М.: Радио и связь. – 1982. – 84 с.

22. Малышев, Н.Г. Нечеткие модели для экспертных систем в САПР / Н.Г. Малышев, Л.С. Берштейн, А.В. Боженюк. – М.: Энергоатомиздат. – 1991. – 136 с.

23. Bezdek, J.C. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J. Keller, R. Krishnapuram, N. Pal. – The Handbooks of Fuzzy Sets. – Kluwer, Dordrecht, Netherlands: Springer. – 1999. – 4. – 776 p.

24. Mendel, J. M. Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions / J.M. Mendel. – NJ: Prentice Hall. – 2001. – 555 p.

25. Steeb, W.-H. The Nonlinear Workbook: Chaos, Fractals, Cellular Automata, Genetic Algorithms, Gene Expression Programming, Support Vector Machine, Wavelets, Hidden Markov Models, Fuzzy Logic with C++, Java and SymbolicC++ Programs, 5th Edition / W.-H. Steeb. – Singapore:World Science Publ. – 2011. – 644 p.

26. Trillas E. Fuzzy Logic: An Introductory Course for Engineering Students / E. Trillas, L. Eciolaza. – Springer. – 2015. – 204 p.

27. Часовских, А. Обзор алгоритмов кластеризации данных / А. Часовских // Хабрахабр. – [https://habrahabr.ru/post/101338/]. – (дата обращения: 25.11.2015).

28. MacQueen, Z.B. Some Methods of Classification and Analysis of Multivariate Observations/ Z.B. MacQueen // Berkely Symposium on Mathematical Statistics and Probability. – 1967. – Vol. 1. – P. 281 - 297.

29. Lloyd, S.P. Least Squares Quantization in PCM / S.P. Lloyd // IEEE Transactions on Information Theory. – 1982. – Vol. IT-28. – P. 129 - 137.

30. Bradley, P.S. Clustering via Concave Minimization / P.S. Bradley, O.L. Mangasarian, W.N. Street // Advances in Neural Information Processing Systems. – 1997. – Vol. 9. – P. 368 - 374.

31. Pandit, V. Local Search Based Approximation Algorithms The k-median problem / V. Pandit, N. Garg, R. Khandekar, V. Arya. – The 2011 School on Approximability. – 2011. – 82 p.

32. Dempster, A.P. Maximum-Likelihood from Incomplete Data via the EM Algorithm/ A.P. Dempster, N.M. Laird, R.D.B. // Journal of the Royal Statistical Society. – 1977. – vol.B. – P. 1 - 38

33. Zhong, S. A Unified Framework for Model-based Clustering / S. Zhong, J. Ghosh // Journal of Machine Learning Research. – 2003. – Vol. 4. – P. 1001 - 1037.

34. Jang, J.-Sh.R. Neuro - Fuzzy and Soft Computing / J.-Sh.R. Jang, Ch.-T. Sun, E. Mizutani. – Upper Saddle River, NJ: Prentice Hall. – 1997. – 614 p.

35. Kohonen, T. Self-Organizing Maps / Kohonen T. - Berlin: Springer-Verlag. – 1995 – 501 p.

36. Gorshkov, Ye. New recursive learning algorithms for fuzzy Kohonen clustering network / Ye.Gorshkov, V. Kolodyazhniy, Ye. Bodyanskiy // Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Rapperswil, 21-24 June, 2009. - Switzerland, 2009. - P. 58 - 61.

37. Aggarwal C. C. Data Mining / C. C. Aggarwal. – Springer. – 2015. – 734 p.

38. Kubat M. An Introduction to Machine Learning / M. Kubat. – Springer. – 2015. – 291 p.

39. Lampropoulos A. S. Machine Learning Paradigms: Applications in Recommender Systems / A S. Lampropoulos, G. A. Tsihrintzis. – Springer. – 2015. – 125 p.

40. Piegorsch W. W. Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery / W. W. Piegorsch. – Wiley. – 2015. – 464 p.

41. Xu, R. Clustering / R. Xu, D. C. Wunsch. – IEEE Press Series on Computational Intelligence. – Hoboken, NJ:John Wiley & Sons, Inc. – 2009. – 370 p.

42. Borgelt, C. Prototype-based Classification and Clustering / C. Borgelt. – Magdeburg. – 2005. – 350 p.

43. Ball, G.H. A Clustering Technique for Summarizing Multivariate Data / G.H.Ball, D.J. Hall. // Behavioral Science. – 12(2). – 1967. – P.153 – 155.

44. MacQueen, J. On convergence of k-means and partitions with minimum average variance / J. MacQueen. – Ann. Math. Statist. – 1965. – 36. – 1084 p.

45. Cover, T.M. Estimates by the nearest-neighbor rule / T.M. Cover // IEEE Trans. on Information Theory. – 1968. – 14. – P. 50 – 55.

46. Ruspini, E.H. A New Approach to Clustering / E.H. Ruspini // Informationand Control. – San Diego, CA: Academic Press. – 1969 – 15(1). – P. 22 – 32.

47. Agresti, A. Categorical Data Analysis / A. Agresti. – Wiley Series in Probability and Statistics. – NY:Wiley-Interscience. – 2012. – 744 p.

48. Murray, J.S. Bayesian Gaussian copula factor models for mixed data / J.S. Murray, D.B. Dunson, L. Carin, Lucas J.E. // Journal of the American Statistical Association. – 2013. – 108(502). – P. 656 - 665.

49. McParland, D. Clustering ordinal data via latent variable models / D. McParland, I.C. Gormley // Studies in Classification, Data Analysis, and Knowledge Organization. – Berlin: Springer. 2013. – V. 547. – P. 127 - 135.

50. Gollini, I. Mixture of latent trait analyzers for model-based clustering of categorical data / I. Gollini, T.B. Murphy // Statistics and Computing. – 2014. – 24(4). – P. 569 - 588.

51. Biernacki, Ch. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm / Ch. Biernacki, J. Jacques // Statistics and Computing. – 2016. – Volume 26. – Issue 5. – P. 929–943.

52. Rojas, R. Neural Networks. A Systematic Introduction / R. Rojas. – Berlin:Springer-Verlag. – 1996. – 502 p.

53. Du K.-L., Neural Networks and Statistical Learning / K.-L. Du, M. N. S. Swamy. – London: Springer-Verlag. – 2014. – 824 p.

54. Haykin, S. Neural Networks. A Comprehensive Foundation / S. Haykin. – Upper Saddle River,N.J.: Prentice Hall, Inc. – 2004. – 842 p.

55. Lakhmi, C. Jain. Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Application / C. Jain Lakhmi, N.M. Martin. – CRC Press. – 1998. – 299 p.

56. Siddique N. Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing / N. Siddique, H. Adeli. – Wiley. – 2013. – 532 p.

57. Kruse R. Computational Intelligence / R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, P. Held. – Berlin: Springer. – 2013. – 488 p.

58. Jang, J.-S. R. ANFIS: Adaptive-network-based fuzzy inference systems / J.-S.R. Jang // IEEE Trans. Syst., Man, and Cybern. – 1993. – 23(3). – P. 665 – 685.

59. Jang, J.-S.R. Functional equivalence between radial basis function networksand fuzzy inference systems / J.-S.R. Jang, C.-T. Sun // IEEE Trans. on Fuzzy Systems. – 1993. – 4(1). – P. 156 – 159.

60. Vuorimaa, P. Fuzzy self-organizing maps / P. Vuorimaa // Fuzzy Sets and Systems. – 1994. – 66. – P. 223 – 231.

61. Nauck, D. A neuro-fuzzy approach to obtain interpretable fuzzy systems for function approximation / D. Nauck, R. Kruse // Proc. IEEE Int. Conf. on Fuzzy Systems. Anchorage, AK. – 1998. – 2. – P. 1106 – 1111.

62. Бодянский, Е.В. Об одном алгоритме обучения нейро-фаззи-предиктора /Е.В. Бодянский, В.В. Колодяжный // Адаптивні системи автоматичного управління: Зб. наук. праць. – 2000. – 3(23). – С. 29 - 36.

63. Agresti, A. Analysis of Ordinal Categorical Data / A. Agresti. – Wiley Series in Probability and Statistics. – NY:Wiley-Interscience. – 2010. – 424 p.

64. Butkiewicz, B.S. Robust fuzzy clustering with fuzzy data / B.S. Butkiewicz // Lecture Notes in Computer Science. – Heidelberg: Springer–Verlag. – 2005. – Vol. 3528. – P. 76 - 82.

65. Brouwer, R.K. Fuzzy set covering of a set of ordinal attributes without parameter sharing / R.K. Brouwer // Fuzzy Sets and Systems. – 2006. – 157. – №13. – P. 1775 – 1786.

66. Brouwer, R.K. A feedforward neural network for mapping vectors to fuzzy sets of vectors / R.K. Brouwer, W. Pedrycz // Proc. Int. Conf. on Artificial Neural Networks and Neural Information Processing ICANN/ICOMIP 2003. – Istanbul, Turkey, 2003. – P. 45 - 48.

67. Mahnhoon, L. Mapping of Ordinal Feature Values to Numerical Values through Fuzzy Clustering / L. Mahnhoon // IEEE Trans. on Fuzzy Systems. – 2008. – P. 732 - 737.

68. Chung, F.L. Fuzzy competitive learning / F.L. Chung, T. Lee // Neural Networks. – 1994. – 7(3) – P. 539 - 552.

69. Park, D.C. Gradient based fuzzy c-means (GBFCM) algorithm / D.C. Park, I. Dagher // IEEE Int. Conf. on Neural Networks. – 1984. – P. 1626 - 1631.

70. Mahnhoon, L. Likelihood based fuzzy clustering for data sets of mixed features / L. Mahnhoon, R. K. Brouwer // IEEE Symp. on Foundations of Comput. Intell. FOCI 2007. – 2007. – P. 544 - 549.

71. Dave R.N. Robust clustering methods: A unified view / R.N. Dave, R. Krishnapuram // IEEE Trans. on Fuzzy Systems. – 1997. – 5. – P. 270 - 293.

72. Tsuda, K. Sequential fuzzy cluster extraction and its robustness against noise / K. Tsuda, S. Senda, M. Minoh, K. Ikeda // Systems and Computers in Japan. – 1997. – 28. – P. 10 - 17.

73. Bodyanskiy, Ye. Computational intelligence techniques for data analysis / Ye. Bodyanskiy // Lecture Notes in Informatics. – Bonn, Germany. – 2005. – 72. – P. 15 - 36.

74. Bodyanskiy, Ye. Robust recursive fuzzy clustering algorithms / Ye. Bodyanskiy, Ye. Gorshkov, I. Kokshenev, V. Kolodyazhniy // Proc. East West Fuzzy Colloquium 2005. – Zittau/Goerlitz. – 2005. – P. 301 - 308.

75. Gorshkov, Ye. Robust recursive fuzzy clustering-based segmentation of biomedical time series / Ye. Gorshkov, I. Kokshenev, Ye. Bodyanskiy, V. Kolodyazhniy, O. Shilo // Proc. 2006 Int. Symp. on Evolving Fuzzy Systems. – Lancaster, UK. – 2006. – P. 101 - 105.

76. Kokshenev, I. Outlier resistant recursive fuzzy clustering algorithm / I. Kokshenev, Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy // Computational Intelligence: Theory and Applications. Advances in Soft Computing. – 2006. – 38. – P. 647 - 652.

77. Hoeppner, F. Fuzzy clustering of sampled functions / F. Hoeppner, F. Klawonn // Proc. 19-th Int. Conf. North American Fuzzy Information Processing Society (NAFIPS). – Atlanta, USA. – 2000. – P. 251 - 255.

78. Georgieva, O. A clustering algorithm for identification of single clusters in large data sets / O. Georgieva, F. Klawonn // Proc. 11-th East-West Fuzzy Coll. – Zittau/Goerlitz: HS. – 2004. – P. 118 - 125.

79. Winkler, R. M-Estimator induced Fuzzy Clustering Algorithms / R. Winkler, F. Klawonn, R. Kruse. // European Society for Fuzzy Logic and Technology – «les rencontres francophones sur la Logique Floue et ses Applications» (EUSFLAT-LFA). – Aix-les-Bains, France. – 2011. – P. 298 - 304.

80. Winkler, R. A new Distance Function for Prototype based Clustering Algorithms in High Dimensional Spaces / R. Winkler, F. Klawonn, R. Kruse. // Statistical

Models for Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. – Springer. – 2013. – P. 371 - 378.

81. Sepkovski, J.J. Quantified coefficients of association and measurement of similarity / J.J. Sepkovski // Int. J. Assoc. Math. – 1974. – 6(2). – P. 135 - 152.

82. Krishnapuram, R. The possibilistic C-means algorithm: Insights and recommendations / R. Krishnapuram, J.M. Keller // IEEE Trans. on Fuzzy Systems. – 1996. – 4. – P. 385 - 393.

83. Sudipto, G. ROCK: A Robust Clustering Algorithm for Categorical Attributes / G. Sudipto, R. Rajeev, S. Kyuseok // Proc. of The IEEE Int. Conf. on Data Engineering. – Sydney. – 1999. – P. 512 - 521.

84. Jaccard, P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines/ P. Jaccard // Bull. Soc. Vaudoise sci. Natur. – 1901. – V. 37 (140). – S. 241 - 272.

85. Huang, Zh. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values / Zh. Huang // Data Mining and Knowledge Discovery. – 1998. – 2. – №2. – P. 283 - 304.

86. He, Z. Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode / Z. He, S. Deng, X. Xu // Computational Intelligence and Security. – 2005. – Lecture Notes in Computer Science. – V.3801. – P. 157 - 162.

87. Lei, M. An improved k-means algorithm for clustering categorical data / M. Lei, P. He, Zh. Li // Communications and Computer. – 2006. – 3. – №8. – P. 20 - 24.

88. Huang, Zh. A fuzzy k-modes algorithm for clustering categorical data/ Zh. Huang, M.K. Ng // IEEE Trans on Fuzzy Systems. – 1999. – 7. – №4. – P. 446 - 452.

89. Kim, D.W. Fuzzy clustering of categorical data using fuzzy centroids / D.W. Kim, K.H. Lee, D. Lee // Pattern Recognition Letters. – 2004. – 25. – P. 1263 - 1271.

90. Lee, M. Fuzzy p-mode prototypes: A generalization of frequency-based cluster prototypes for clustering categorical objects / M. Lee // Computational Intelligence and Data Mining. – Nashville, TN. – 2009. – P. 320 - 323.

91. Bodyanskiy, Ye. Recursive Fuzzy Clustering Algorithms / Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan // Proc. 10th East–West Fuzzy Colloquium. – Zittau, Germany. – 2002. – P. 276 - 283.

92. Krishnapuram, R. A possibilistic approach to clustering / R. Krishnapuram, J. Keller // IEEE Trans. on Fuzzy Systems. – 1993. – 2. – №1. – P. 98 - 110.

93. Айзерман, М.А. Метод потенциальных функций в теории обучения машин. / М.А. Айзерман, Э.М. Браверман, Л.И. Розоноэр. – М.: Наука. – 1970. – 384 с.

94. Надарая, Э.А. О непараметрических оценках плотности вероятности и регрессии / Э.А. Надарая // Теория вероятностей и ее применение. – 1965. – 10. – № 1. – С. 199 - 203.

95. Варядченко, Т.В. Непараметрический метод обращения функций регрессии / Т.В. Варядченко, В.Я. Катковник // Стохастические системы управления. – Новосибирск: Наука. – 1979. – С. 4 - 14.

96. Parzen, E. On the estimation of a probability density function and the mode / E. Parzen // Ann. Math. Statist. – 1962. – 38. – P. 1065 - 1076.

97. Живоглядов, В.П. Непараметрические алгоритмы адаптации / В.П. Живоглядов, А.В. Медведев. – Фрунзе: Илим. – 1974. – 214 с.

98. Медведев, А.В. Адаптация в условиях непараметрической неопределенности / А.В. Медведев // Адаптивные системы и их приложения. – Новосибирск: Наука. – 1978. – С. 4 - 34.

99. Раудис, Ш.Ю. Оптимизация непараметрического алгоритма классификации / Ш.Ю. Раудис // Адаптивные системы и их приложения. – Новосибирск: Наука. – 1978. – С. 57 - 61.

100. Медведев, А.В. Непараметрические алгоритмы идентификации нелинейных динамических систем / А.В. Медведев // Стохастические системы управления. – Новосибирск: Наука. – 1979. – С. 15 - 22.

101. Hartman, E. J. Layered neural networks with Gaussian hidden units as universal approximations / E.J. Hartman, J.D. Keeler, J. Kowalski // Neural Computation. – 1990. – 2. – P. 210 - 215.

102. Park, J. Universal approximation using radial-basis-function networks / J. Park, I.W. Sandberg // Neural Computation. – 1991. – 3. – P. 246 - 257.

103. Leonard, J.A. Using radial basis functions to approximate a function and its error bounds / J.A. Leonard, M.A. Kramer, L.H. Ungar // IEEE Trans. on Neural Networks. – 1992. – 3. – P. 614 - 627.

104. Sunil, E.V.T. Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems / E.V.T. Sunil, C.Sh. Yung // IEEE Trans. on Neural Networks. – 1994. – 5. – P. 594 - 603.

105. Poggio, T. A Theory of Networks for Approximation and Learning / T. Poggio, F. Girosi. – A. I. Memo № 1140, C.B.I.P. Paper № 31. – Massachussetts Institute of Technology. – 1994. – 63 p.

106. Круглов, В.В. Нечеткая логика и искусственные нейронные сети / В.В. Круглов, М.И. Дли, Р.Ю. Голунов. – М.: Физматлит. – 2001. – 224 с.

107. Осовский, С. Нейронные сети для обработки информации/ С. Осовский. – М.:Финансы и статистика. – 2002. – 344 с.

108. Аксенов, С.В. Организация и использование нейронных сетей (методы и технологии) / С.В. Аксенов, В.Б. Новосельцев. – Томск: НТЛ. – 2006. – 128 с.

109. Батыршин, И.З. Нечеткие гибридные системы. Теория и практика / И.З. Батыршин. – М:ФИЗМАТЛИТ. – 2007. – 208 с.

110. Yamakawa, T. A Neo Fuzzy Neuron and Its Applications to System Identification and Prediction of the System Behavior / T. Yamakawa, E. Uchino, T. Miki, H. Kusanagi // Proc. 2-nd Int. Conf. on Fuzzy Logic and Neural Networks “IIZUKA-92”. – Iizuka, Japan. – 1992. – P. 477 - 483.

111. Uchino, E. Soft Computing Based Signal Prediction, Restoration, and Filtering / E. Uchino, T. Yamakawa // Ed. Da Ruan. Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks, and Genetic Algorithms. – Boston: Kluwer Academic Publishers. – 1997. – P. 331 - 349.

112. Miki, T. Analog Implementation of Neo-Fuzzy Neuron and Its On-board Learning / T. Miki, T. Yamakawa // Ed. N.E. Mastorakis. Computational Intelligence and Applications. – Piraeus: WSES Press. – 1999. – P. 144 - 149.

113. Howlett, R.J. Radial basis functions networks. Recent developments in theory and applications / R.J. Howlett, L.C. Jain. − Berlin: Springer. – 2001. − 318 p.

114. Chen, S. Recursive hybrid algorithm for nonlinear system identification using radial basis functions networks / S. Chen, S.A. Billings, P.M. Grant // Int. J. Control. − 1992. − Vol. 55. – № 5. − P. 1051 - 1070.

115. Sugeno, M. An Introductory Survey of Fuzzy Control / M. Sugeno // Information Sciences. – N.Y.: ACM Press. – 1985. – Vol. 36. – № 3. – P. 59 - 83.

116. Bodyanskiy, Ye. Evolving network based on double neo-fuzzy neurons / Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, P. Otto// Proc. 52nd Int. Scientific Coll. «Computer Science Meets Automation". – TU Ilmenau (Thuer). – 2007. – P. 35 - 40.

117. Лбов, Г.Н. Методы обработки разнотипных экспериментальных данных / Г.Н. Лбов. – Новосибирск: Наука. – 1985. – 160 с.

118. Бодянский, Е.В. Многошаговые оптимальные упредители многомерных нестационарных стохастических процессов / Е.В. Бодянский, И.П. Плисс, Т.В. Соловьева // Доклады АН УССР. – 1986. – Сер. А. – №12. – С. 47 - 49.

119. Kaczmarz, S. Approximate solution of systems of linear equations / S. Kaczmarz // Int. J. Control. – 1993. – 53. – P. 1269 - 1271.

120. Widrow, B. Adaptive switching circuits / B. Widrow, Jr. M.E. Hoff // 1960 IRE Western Electric Show and Connection Record. – 1960. – Part 4. – P. 96 – 104.

121. Goodwin, G.C. Discrete time stochastic adaptive control / G.C. Goodwin, P.J. Ramadge, P.E. Caines // SIAM J. Control and Optimization. – 1981. – 19. – P. 829 - 853.

122. Goodwin, G.C. A globally convergent adaptive predictor / G.C. Goodwin, P.J. Ramadge, P.E. Caines // Automatica. – 1981. – 17. – P. 135 - 140.

123. Бодянский, Е.В. Адаптивные алгоритмы идентификации нелинейных объектов управления / Е.В. Бодянский //АСУ и приборы автоматики. – Харьков: Выща шк. – 1987. – Вып. 81. – С. 43 - 46.

124. Bodyanskiy, Ye. An adaptive learning algorithm for neuro-fuzzy network / Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan // Computational Intelligence. Theory and Applications. – Berlin, Heidelberg. – NY: Springer. – 2001. – P. 68 - 75.

125. Frank, A. UCI Machine Learning Repository / A. Frank, A. Asuncion // [http://archive.ics.uci.edu/ml]. – Irvine, CA: University of California, School of Information and Computer Science, 2013.

126. Черкашин, П.А. Готовы ли Вы к войне за клиента? Стратегия управления взаимоотношениями с клиентами / П. А. Черкашин. – М.:ИНТУИТ. – 2004 – 384 с.

127. Бажанов, С.А. Инфракрасная диагностика электрооборудования распределительных устройств / С.А. Бажанов. – М.: НТФ "Энерго-прогресс". – 2000. – 76 с.

128. Вавилов, В.П. Тепловые методы неразрушающего контроля: Справочник / В.П. Вавилов – М.: Машиностроение. – 1991. – 240 с.

129. Опанасенко, В.А. (Самитова В.А.) Алгоритм нечеткой кластеризации данных, представленных порядковыми атрибутами / В.А. Опанасенко, А.Н. Слипченко // Системный анализ и информационные технологии: IX международная научно-техническая конференция, 15-19 апреля 2007г.: тез. докл. – Киев. – 2007. – С. 126.

130. Самитова, В.А. Нечеткая кластеризация порядковых данных с помощью двойного нео-фаззи нейрона / В.А. Самитова // Радиоэлектроника и молодежь в XXI веке: 12-й Международный молодежный форум, 1 – 3 апреля 2008 г.: мат. конф. – Харьков. – 2008. – С. 144.

131. Самитова, В.А. Нечеткая робастная кластеризация порядковых данных на основе мер схожести / В.А. Самитова // Радиоэлектроника и молодежь в XXI веке: 19-й Международный молодежный форум, 20 – 22 апреля 2015 г.: мат. конф. – Харьков. – 2015. – С. 58 - 59.

132. Самитова, В.А. Нечеткая кластеризация порядковых данных на основе функций принадлежности и функций правдоподобия / В.А. Самитова //

Радиоэлектроника и молодежь в XXI веке: XX Юбилейный Международный молодежный форум, 19 – 21 мая 2016 г.: мат. конф. – Харьков. – 2016. – С. 47 - 48.

133. Бодянский, Е.В. Возможностная нечеткая кластеризация категориальных данных на основе частотных прототипов и мер несходства / Е.В. Бодянский, В.А. Самитова // Полиграфические, мультимедийные и web-технологии: I Международная научно-техническая конференция, 16 – 20 мая 2016 г.: мат. конф. – Харьков. – 2016. – С. 39 - 40.

134. Бодянский, Е.В. Рекуррентная нечеткая кластеризация данных на основе отображения порядковых характеристик в цифровую шкалу / Е.В. Бодянский, В.А. Самитова // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта (ISDMCI’2016): XII международная научная конференция, 24 – 28 мая 2016 г.: мат. конф. – Железный порт. – 2016. – С. 258 - 260.

# ПРИЛОЖЕНИЕ А.
# АКТЫ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

Затверджую
Перший проректор
Харківського національного
університету радіоелектроніки
_____ І.І. Ключник
«24» 05 2016 р.

**А К Т**

про впровадження в навчальний процес результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук «Класифікація та кластеризація даних, що задані в нечислових шкалах» аспірантки кафедри штучного інтелекту Харківського національного університету радіоелектроніки Самітової Вікторії Олександрівни

Комісія у складі декана факультету комп'ютерних наук, д.т.н., проф. Єрохіна А.Л., завідувача кафедри штучного інтелекту, д.т.н., проф. Філатова В.О., проф. каф. штучного інтелекту, к.т.н., доц. Рябової Н.В. підтверджує, що результати дисертаційної роботи Самітової В.О., що пов'язані із розробкою адаптивних методів навчання нейро-фаззі систем для інтелектуального аналізу даних, впроваджені в навчальний процес на кафедрі штучного інтелекту в курсах «Штучні нейронні мережі: архітектури, навчання та застосування» та «Нейромережеві методі обчислювального інтелекту».

Декан факультету КН, д.т.н., проф. _____ А.Л. Єрохін

Зав. каф. ШІ, д.т.н., проф. _____ В.О. Філатов

Проф. каф. ШІ, к.т.н., доц. _____ Н.В. Рябова

ISO 9001:2008 MIDIEL ДСТУ ISO 9001-2009

| науково-виробнича фірма «МІДІЕЛ» | | научно-производственная фирма «МИДИЭЛ» | |
|---|---|---|---|
| | Україна, 61037, м. Харьків, пр. Московський, буд. 199, корпус цеха ШЦ 28, кімната 10 | | Украина, 61037, г. Харьков, пр. Московский, д. 199, корпус цеха ШЦ 28, комната 10 |
| тел.: | +380 (62) 385-34-54, +380 (50) 422-25-22 | тел.: | +380 (62) 385-34-54, +380 (50) 422-25-22 |
| факс: | +380 (62) 385-33-59 | факс: | +380 (62) 385-33-59 |

**ЗАТВЕРДЖУЮ**
Директор ТОВ НВФ «МІДІЕЛ»
Сульженко В.О.
«20» 05 2016

№20348024

# АКТ
про впровадження результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук
Самітової Вікторії Олександрівни

Комісія у складі:

Голова Григор'єв С.В.
Члени комісії Тимошенко І.М.
Кириченко М.О.

склала цей акт про те, що метод кластерування, розроблений в дисертаційній роботі Самітової В.О., використано для проведення автоматичної діагностики електрообладнання на підприємстві. Це дозволило виявляти дефекти електрообладнання на ранніх стадіях їх розвитку, підвищити швидкість, точність та зручність діагностики, порівняно з методами, що застосовувалися раніше.

Результати впровадження довели, що розроблений Самітовою В.О. метод, який ґрунтуються на сучасних інтелектуальних технологіях, виконаний на високому науково-технічному рівні та має переваги над існуючими рішеннями.

Комісія підтверджує доцільність використання розробленого методу для розв'язання задачі автоматичної діагностики електрообладнання, що виготовляється на підприємстві.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Голова комісії Григор'єв С.В.

Члени комісії Тимошенко І.М.

Кириченко М.О.

**Товариство з обмеженою відповідальністю «Південелектропроект»**
Юридична адреса: 61037, м. Харків, пр. Московський 199
Поштова адреса: а/с 11868, м. Харків, 61037
Тел./факс: (057) 728-13-32
ЄДРПОУ 35700489
р/р 26000000077999
в ПАТ «УКРСОЦБАНК» МФО 300023
інд. под. № 357004820344
E-mail: info@ooo-yuep.com

**Общество с ограниченной ответственностью «Южэлектропроект»**
Юридический адрес: 61037, г. Харьков, пр. Московский, 199
Почтовый адрес: а/я 11868, г. Харьков, 61037
Тел./факс: (057) 728-13-32
ЕГПРОУ 35700489
р/с 26000000077999
в ПАО «УКРСОЦБАНК» МФО 300023
инд. нал. № 357004820344
E-mail: info@ooo-yuep.com

**ЗАТВЕРЖУЮ**
Директор
ТОВ «Південелектропроект»
_____ Б.О. Кіяшко
«25» 05.2016

АКТ
про впровадження результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук
Самітової Вікторії Олександрівни

Комісія у складі:
Голова — Моргунов О.В. – комерційний директор
Члени комісії — Перельмутер Д.В. – технічний директор
Єфімченко М.В. – керівник відділу програмувальників

склала цей акт про те, що розроблений в дисертаційній роботі Самітової В.О. метод кластерування, застосовано для аналізу клієнтської бази підприємства. Використання розробленого методу дозволило збільшити прибуток підприємства за рахунок збільшення об'ємів продажу і оптимізації операційних витрат, а також зменшити час, необхідний для аналізу даних, порівняно з методами, що застосовувалися раніше.

Комісія підтверджує працездатність програмного засобу, розробленого на засаді запропонованого методу у вигляді програмного модулю, для розв'язання задачі аналізу клієнтської бази підприємства.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Голова комісії _____ О.В. Моргунов

Члени комісії _____ Д.В.Перельмутер

_____ М.В. Єфімченко