Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Институт кибернетики
Направление подготовки 09.03.01 «Информатика и вычислительная техника»
Кафедра вычислительной техники

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Кластеризация числовых данных рекуррентной нейронной сетью

УДК 004.67: 004.93’14.032.26

Студент

Группа	ФИО	Подпись	Дата
8В3Б	Имигеев Евгений Иннокентьевич		

Руководитель

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент кафедры ИСТ	Немировский В.Б.	к.ф.-м.н		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. МЕН	Спицын В.В.	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент каф. ЭБЖ	Невский Е.С.	-		

ДОПУСТИТЬ К ЗАЩИТЕ:

Зав. кафедрой	ФИО	Ученая степень, звание	Подпись	Дата
ИСТ	Мальчуков А.Н.	к.т.н.		

Томск – 2017

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Институт кибернетики
Направление подготовки 09.03.01 «Информатика и вычислительная техника»
Кафедра вычислительной техники

УТВЕРЖДАЮ:
Зав. кафедрой

_____ _____ Мальчуков А.Н.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

Студенту:

Группа	**ФИО**
8В3Б	Имигееву Евгению Иннокентьевичу

Тема работы:

Кластеризация числовых данных рекуррентной нейронной сетью	
Утверждена приказом директора (дата, номер)	07.02.2017 г., № 789/с

Срок сдачи студентом выполненной работы:	16.06.2017

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Разработка метода кластеризации числовых данных рекуррентной нейронной сетью.
Перечень подлежащих исследованию, проектированию и разработке вопросов	Выбор языка реализации приложения, параметров искусственной нейронной сети, изучение предложенного метода, проектирование приложения. Исследование полученных результатов и оценка эффективности применённого алгоритма. Расчет ресурсоэффективности и ресурсосбережения; Анализ вредных производственных факторов.
Перечень графического материала	Структурная схема нейронной сети; Структурные схемы нейрона; Функции активации нейрона; Одномерное точечное отображение; Реализованное приложение;

	Ирисы Фишера

Консультанты по разделам выпускной квалификационной работы	
Раздел	**Консультант**
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Спицын Владислав Владимирович
Социальная ответственность	Невский Егор Сергеевич

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	07.02.2017

Задание выдал руководитель:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент кафедры ИСТ	Немировский В.Б.	к.ф.-м.н		07.02.2017

Задание принял к исполнению студент:

Группа	**ФИО**	**Подпись**	**Дата**
8В3Б	Имигеев Евгений Иннокентьевич		07.02.2017

Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Институт кибернетики
Направление подготовки 09.03.01 «Информатика и вычислительная техника»
Уровень образования Бакалавриат
Кафедра Информационных систем и технологий
Период выполнения осенний / весенний семестр 2016/2017 учебного года
Форма представления работы:

Бакалаврская работа

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	16.06.2017

Дата контроля	**Название раздела (модуля) / вид работы (исследования)**	**Максимальный балл раздела (модуля)**
24.02.2017	Анализ предметной области	10
17.03.2017	Исследование предложенного алгоритма для реализации	10
24.04.2017	Реализация метода	20
19.05.2017	Отладка программы и проведение экспериментов на тестовых данных	20
30.05.2017	Оформление пояснительной записки	20
14.06.2017	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
14.06.2017	Социальная ответственность	10

Составил преподаватель:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент каф. ИСТ	Немировский Виктор Борисович	к.ф.-м.н.		07.02.2017

СОГЛАСОВАНО:

Зав. кафедрой	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Информационные системы и технологии	Мальчуков Андрей Николаевич	к.т.н.		07.02.2017

ЗАПЛАНИРОВАННЫЕ РЕЗУЛЬТАТЫ ПО ОСНОВНОЙ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЕ ПОДГОТОВКИ БАКАЛАВРОВ 09.03.01 «ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА», ИК ТПУ, ПРОФИЛЬ «ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ, КОМПЛЕКСЫ, СИСТЕМЫ И СЕТИ»

Код результата	Результат обучения (выпускник должен быть готов)	Требования ФГОС, критерии АИОР
Р1	Применять базовые и специальные естественнонаучные и математические знания в области информатики и вычислительной техники, достаточные для комплексной инженерной деятельности.	Требования ФГОС (ОК-1, 10, ПК-4, 5, 6), критерий 5 АИОР (п. 1.1)
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.	Требования ФГОС (ОК-11, 12, 13, ПК-1, 2, 11), критерий 5 АИОР (п.1.1, 1.2)
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием аппаратно-программных средств информационных и автоматизированных систем, с использованием базовых и специальных знаний, современных аналитических методов и моделей.	Требования ФГОС (ОК-1, 8, ПК-2, 4, 6), критерий 5 АИОР (п. 1.2)
Р4	Разрабатывать программные и аппаратные средства (системы, устройства, блоки, программы, базы данных и т. п.) в соответствии с техническим заданием и с использованием средств автоматизации проектирования.	Требования ФГОС (ОК-2, 3, ПК-3, 4, 5), критерий 5 АИОР (п. 1.3)
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания аппаратных и программных средств информационных и автоматизированных систем.	Требования ФГОС (ОК-6, ПК-6, 7), критерий 5 АИОР (п.1.4)
Р6	Внедрять, эксплуатировать и обслуживать современные программно-аппаратные комплексы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.	Требования ФГОС (ОК-4, 15, 16, ПК-9, 10, 11), критерий 5 АИОР (п. 1.5)

Универсальные компетенции		
Р7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности.	Требования ФГОС (ОК-1, 4, ПК-1, 6, 7), критерий 5 АИОР (п. 2.1)
Р8	Владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности.	Требования ФГОС (ОК-14, ПК-7), критерий 5 АИОР (п. 2.2)
Р9	Эффективно работать индивидуально и в качестве члена группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.	Требования ФГОС (ОК-2, 3, 4), критерий 5 АИОР (п. 2.3, 2.4)
Р10	Демонстрировать знания правовых, социальных, экономических и культурных аспектов комплексной инженерной деятельности.	Требования ФГОС (ОК-1, 5, 9), критерий 5 АИОР (п. 2.5)
Р11	Демонстрировать способность к самостоятельной к самостоятельному обучению в течение всей жизни и непрерывному самосовершенствованию в инженерной профессии.	Требования ФГОС (ОК-6, 7), критерий 5 АИОР (п. 2.6)

РЕФЕРАТ

Выпускная квалификационная работа содержит 63 с., 16 рис., 11 табл., 22 источников.

Ключевые слова: рекуррентная нейронная сеть, функция активации нейрона, сигмоида, кластеризация, классификация, ирисы Фишера, евклидовы расстояния, информационная энтропия.

Объектом исследования является задача кластеризации многомерных числовых данных.

Цель работы – исследование и программная реализация метода кластеризации числовых данных рекуррентной нейронной сетью.

В процессе исследования были изучены и проанализированы существующие методики кластеризации.

В результате исследования был реализован алгоритм кластеризации анализируемых данных.

Область применения: анализ числовых данных объектов, процессов или явлений, для поиска и выделения в них существующих закономерностей, для понимания данных путём выявления кластерной структуры.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ И НОРМАТИВНЫЕ ССЫЛКИ

Кластеризация (кластерный анализ) – это прогресс разбиения объектов по некоторому признаку на однородные группы (кластеры). «Кластер» в английском языке означает «сгусток», «гроздь (винограда)», «скопление (звезд)», что говорит о том, что объекты кластеризации представляются точками в пространстве, где и происходит определение всех «скоплений» этих точек.

Искусственная нейронная сеть (в дальнейшем – НС) – математически упрощенная модель головного мозга, программа, предназначенная для выполнения определенной задачи так, как бы выполнил её мозг.

Нейрон – это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передает её дальше.

Функция активации нейрона – функция, принимающая сумму входных сигналов нейрона как аргумент, которая выполняет линейное или нелинейное преобразование.

Точечное отображение – самостоятельный раздел теории динамических систем, где изучаются объекты не с непрерывным, а с дискретным временем.

Ирисы Фишера – это известный набор данных о цветках ирисов, состоящий из 3 классов по 50 объектов каждый. Каждый ирис имеет 4 параметра: длина и ширина лепестка, длина и ширина чашелистика.

Евклидовы расстояния – это расстояния между точками многомерного пространства.

Информационная энтропия – это мера упорядоченности набора данных.

Оглавление

ВВЕДЕНИЕ

В биоинформатике исследуются сложные сети взаимодействующих генов [1], состоящих из сотен или даже тысяч элементов. «Беглым взглядом» невозможно выделить подсети, узкие места, концентраторы и другие скрытые свойства изучаемой системы, а анализ «вручную» займет достаточно много ресурсов и времени.

Чтобы облегчить понимание закономерностей функционирования слабо изученных сложных процессов и явлений следует разработать и использовать специальные методы статистического анализа многомерных данных [2].

Данная работа посвящена одному из наиболее обещающих подходов к анализу многомерных процессов и явлений – кластерному анализу.

Кластеризация является одним из методов интеллектуального анализа данных, входит в перечень задач Data Mining.

Термин Data Mining состоит из двух понятий: добыча ценной информации в большой базе данных (data) как добыча горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Data Mining переводится как добыча данных, интеллектуальный или глубинный анализ данных. «Поиск знаний в большой базе данных» - это можно назвать синонимом Data Mining [3].

В основе кластеризации находится классификация, которая так же является одной из задач Data Mining.

Классификация – это системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам или критериям для удобства их исследования; группировка множества объектов и расположение их в определенном порядке, отражающем их сходные классификационные признаки, выбранных для определения сходства или различия между этими объектами.

Процесс классификации заключается в разбиении множества объектов на существующие, уже определенные классы по некоторому критерию.

Кластеризация является более сложной задачей, её особенность заключается в том, что классы объектов изначально не предопределены. В результате кластеризации объекты разбиваются на группы, разделёнными по определенному признаку.

Кластеризация – это процесс разбиения объектов по некоторому признаку на однородные группы (кластеры), основанный на представлении объектов точками геометрического пространства с последующим выделением групп объектов как «сгустков» этих точек. Собственно, «кластер» (cluster) в английском языке и означает «сгусток», «гроздь (винограда)», «скопление (звезд)» и т.п.

Характеристиками кластера являются два признака: внутренняя однородность – объекты одного кластера однородны и схожи; внешняя изолированность – кластеры заметно отделимы друг от друга, то есть объекты одного кластера существенно отличаются от объектов других кластеров.

Цели кластеризации [4]:

- Понимание данных путём выявления кластерной структуры – разбиение выборки на группы позволяет упростить дальнейшую обработку данных;
- Сжатие данных – если исходная выборка достаточно избыточна, то можно сократить её до числа кластеров, взяв из каждого по одному типичному представителю;
- Обнаружение новизны – в процессе кластеризации выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1. Определение выборки объектов для анализа;
2. Вычисление меры сходства между объектами;
3. Применение метода кластеризации для создания групп (кластеров) сходных объектов;
4. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации для получения оптимального результата [5].

Метод кластеризации, исследуемый в этой работе, основан на кластеризующих свойствах рекуррентного нейрона. С помощью одномерного точечного отображения входных сигналов нейрона на его функции активации возможно поделить выборку на кластеры [6].

Для оценки качества работы алгоритма будет произведена кластеризация уже известного набора данных, а именно – ирисов Фишера, которые имеют 4 параметра (4-мерное пространство) и содержат 3 класса по 50 цветков каждый. По результатам работы программы с данным набором данных будет ясно, насколько точно работает данный метод.

1 АНАЛИТИЧЕСКИЙ ОБЗОР

В данный момент разработано достаточно большое количество методов кластеризации. У каждого есть свои плюсы и минусы, а для разработки нового метода следует провести обзор этих уже существующих алгоритмов.

1.1 Методы кластеризации данных

В общем выделяются две основные классификации алгоритмов кластеризации [5]:

- Иерархические и плоские:
  - Иерархические алгоритмы (или алгоритмы таксономии) строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений, т.е. на выходе получается дерево кластеров, корнем которого является вся выборка, а листьями – наиболее мелкие кластеры;
  - Плоские алгоритмы строят одно разбиение объектов на кластеры;
- Чёткие и нечёткие:
  - Чёткие (или непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит только одному кластеру;
  - Нечёткие (пересекающиеся) алгоритмы каждому объекту ставят в соответствие набор вещественных значений, показывающих степень отношения объекта к кластерам – каждый объект относится к каждому кластеру с некоторой вероятностью.

Рассмотрим кратко некоторые методы.

1.1.1 Алгоритмы иерархической кластеризации

Алгоритмы иерархической кластеризации делятся на 2 типа: восходящие и нисходящие алгоритмы. Нисходящие алгоритмы работают по принципы «сверху-вниз»: вначале все объекты помещаются в один кластер, который затем разбивается на всё более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева – дендрограммы. Классический пример такого дерева – классификация животных и растений.

К недостатку иерархических алгоритмов можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

1.1.2 Алгоритмы квадратичной ошибки

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения.

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Самым распространенным алгоритмом этой категории является метод k-средних. Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

1. Случайно выбрать k точек, являющихся начальными «центрами масс» кластеров;
2. Отнести каждый объект к кластеру с ближайшим «центром масс» (точка со средними значениями характеристик для определенного кластера);
3. Пересчитать «центры масс» кластеров согласно их текущему составу;
4. Если критерий остановки алгоритма не удовлетворен, вернуться к п.2.

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Так же возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер.

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

1.1.3 Нечеткие алгоритмы

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм с-средних (c-means). Он представляет собой модификацию метода k-средних. Шаги работы алгоритма:

1. Выбрать начальное нечеткое разбиение n объектов на k кластеров путем выбора матрицы принадлежности U размера n x k.
2. Используя матрицу U, найти значение критерия нечеткой ошибки:
3. Перегруппировать объекты с целью уменьшения этого значения критерия нечеткой ошибки.
4. Возвращаться в п. 2 до тех пор, пока изменения матрицы U не станут незначительными

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

1.1.4 Алгоритмы, основанные на теории графов

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа $G = (V, E)$, вершинам которого соответствуют объекты, а ребра имеют вес, равный «расстоянию» между объектами. Достоинством графовых алгоритмов кластеризации являются наглядность, относительная простота реализации и возможность вносения различных усовершенствований, основанные на геометрических соображениях. Основными алгоритмам являются алгоритм выделения связных компонент, алгоритм

построения минимального покрывающего (остовного) дерева и алгоритм послойной кластеризации.

Из минусов этих алгоритмов можно отметить следующие: сложности управления количеством кластеров; достаточно плохое обнаружение кластеров с большим диаметром; часто кластеры получаются разного размера; огромные вычислительные сложности с многомиллионными выборками из-за необходимости вычисления между всеми объектами их расстояний [7].

1.2 Сравнение алгоритмов

Из приведенных алгоритмов видно, что все они либо имеют какие-то важные параметры, которые необходимо задать перед началом анализа, либо требуют тщательного подбора этих параметров, либо требуют достаточно сложную вычислительную подготовку данных к анализу.

В таблице 1 показана вычислительная сложность алгоритмов кластеризации.

Табл.1.1 – Вычислительная сложность алгоритмов кластеризации

Алгоритм кластеризации	**Вычислительная сложность**
Иерархический	$O(n^2)$
k-средних	$O(nkl)$, где k – число кластеров, l – число итераций
c-средних	
Выделение связных компонент	*Зависит от алгоритма*
Минимальное покрывающее дерево	$O(n^2 \log n)$
Послойная кластеризация	$O(max(n, m))$, где $m < n(n-1)/2$

В данной работе предлагается создать такой алгоритм, который будет иметь преимущественно автоматический характер работы – ввод как можно меньшего количества параметров для кластеризации с небольшой вычислительной сложностью, и, конечно, высокой точностью работы.

2 МЕТОД КЛАСТЕРИЗАЦИИ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТЬЮ

Метод кластеризации, исследуемый в этой работе, основан на кластеризующих свойствах рекуррентного нейрона [6]. С помощью одномерного точечного отображения входных сигналов нейрона на его функции активации возможно поделить выборку на кластеры в соответствии с гипотезой компактности о принадлежности близких и схожих объектов к одному классу.

2.1 Нейронная сеть

Мозг представляет собой сложный, нелинейный, параллельный компьютер (систему обработки данных). Он способен самоструктурироваться, чтобы его сети из нейронов могли выполнять конкретные задачи (например, распознавание образов, обработка сигналов органов чувств, моторные функции) во много раз быстрее, чем могут позволить самые быстродействующие современные компьютеры. На распознавание знакомого лица у человеческого мозга уходит около 100-200 миллисекунд, в то время как выполнение аналогичных задач даже меньшей сложности на компьютере может занять несколько дней [8].

Искусственная нейронная сеть (ИНС) – математическая упрощенная модель биологической нейронной сети - мозга. ИНС представляет собой программу, предназначенную для выполнения определенной задачи так, как бы выполнил ее мозг. Чтобы добиться этого, нейронную сеть необходимо обучать, чтобы она накапливала «опыт» работы с интересующей нас задачей. Чем сложнее задача, тем дольше обучение и больше нейросеть [17].

2.1.1 Структура нейронной сети

ИНС состоит из 3-х слой: входной, скрытый, выходной (рис.2.1).

Входной слой принимает данные для дальнейшей обработки в скрытом слое. После чего на выходном слое формируется вывод сети.

Скрытые слои, выполняющие всю работу, представляют собой «черный ящик», который в процессе обучения настраивается на достижение результата. Неизвестно как работает скрытый слой. Данное свойство нейронных сетей называется непрозрачностью. В зависимости от сложности задачи скрытых слоев может быть больше чем один.

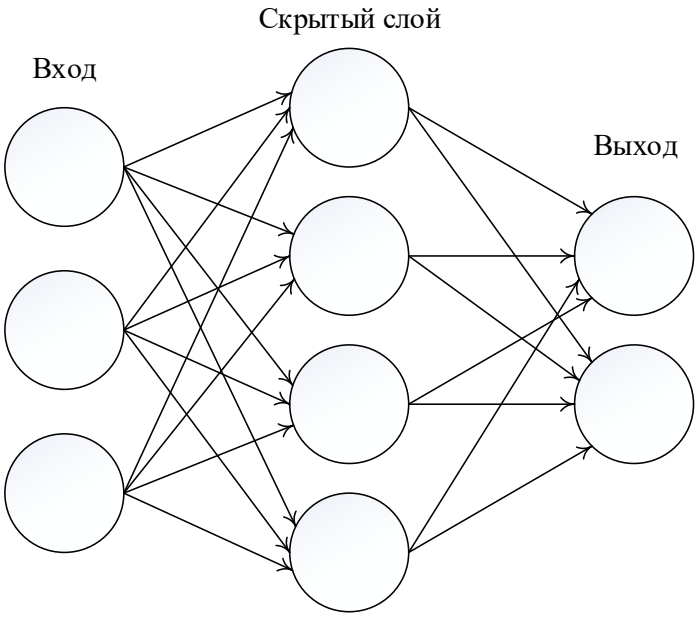

Рис.2.1. Структура нейронной сети прямого распространения

2.1.2 Структура нейрона

Структура искусственного нейрона: у нейрона есть входы (у каждого входа свой вес, который говорит о важности этого входа, называемый синапсом), сумматор, функция активации (может быть не у всех нейронов) и выход (рис.2.2).

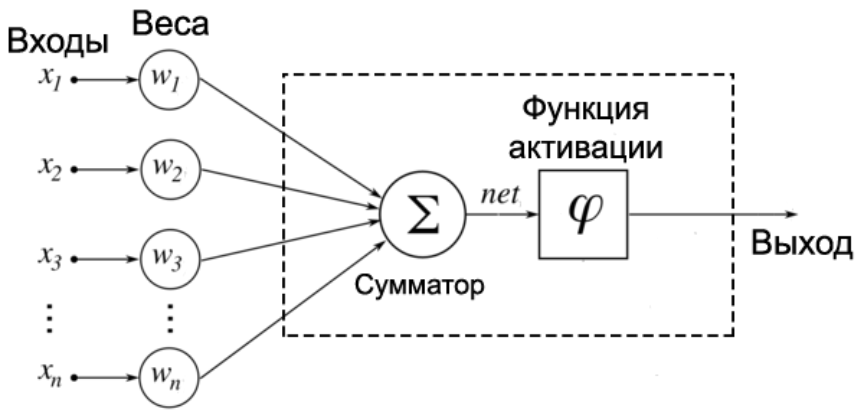

Рис.2.2. Внутренняя структура нейрона

Входы умножаются на свои веса, складываются в сумматоре и передаются в функцию активации нейрона, которая формирует выход нейрона.

2.1.3 Функция активации нейрона

Функция активации (активационная функция, функция возбуждения) – функция, вычисляющая выходной сигнал искусственного нейрона. В качестве аргумента принимает сигнал, получаемый на выходе входного сумматора. Наиболее часто используются следующие функции активации:

- Единичный скачок или жесткая пороговая функция (рис.2.3) – самая простая функция, не обладает гибкостью;

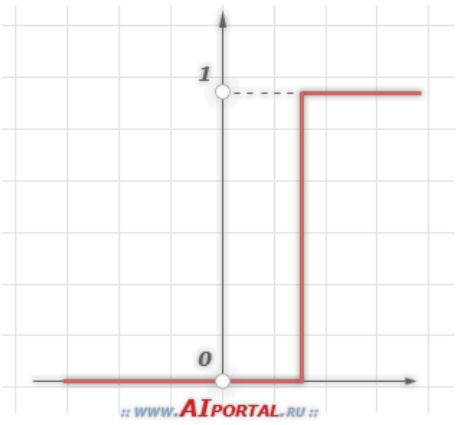

Рис.2.3. Жесткая пороговая функция

- Линейный порог или гистерезис (рис.2.4) – кусочно-линейная функция; не имеет недостатка пороговой функции; невысокая вычислительная сложность;

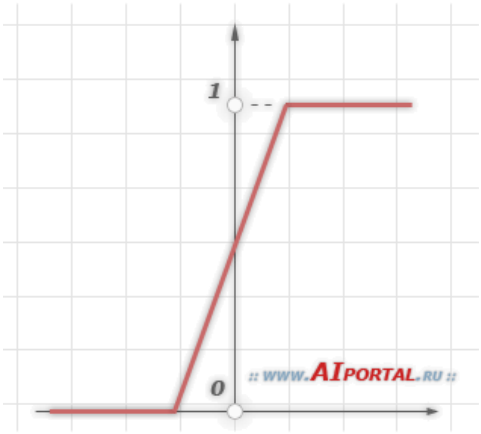

Рис.2.4. Гистерезис

- Сигмоидальная функция или сигмоида (рис.2.5) – непрерывная, монотонно возрастающая, дифференцируемая функция активации,

$$f(x) = \frac{1}{1 + e^{-x}}. \qquad (1)$$

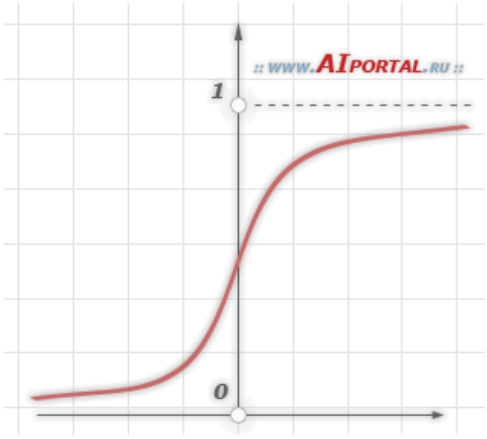

Рис.2.5. Сигмоид – логистическая функция

Слабые входные сигналы нуждаются в большом усилении, чтобы дать пригодный к использованию выходной сигнал нейрона, а сильные входные сигналы необходимо наоборот ослабить. Сигмоида обладает свойством

усиливать слабые сигналы лучше, чем сильные, и может насыщаться от сильных сигналов. Таким образом одна и та же сеть может обрабатывать как слабые сигналы, так и сильные.

Еще одним примером сигмоидальной функции активации является гиперболический тангенс (рис.2.6), задаваемая следующим выражением:

$$f(x) = th\left(\frac{x}{\alpha}\right)$$

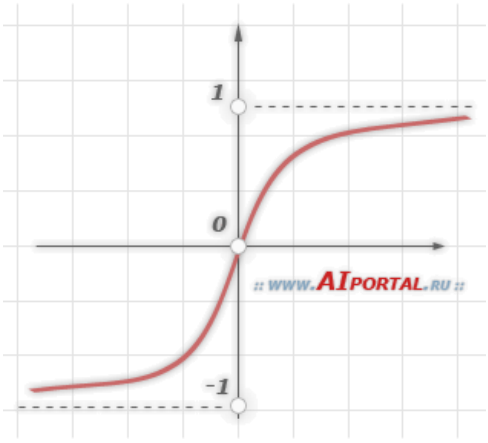

Рис.2.6. Сигмоид – гиперболический тангенс

В отличии от логистической функции гиперболический тангенс позволяет выдавать выходной сигнал различных знаков, что для ряда сетей оказывается выгодным. Эту функцию используют, когда выводом может быть три, например, когда сеть должна определить движение – идти вперед, назад или стоять на месте.

Функции активации, такие как жесткая пороговая и линейная, встречаются очень редко и, как правило, используются на учебных примерах. В практических задачах, например, прогнозировании, классификации и др., всегда применяется сигмоидальная функция активации.

2.2 Рекуррентная нейронная сеть. Точечное отображение

Рекуррентные нейронные сети – вид нейронных сетей, в которых имеется обратная связь. Обратные связи могут быть как между нейронами разных слоев, так и одного слоя [18].

2.2.1 Рекуррентный нейрон

Если в цепи нейронов сети каждый нейрон выполняет одно и то же действие, то такую цепь можно преобразовать в один нейрон с локальной обратной связью (рис.2.7). Так поступающий сигнал будет циркулировать по нейрону, многократно проходя через одну и ту же функцию активации [6].

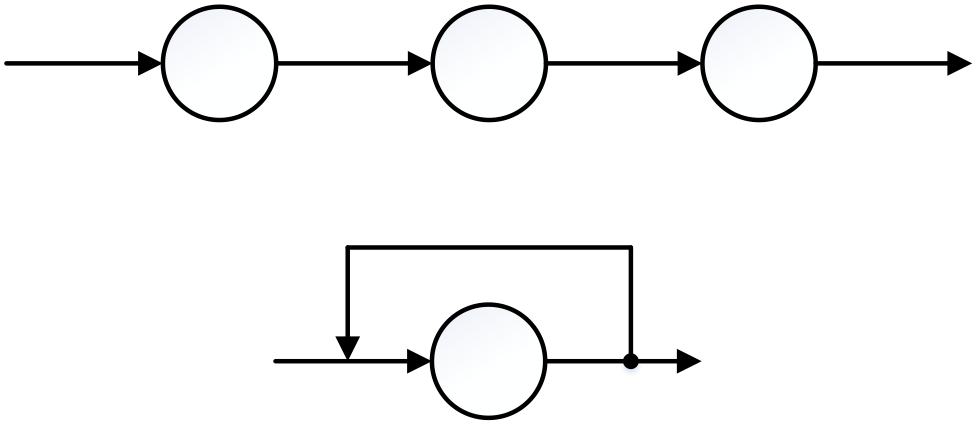

Рис.2.7. Структура рекуррентного нейрона

2.2.2 Одномерное точечное отображение

Модель сети данной работы основана на точечных отображениях.

С математической точки зрения нейрон с локальной обратной связью реализует точечное отображение входного значения нейрона на его активационной функции.

Точечные отображения – это самостоятельный раздел теории динамических систем, где изучаются объекты не с непрерывным, а с дискретным временем [9].

Следующее соотношение определяет точечное отображение

$$x_{n+1} = f(x_n),$$

где n – номер текущей итерации. Поочередно применяя отображение получаем бесконечную последовательность точек

$$x_0, x_1, \dots, x_n, x_{n+1}, \dots$$

Точечное отображение входов и выходов рекуррентного нейрона можно показать на диаграмме Ламерея (рис.2.8), где x_1^*, x_3^* - неподвижная устойчивые точки, x_2^* - неподвижная неустойчивая точка.

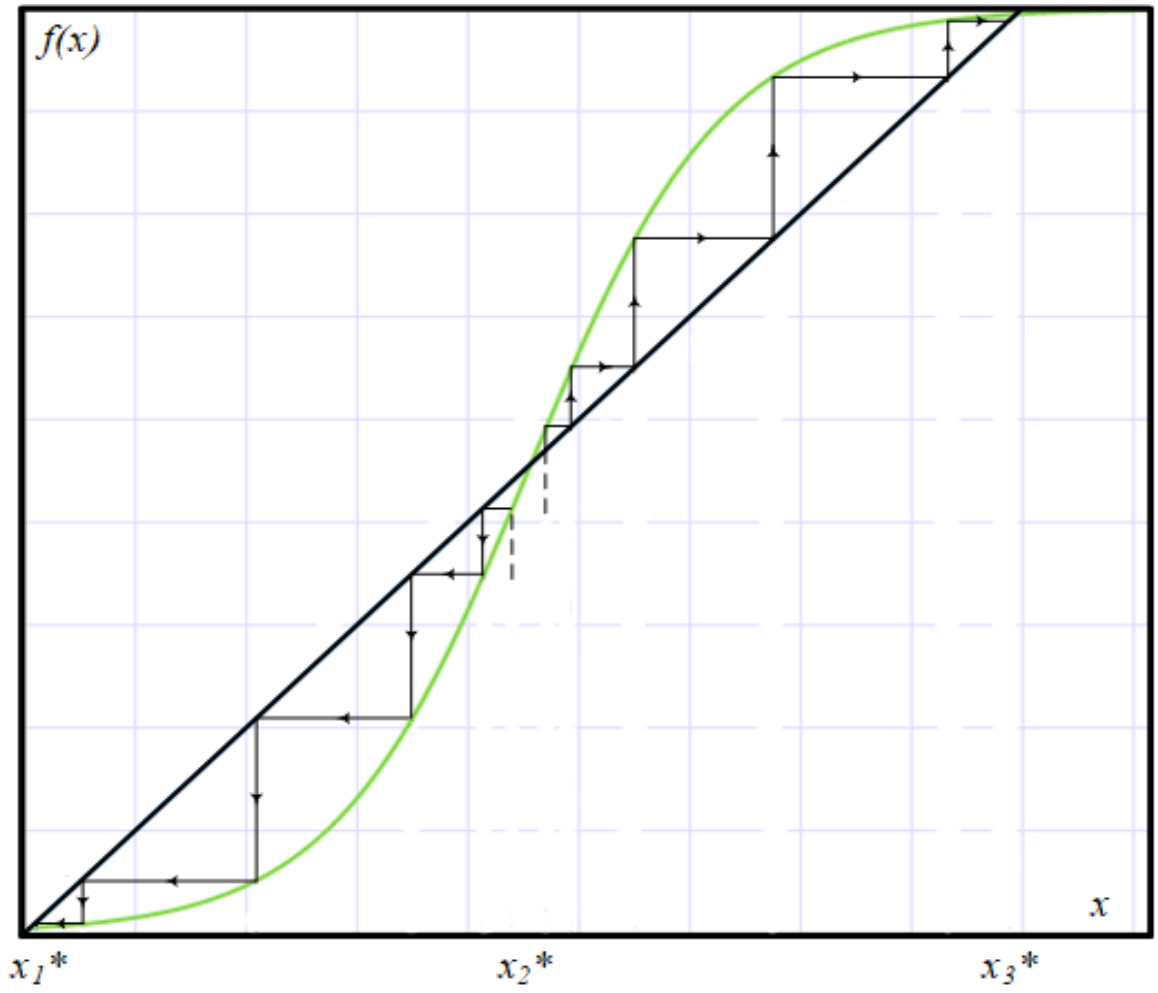

Рис.2.8. Диаграмма Ламерея для некоторого одномерного отображения f

Такое отображение называется сжимающим.

Если значение функции равняется самой функции, то такая точка является неподвижной (значение *x* в месте пересечения функции и биссектрисы):

$$x = f(x).$$

Неподвижные точки x^* также делятся на устойчивые и неустойчивые.

Неподвижная точка x^* точечного отображения устойчива, если

$$|f'(x^*)| < 1,$$

и неустойчива, если

$$|f'(x^*)| > 1.$$

Если $f'(x) = 1$, то вопрос об устойчивости неподвижной точки определяется высшими производными [10].

Производную сигмоиды можно рассчитать следующим образом:

$$f'(x) = \frac{\alpha * \left(f(x)\right)^2 * e^{-\alpha * x + \beta}}{\mu}$$

2.3 Кластеризующие свойства рекуррентного нейрона

В процессе отображения $x_{n+1} = f(x_n)$ на активационную функцию нейрона $f(x)$ любое значение *x* из интервала между точками *A* и *B*, достигнет устойчивой точки за одно и то же количество итераций (см.рис.2.9). Это, в соответствии с гипотезой компактности, позволяет отнести подмножество значений *x* из интервала *(A, B)* к одному кластеру.

Гипотеза компактности предполагает, что схожие объекты гораздо чаще лежат в одном классе, чем в разных; или, другими словами, классы образуют компактно локализованные подмножества в пространстве объектов [11].

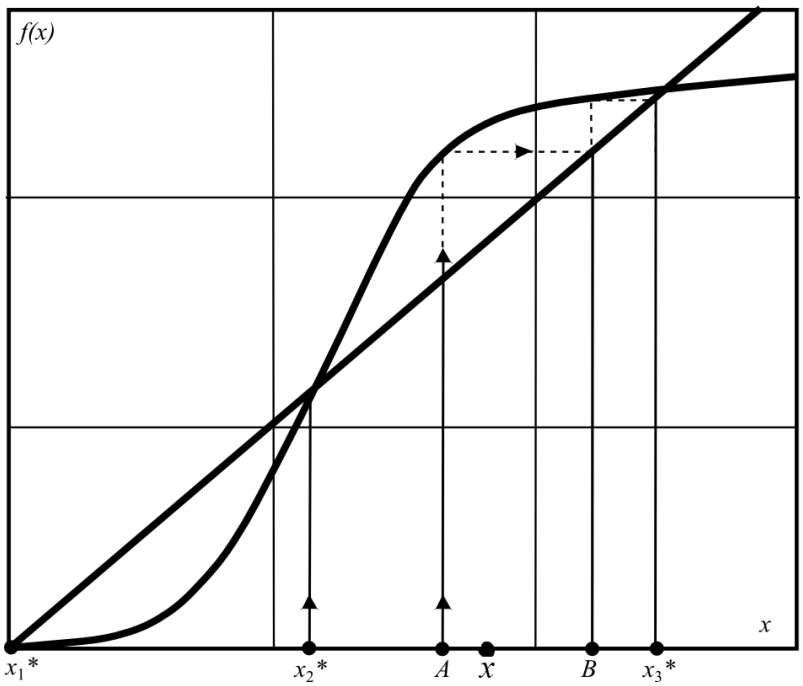

Рис.2.9. Отображение входного сигнала на функции активации нейрона

С помощью точечного отображения входного сигнала нейрона на его функции активации мы получаем кластеры. Следовательно, чем лучше настроена функция активации, тем точнее будет кластеризация.

Однако для проведения точечного отображения входов нейрона на функции активации должно быть 3 неподвижные точки, между которыми и находятся кластеры.

Для этого необходимо настроить функцию активации так, чтобы она пересекалась с биссектрисой 3 раза. Так же данные кластеризации необходимо нормировать к единичному интервалу для удобства работы с ними.

2.3.1 Настройка параметров активационной функции

Чтобы настроить сигмоиду под входные данные следует использовать не классическую сигмоиду (1), а сигмоиду с двумя параметрами: коэффициентом наклона функции и коэффициентом смещения функции [12].

$$f(x) = \frac{1}{1 + e^{-\alpha x + \beta}}, \quad (2)$$

где α - коэффициент наклона, β - коэффициент смещения.

Зная минимальное и максимальное значения интервала входных данных кластеризации, меняя параметры α и β сигмоиды (2), ее можно настроить так, чтобы нелинейная часть функции покрывала весь интервал входных данных. Так это даст кластеризацию всех входных данных.

Однако высока вероятность, что такая функция активации приведет к некачественной кластеризации. Поэтому следует использовать сигмоиду с тремя параметрами:

$$f(x) = \frac{\mu}{1 + e^{-\alpha \cdot x + \beta}}, \quad (3)$$

где μ - коэффициент усиления.

Выражение (3) имеет 3 неизвестных. Так как нам известны границы интервала входных данных, мы можем вывести выражения для двух коэффициентов α и β, а третью неизвестную μ примем как изменяемый параметр.

Найдем выражения коэффициентов α и β для неподвижной точки *x:*

$$x = f(x):$$

$$x = \frac{\mu}{1 + e^{-\alpha \cdot x + \beta}}$$

$$1 + e^{-\alpha \cdot x + \beta} = \frac{\mu}{x}$$

$$e^{-\alpha \cdot x + \beta} = \frac{\mu}{x} - 1$$

$$\ln\left(e^{-\alpha \cdot x + \beta}\right) = \ln\left(\frac{\mu}{x} - 1\right)$$

$$-\alpha \cdot x + \beta = \ln\left(\frac{\mu}{x} - 1\right) \quad (4)$$

Из выражения (4) напишем систему уравнений для x_{max} и x_{min}:

$$\begin{cases} -\alpha \cdot x_{max} + \beta = \ln\left(\frac{\mu}{x_{max}} - 1\right) \\ -\alpha \cdot x_{min} + \beta = \ln\left(\frac{\mu}{x_{min}} - 1\right) \end{cases}$$

Вычтем уравнения друг из друга:

$$-\alpha \cdot (x_{max} - x_{min}) = \ln\left(\frac{\mu}{x_{max}} - 1\right) - \ln\left(\frac{\mu}{x_{min}} - 1\right)$$

$$\alpha = -\ln\left(\frac{\frac{\mu}{x_{max}} - 1}{\frac{\mu}{x_{min}} - 1}\right) \cdot \frac{1}{x_{max} - x_{min}} \quad (5)$$

Теперь из (4) выведем выражение для β для x_{max} и x_{min}:

$$-\alpha \cdot x + \beta = \ln\left(\frac{\mu}{x} - 1\right)$$

$$\beta = \ln\left(\frac{\mu}{x} - 1\right) + \alpha \cdot x \quad (6)$$

В формулах (5) и (6) коэффициенты активационной функции α и β зависят от неизвестного параметра μ, варьируя который, можно добиваться оптимальной кластеризации (см. раздел 2.4). При этом рассчитанные α и β функции активации нейрона будут автоматически подстраиваться к новому входному вектору.

Так же заметим, что значения x_{min} и x_{max} должны отличатся от нуля и быть больше значения параметра μ, так как это приведет к неопределенности (деление на нуль, отрицательный логарифм или логарифм нуля). Поэтому не следует выполнять нормировку данных к единичному интервалу, вместо этого необходимо провести коэффициентную нормировку входных значений.

2.3.2 Максимальное значение коэффициента усиления

При изменении параметра μ меняются коэффициенты сигмоиды α и β, следовательно, меняется и сама кривая. При увеличении μ кривая будет вырождаться – неустойчивая и устойчивая точки будут совпадать. Поэтому у параметра μ следует определять максимальное значение (параметр μ не может превышать единицу), такое, после которого кривая может вырождаться.

Чтобы определить максимально допустимое значение параметра μ, необходимо найти такое μ, при котором условие устойчивости (глава 2.2.2) неподвижной точки нарушается.

Как было выявлено (глава 2.3.1), значение μ должно быть больше x_{max}, иначе расчет коэффициентов α и β будет невозможен.

Из всего выше перечисленного интервал допустимых значений параметра μ будет выглядеть следующим образом:

$$\mu \in (x_{max}, \mu_{max}]$$

где μ_{max} – такое μ, после которого кривая может вырождаться.

2.3.3 Нормировка

Чтобы нормировать данные интервала можно использовать следующую формулу:

$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Однако такая нормировка приведет значения x в диапазон от 0 до 1, чего необходимо избежать. Поэтому следует добавить некий коэффициент сжатия K_1 и коэффициент сдвига от нуля K_2:

$$\bar{x} = K_1 \frac{x - x_{min}}{x_{max} - x_{min}} + K_2 \qquad (7)$$

Тогда интервал входных значений x из [0, 1] станет [K_2, K_1+K_2].

2.4 Условие оптимальности кластеризации

Поскольку два параметра сигмоиды из трёх определены, то третий параметр остается переменной величиной. Изменяя параметр μ, кластеризацию одного набора данных можно произвести множество раз. При другом значении параметра сигмоиды будет меняться как кривая, так и результаты кластеризации. Тем самым появляется необходимость выбрать самую оптимальную кластеризацию из их множества.

Экспериментально было установлено [13], что оптимальность кластеризации наступает при таком значении μ, которое находится в окрестности максимальной скорости изменения энтропии

$$\mu_{\text{опт}} = \max \left| \frac{\Delta H}{\Delta \mu} \right|$$

где ΔH – скорость изменения энтропии, соответствующее изменению μ.

Информационная энтропия – это мера упорядоченности набора данных [21]. Энтропия отображения H рассчитывается по формуле Шеннона

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

где $p_i = \frac{N_i}{N}$ – относительное число попаданий входного сигнала в i-й кластер, N – общее количество всех значений сигнала.

5 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

В данной работе рассматривается разработка, которая представляет из себя компьютерное приложение. Созданная программа содержит в себе функции, реализующие кластерный анализ многомерных числовых данных рекуррентной нейронной сетью. Данный алгоритм использует в качестве входных данных вектор евклидовых расстояний от одного объекта кластеризации до всех остальных объектов выборки – числовые параметры многомерных объектов представляются как координаты точек в многомерном пространстве, позволяя определить между ними расстояние – меру близости, по которой и происходит кластерный анализ [22].

Главная особенность разработки состоит в том, что определение кластеров происходит с большой точностью, а также заранее не требуется задавать число кластеров, так как оно определяется автоматически, что позволяет использовать её в широком кругу задач и сфер деятельности, таких как археология, медицина, психология, химия, биология, государственное управление, филология, антропология, маркетинг, социология, геология, информатика и т.д.

Рассматриваемая разработка написана на высокоуровневом кроссплатформенном языке, что позволяет ее использовать на различных операционных системах.

Целью данного раздела выпускной квалификационной работы является создание и проектирование конкурентоспособных разработок, отвечающих требованиям в области ресурсосбережения и ресурсоэффективности. Также целью является оценка коммерческого потенциала разработки, перспективности проекта.

5.1 Анализ конкурентных технических решений

Методов кластеризации многомерных данных насчитывается достаточно немного, каждый из которых имеет свои преимущества и недостатки перед другими методами, решая задачи почти любой сложности. Однако, если рассматривать общий функционал этих алгоритмов, то можно сказать, что идеального алгоритма не существует. Кластеризация – это первоначальный анализ данных, предназначенный для лучшего понимания изучаемых процессов и явлений, результаты которого требуются в дальнейшем анализе и проверке областными экспертами.

В качестве конкурентов разработки целесообразно рассмотреть конкурентные технические решения, представляющие собой методы и алгоритмы, с помощью которых проводится кластерный анализ. Сравним предложенный метод с такими известными методами как k-means и DBSCAN, которые уже устоялись и повсеместно используются.

Первый метод, называемый k-means, является алгоритмом квадратичной ошибки, где задача кластеризации рассматривается как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратичной ошибки разбиения.

Алгоритмы среднеквадратичной ошибки относятся к типу плоских алгоритмов, которые строят одно разбиение объектов на кластеры (в отличие от иерархических). Самый распространенный среди них – k-means. Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга, после чего определяет оптимальные «центры масс» для этих кластеров.

Второй метод-конкурент был предложен как решение проблемы пространственного разбиения данных на кластеры произвольной формы. Большинство алгоритмов, производящих плоское разбиение, создают кластеры по форме близкие к сферическим, так как минимизируют расстояние объектов до центра кластера. Авторы DBSCAN (Density Based Spatial Clustering of

Applications with Noise) экспериментально доказали, что их алгоритм способен распознать кластера различной формы.

Особенностью метода DBSCAN является достаточно сильное разбиение всей выборки как на маленькие, так и на большие группы (в зависимости от установленного порогового значения), при чем различных форм с выделением сильных концентраторов.

Экспертная оценка основных технических характеристик данных продуктов представлена в таблице 5.1.

Табл.5.1 – Оценочная карта сравнения конкурентных технических решений

Критерии оценки	**Вес**	**Баллы**			**Конкуренто-способность**		
		$Б_Ф$	$Б_1$	$Б_2$	$К_Ф$	$К_1$	$К_2$
Технические оценки ресурсоэффективности							
Вычислительная сложность	0,2	5	3	3	1	0,6	0,6
Скорость работы	0,15	5	4	3	0,75	0,6	0,45
Точность вычислений	0,3	4	3	4	1	0,6	1,2
Работа с большими объемами данных	0,1	5	4	3	0,5	0,4	0,3
Универсальность	0,25	3	4	3	0,75	1	0,75
Итого	**1**	22	18	16	4	3,2	3,3
Экономические критерии оценки ресуроэффективности							
Конкурентоспособность продукта	0,5	5	4	4	2,5	2	2
Поддержка продукта	0,1	4	3	2	0,4	0,3	0,2
Срок выхода на рынок	0,1	1	5	5	0,1	0,5	0,5
Сертификат разработки	0,3	5	4	5	1,5	1,2	1,5
Итого	**1**	15	16	16	4,5	4	4,5

Из анализа можно сделать вывод, что продукты конкурентов проигрывают по техническим критериям. Однако по экономическим оценкам видна одинаковая оценка конкурентоспособности, так как из-за того, что предложенный разрабатываемый алгоритм все еще находится в стадии разработки и не был выпущен в массовое потребительство.

5.2 Технология QuaD

Чтобы проанализировать разработку были выделены следующие критерии оценки. Показатели качества разработки:

1) Надежность – способность системы работать без отказа и ошибок во время обработки текущих задач. Способность выполнять требуемые функции в заданных режимах и условиях применения.
2) Качество пользовательского интерфейса – удобство работы пользователя с системой, интуитивность использования элементов и их адекватность, возможность запомнить расположение элементов.
3) Точность вычислений – способность системы обеспечить определенную точность вычислений на заданных данных и минимизация случайных и систематических ошибок. Предоставление результирующих данных с заданной точностью.
4) Время обработки – необходимый интервал времени, в течении которого система обрабатывает данные.
5) Универсальность выполняемых задач – способность системы производить операции над разнородными данными, независимость от типа исследуемого объекта, однородность представленных результатов.
6) Требования к ресурсам – требования, которые выдвигает построенная система к аппаратно-программным средствам, на которых она производит вычисления.

Показатели коммерческого потенциала разработки:

1) Перспективность рынка – оценка заинтересованности клиентов к данной разработке в рассматриваемом сегменте рынка

2) Законченность работы – характеризует ту стадию разработки, на которой находится система.

3) Конкурентоспособность – свойство, характеризующее степень удовлетворения разработкой в сравнении с аналогичными, представляемыми на данном рынке

4) Доступность – свобода продавать на рынке.

5) Финансовая эффективность – соотношение реальной или предполагаемой прибыли к затратам.

Табл.5.2 – Оценочная карта

Критерии оценки	Вес	Баллы	Макс. балл	Отн. знач.	Ср.-взвеш. знач.
Показатели оценки качества разработки					
1. Надежность	0,1	80	100	0,8	0,08
2. Качество пользовательского интерфейса	0,05	40	100	0,4	0,02
3. Точность вычислений	0,2	90	100	0,9	0,18
4. Время обработки	0,1	90	100	0,9	0,09
5. Универсальность выполняемых задач	0,15	70	100	0,7	0,105
6. Требования к ресурсам	0,2	90	100	0,9	0,18
Показатели оценки коммерческого потенциала					
1. Перспективность рынка	0,04	100	100	1	0,04
2. Законченность работы	0,01	100	100	1	0,01
3. Конкурентоспособность	0,1	100	100	1	0,1
4. Доступность	0,04	20	100	0,2	0,008
5. Финансовая эффективность	0,01	50	100	0,5	0,005
				Итог:	0,818

Перспективность и качество разрабатываемого продукта технологии QuaD определяется по формуле:

$$П_{ср} = \sum В_i Б_i,$$

где $П_{ср}$ – взвешенное значение показателя перспективности и качества разработанного продукта, $В_i$ – вес показателя (в долях единицы), $Б_i$ – взвешенное значение i-го показателя.

Полученная оценка равна 0,818, что соответствует перспективному проекту. Такая оценка качества означает, что в данный проект рекомендуется инвестирование в её дальнейшее улучшение.

5.3 SWOT-анализ

Результаты анализа представлены в таблице 5.2.

Табл.5.3 – SWOT-анализ проекта

	Сильные стороны: 1. Высокая точность 2. Кроссплатформенность 3. Быстрота работы 4. Работа с большим количеством данных 5. Готовность к работе	**Слабые стороны:** 1. Сложный интерфейс пользователя 2. Малая доступность 3. Основан на эмпирическом законе 4. Недостаточный функционал 5. Малая доступность
Возможности: 1. Улучшение существующего алгоритма	В3С1: коммерциализация разработки в сочетании с высокой точностью может обеспечить прибыль,	В1Сл3: улучшение алгоритма позволит уйти от эмпирического до полного закона

2. Улучшение пользовательского интерфейса 3. Коммерциализация разработки 4. Увеличение функционала 5. Визуализация объектов кластеризации	улучшить конкурентоспособность и снабдить средствами для дальнейших улучшений В2С2: улучшение интерфейса позволит быстрее распространить программу на все платформы	В2Сл1: улучшение пользовательского интерфейса программы позволит улучшить эффективность работы в ней В1Сл4: добавление нового функционала позволить значительно расширить область применения
Угрозы: 1. Недостаточная точность в сравнении с конкурентами 2. Эмпирический закон окажется неверным 3. Запрет на распространение 4. Понижение стоимости конкурентных разработок 5. Развитая конкуренция технологий	У1С5: несмотря на высокую точность алгоритма, не идеальность метода может привести к тому, что потребители будут выбирать конкурирующие продукты У2С1: эмпирически сложилось, что метод оказался высокоточным, однако это не может означать, что точность останется такой же во всех остальных случаях	У1Сл4: Недостаточная точность в купе с небогатым функционалом приведет к уходу от предложенного продукта к продуктам конкурентов У2Сл3: если эмпирический закон окажется неверным, то это приведет к полному отвержению предложенного метода

Таким образом можно прийти к выводу, что основными рисками при дальнейшей разработке и продвижении системы является возможная ошибка в эмпирическом законе и недостаточная точность вычислений, предоставляемых разработкой, а также возможность появления метода, лишённого всех недостатков предложенного алгоритма.

5.4 Определение возможных альтернатив проведения научных исследований

В данной работе рассматривается реализация алгоритма кластеризации (группировки) многомерных числовых данных. Существует множество методик для решения данной задачи. Для анализа уместно использовать морфологический подход как инструмент для определения возможных альтернатив, которые, на первый взгляд, кажутся незаметными.

Раскроем наиболее значимые характеристики объекта исследования.

Для реализации используется алгоритм кластеризации многомерных числовых данных. Существует классификация этих алгоритмов, каждый из которых имеет свои недостатки и преимущества.

Иерархические методы хоть и имеют высокую точность, но для правильного выбора кластеризации требуется экспертный анализ результатов кластеризации, так как они достаточно излишни – представляют огромное множество вариантов разбиения на группы. Алгоритм квадратичной ошибки k-mean требует задания количества кластеров перед началом анализа, так еще и отчасти содержит в себе метод случайного разбиение. Предложенный метод на рекуррентной нейронной сети не требует задания каких-либо параметров, так как они определяются автоматически при передаче объектов в программу

Для языка реализации может использоваться как платформонезависемый язык Java, так и язык высокого уровня C++, который имеет большие возможности, как и по скорости работы, так и по платформонезависимости.

Так как реализуемый метод должен иметь высокую точность, то функция активации нейрона должна быть нелинейной, плавной, и должна находится в только в положительной части системы координат [0, 1]. Сигмоида полностью устраивает по условиям. Функция гиперболического тангенса покрывает интервал [-1, 1], что потребует введение коэффициентов смещения и сжатия функции, чтобы она стала удовлетворять условиям работы. Кусочно-линейная функция сильно снизила точность, то так же сильно упростила вычисления.

Для удобства, выходные данные должны быть представлены в удобном формате. Из возможных, данные могут быть представлены в численном варианте, в виде текстового файла, или же в виде визуализации как изображение, на котором показана проекция многомерного пространства (пространство параметров объектов) на двумерное (плоскость проекции – изображение). Существует альтернативный метод, который сочетает в себе разнородные данные – иерархическая структура данных. В результате, выходные данные могут храниться вместе сходными и могут быть представлены визуализатором в дальнейшем или же другими аналогичными программами.

Для оценки качества кластеризации можно использовать 3 метода: скорость работы, точность кластеризации, работа с огромной выборкой. Один алгоритм кластеризации может работать очень быстро, но не сможет выполнить кластеризацию большой выборки, ведь кластеризация большой выборки – это не относящийся к скорости критерий, а относящийся к возможности алгоритма. Любой алгоритм будет перспективен, если хотя бы по одному критерию он будет иметь хорошую оценку.

Далее представлена таблица 5.4, в которой описаны различные морфологические характеристики и их возможные варианты.

Табл.5.4 – Морфологическая матрица

	1	2	3
А. Алгоритм	Рекуррентная нейронная сеть	Иерархический	Квадратичная ошибка
Б. Язык реализации	C#	C++	Java
В. Функция активации	Сигмоида	Гиперболический тангенс	Линейно-кусочная
Г. Представление	Текстовый файл	Изображение	Иерархическая структура данных
Д. Оценка качества	Точность вычислений	Скорость выполнения	Работа с огромной выборкой

5.5 Календарный план-график работ

Ниже приведен план-график работ (табл.5.5) по экономической части ВКР.

Табл.5.5 – План-график работ по экономической части ВКР

Описание работ		Недели					
		Апрель		Май		Июнь	
		1-2	3-4	5-6	7-8	9-10	11-12
1	Введение						
2	Анализ конкурентов						
3	QuaD						
4	SWOT						
5	Морфологический анализ						
6	Выводы						

Так же представим диаграмму Ганта на рисунке 5.1.

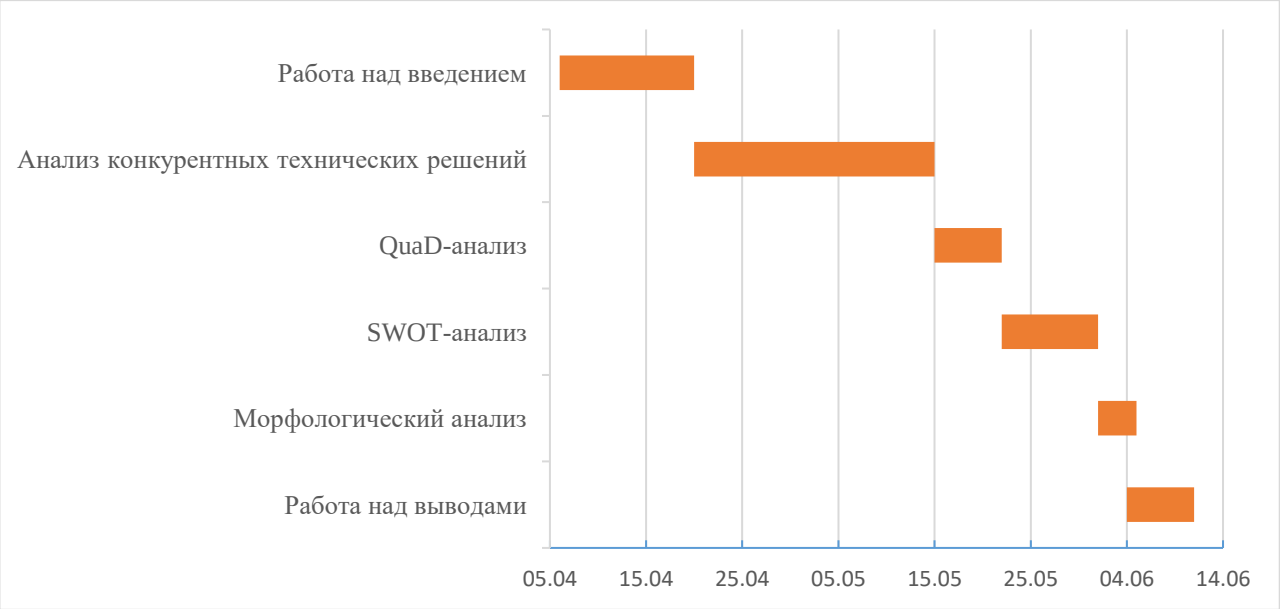

Рис.5.1 – Диаграмма Ганта. График работ

5.6 Вывод

С помощью рассмотренных методов был проведен практический анализ проекта и выявлен его коммерческий потенциал и перспективность. Проведен SWOT и QuaD-анализ разработанной системы, а также сравнение с конкурирующими технологиями. Дополнительно проведен морфологический анализ для выявления перспективных направлений проведения исследований.

В результате SWOT-анализа выявлены сильные и слабые стороны разработанной системы, позитивные и негативные риски, связанные с ее жизненным циклом. Выявлены и описаны зависимости между сильными и слабыми сторонами и рисками.

В ходе QuaD-анализа составлена оценочная карта разработанного продукта. В результате проект получил достаточно высокую оценку.

Морфологический подход определения возможных альтернатив проведения научных исследований показал, что у проекта есть множество путей развития, при этом каждый из вариантов имеет свои преимущества.