На правах рукописи

Харахинов Владимир Александрович

НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ РЕШЕНИЯ ЗАДАЧ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ ДАННЫХ В ТЕХНИЧЕСКИХ СИСТЕМАХ

Специальность 2.3.1 – Системный анализ, управление и обработка информации, статистика (технические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Иркутск – 2023

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Иркутский национальный исследовательский технический университет» (ФГБОУ ВО «ИРНИТУ»).

Научный руководитель: **Сосинская Софья Соломоновна**, кандидат технических наук, доцент кафедры информационных технологий института математики и информационных технологий ФГБОУ ВО «Иркутский государственный университет»

Официальные оппоненты: **Дегтярёв Александр Борисович**, доктор технических наук, доцент, профессор кафедры компьютерного моделирования и многопроцессорных систем Федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет», г. Санкт-Петербург

Найханова Лариса Владимировна, доктор технических наук, профессор, профессор кафедры программной инженерии и искусственного интеллекта Федерального государственного бюджетного образовательного учреждения высшего образования «Восточно-Сибирский государственный университет технологий и управления», г. Улан-Удэ

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный аграрный университет имени А.А. Ежевского»

Защита диссертации состоится 28 сентября 2023 г. в 13.00 часов на заседании диссертационного совета 44.2.002.01, созданного на базе ФГБОУ ВО «Иркутский государственный университет путей сообщения» по адресу: 664074, Иркутская область, г. Иркутск, ул. Чернышевского, 15, ауд. А-803, тел. 8 (3952) 63-83-94, e-mail: diss_sovet@irgups.ru.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Иркутский государственный университет путей сообщения» и на сайте https://www.irgups.ru.

Отзывы на автореферат в двух экземплярах с подписью составителя, заверенной печатью организации, просим направлять по адресу: 664074, Иркутская область, г. Иркутск, ул. Чернышевского, 15, на имя Ученого секретаря диссертационного совета.

Автореферат разослан «__» _____ 2023 г.

Ученый секретарь
диссертационного совета,
доктор технических наук, доцент Л.В. Аршинский

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. На сегодняшний день интеллектуальный анализ данных является междисциплинарной областью знаний, основу которой заложили следующие научные дисциплины: математическая статистика, искусственный интеллект, машинное обучение, визуализация данных. В число задач интеллектуального анализа данных входят задачи классификации и кластеризации.

Задача классификации является наиболее распространенной задачей анализа данных. Для ее решения используются признаки, которыми описываются объекты исследуемого набора данных и выделяются группы объектов - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Кластеризация является логическим продолжением идеи классификации. Особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на классы.

Необходимость решения задач классификации и кластеризации в той или иной сфере человеческой деятельности существует постоянно: определения спама, распознавания голоса, кредитоспособности заемщика и т.д. Нередко решение этих задач связано с потребностью принимать решения различной сложности. Для снижения вероятности принятия неверных решений применяют СППР (системы поддержки принятия решений), в ряде которых активно используются совместно различные методы интеллектуального анализа данных. Их решение зачастую позволяет повысить уровень автоматизации того или иного процесса, тем самым оказывая положительный эффект на всю систему, к которой относится этот процесс.

Вопросы классификации и кластерного анализа рассматриваются в работах многих отечественных и зарубежных авторов: С.А. Айвазяна, М.А. Айзермана, М.М. Бонгарда, С. Виежхона, Б. Дюрана, М. Жамбю, Ю.И. Журавлёва, Н.Г. Загоруйко, Л. Кауфмана, И.Д. Манделя, Б.Г. Миркина, М.С. Олдендерфера, В. Рэнда, Р. Триона, М. Штейнбаха и многих других.

Приведенные выше задачи анализа данных решаются путем применения самых разнообразных подходов. На текущий момент одним из самых эффективных является нейросетевой подход.

Вопросам применения нейронных сетей для решения различных задач анализа данных уделено внимание в работах: А.Н. Горбаня, А. Кофмана, Т. Кохонена, В. Маккалока, М. Моллера, У. Питтса, Ф. Розенблатта, Ф. Уоссермена, С. Хайкина, Д. Хинтона, Д. Хопфилда.

Несмотря на известные достоинства нейронных сетей, в ряде случаев они обладают свойствами, приводящими к нежелательным результатам, например, от обучающей выборки зависит архитектура сети, количество слоев, количество нейронов в каждом слое. Процесс обучения сети может протекать достаточно медленно из-за большой размерности входных данных. Одним из основных способов уменьшения размерности данных является проведение процесса редукции данных. Обучение на ненормализованных данных может снизить качество решения задач классификации и кластеризации объектов. Качество решения может также зависеть от выбора начальной конфигурации сети, начальных весов нейронов, выбора необходимых функций активации.

Для того, чтобы избавиться от этих нежелательных свойств совместно с нейронными сетями, используют другие методы интеллектуального анализа данных. Вопросами совместного применения различных методов интеллектуального анализа данных занимались следующие авторы: З. Гуо, Н. Кадаба, Д. Келли, Ж. Корбич, К. Сузуки, Л. Миддлтон, Д. Хинтон, Р. Эберхарт.

Таким образом, актуальность выбранной темы диссертационной работы обусловлена необходимостью разработки рациональной методики обработки данных при инициализации параметров нейронных сетей в задачах классификации и кластеризации.

Цель диссертационной работы – повышение качества решения задач классификации и кластеризации в технических системах за счет совместного использования редукции, нормализации анализируемых данных автокодировщиком и настройки параметров слоя Кохонена нейронной сети с применением генетического алгоритма.

Для достижения поставленной цели в работе решались следующие задачи:

1) Разработка методики решения задачи кластеризации на основе нейросетевой технологии с применением генетического алгоритма настройки параметров нейронной сети;

2) Редукция и нормализация анализируемых данных на основе нейросетевых технологий с последующей оценкой влияния на качество кластерного анализа;

3) Разработка модели функционирования системы анализа данных с помощью математического аппарата сетей Петри;

4) Разработка алгоритмического и программного обеспечения, нейронной сети для решения задач классификации и кластеризации, а также для принятия управленческих решений по оперативному реагированию на дорожно-транспортные происшествия;

5) Апробация разработанной методики, предложенных методов и СППР в предметных областях: транспорт, машиностроение, биология, сельское хозяйство, банковское дело.

Объект и предмет исследования. Объект исследования – технические объекты, характеризуемые множеством признаков состояния. Предмет исследования – нейросетевые технологии решения задач классификации и кластеризации.

Методы исследования: теория сетей Петри, теория информации, теория искусственных нейронных сетей, методы редукции пространства признаков состояния, генетические алгоритмы.

Достоверность результатов. Достоверность полученных результатов подтверждена совпадением результатов решения задач классификации и кластеризации, которые были получены другими авторами, а также корректностью архитектуры спроектированной системы «Анализ данных».

Тематика работы соответствует следующим пунктам паспорта специальности 2.3.1: п. 3 «Разработка критериев и моделей описания и оценки эффективности решения задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта», п. 5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта», п. 10 «Методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических системах», п. 12 «Визуализация, трансфор-

мация и анализ информации на основе компьютерных методов обработки информации».

Научная новизна диссертационной работы представлена следующими положениями, выносимыми на защиту:

1) Впервые разработана методика совместного использования слоя Кохонена и генетического алгоритма с редукцией данных, повышающая качество результатов проводимого кластерного анализа объектов.

2) Предложено использование автокодировщика в качестве эффективного альтернативного способа нормализации анализируемых данных по отношению к общеизвестным способам.

3) Реализован новый подход к редукции данных для задач классификации и кластеризации.

4) Спроектировано и разработано специальное алгоритмическое обеспечение системы анализа, управления, принятия решений и обработки информации, отличающее совместным использованием общеизвестных и предложенных автором методик для классификации и кластеризации технических объектов.

Теоретическая значимость работы заключается в разработке методики совместного использования генетического алгоритма и алгоритма К-средних для настройки параметров самоорганизующегося слоя Кохонена с целью повышения качества результатов кластеризации. Предложено использование автокодировщиков в качестве эффективного альтернативного способа нормализации анализируемых данных.

Практическая значимость состоит в разработке инструментальных средств, позволяющих исследователям проводить обработку и анализ данных с различными методами классификации и кластеризации, а также для принятия управленческих решений по оперативному реагированию на серьезные ДТП, существенно затрудняющих пропускную способность автомобильной дороги. Разработанные в диссертации концепция, алгоритмы и методы, а также программные модули могут использоваться при разработке математического и программного обеспечения систем анализа данных в различных отраслях промышленности и иных предметных областях.

Разработанные методы классификации и кластеризации обработки данных использовались при моделировании транспортных потоков в компании ООО «Центр транспортных технологий»; компанией ООО НПО ССЦ «Ангара» при анализе массивов данных в различных районах Иркутской области; в учебном процессе Института высоких технологий Иркутского национального исследовательского технического университета (ИРНИТУ) при организации учебного курса «Технологии обработки информации».

Результаты исследования подтверждаются наличием соответствующих актов о внедрении.

Апробация работы. Работа выполнялась в ИРНИТУ на кафедре технологии и оборудования машиностроительных производств и вычислительной техники. Основные положения проведенных исследований докладывались на научных семинарах кафедры технологий и оборудования машиностроительных производств и вычислительной техники ИРНИТУ, на Всероссийских молодежных научно-практических конференциях «Винеровские чтения» (г. Иркутск, 2016, 2019), на Международной научной конференции «Applied Physics, Information Technologies and Engineering» (г. Красноярск, APITECH-2019), на VIII Всероссийской научной

конференции с международным участием «Информационные технологии интеллектуальной поддержки принятия решений» (г. Уфа, ITIDS 2020), на VIII международном семинаре «Критические инфраструктуры в цифровом мире» (г. Байкальск, IWCI 2021), на XII международной научно-практической конференции «Транспортная инфраструктура Сибирского региона» (г. Иркутск, 2021).

Личный вклад. Результаты, составляющие научную новизну и выносимые на защиту, получены лично автором. В остальных работах, полученных совместно другими авторами, автору принадлежат от 40 до 90% полученных научных результатов.

Сведения о публикациях. Результаты диссертационного исследования опубликованы в 10 научных работах, из них 1 статья в журнале, индексируемом международной базой Scopus; 4 статьи в изданиях, входящих в Перечень ВАК. Получено свидетельство о государственной регистрации программы для ЭВМ № 2017617294.

Структура и объем работы. Диссертационная работа состоит из введения, трех глав, заключения и списка литературы из 107 наименований, 5 приложений. Объем работы составляет 139 страниц, 41 рисунок и 17 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, на основании чего сформулированы цель и задачи исследования, определены объект, предмет, методы и средства исследования, научная и практическая значимость работы, изложены научные положения, выносимые на защиту.

В первой главе описаны основные задачи интеллектуального анализа данных и подходы к их решению. Более детально описаны задачи классификации и кластеризации, особое внимание уделяется алгоритму К-средних. Выполнена проверка качества решения задачи кластеризации с помощью индексов Рэнда.

Изложена концепция искусственных нейронных сетей, даны основные понятия и описаны три фундаментальных класса архитектур сетей, а также две парадигмы их обучения. Более детально описано применение нейронных сетей для решения задач классификации и кластеризации – рассмотрены архитектуры основных типов сетей и наиболее популярные алгоритмы для их обучения.

Раскрывается важность процесса редукции данных в задачах интеллектуального анализа данных, решаемых при помощи нейронных сетей.

Вторая глава посвящена проектированию и разработке системы для классификации и кластеризации технических объектов, а также созданию методики предварительной инициализации матрицы весов слоя Кохонена для повышения качества результатов кластерного анализа при использовании данной сети, с последующим включением созданной методики в разрабатываемую систему.

При проектировании была построена структурная схема системы, позволяющая наглядно отобразить общий принцип работы системы (рисунок 1).

В ней выделены 4 подсистемы: препроцессинга; редукции данных; обучения и тестирования нейронных сетей; экспорта результатов анализа.

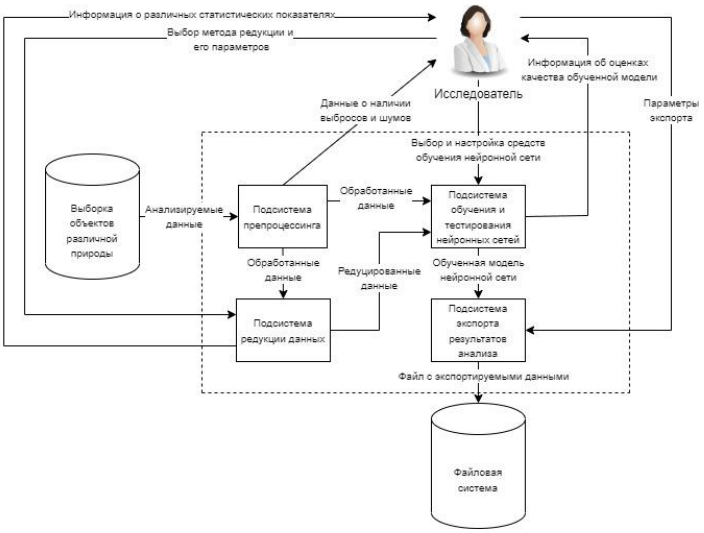

Рисунок 1 – Структурная схема СППР

После определения структурной схемы системы были более детально описаны процессы, происходящие в ней, в особенности в подсистеме обучения и тестирования нейронных сетей, поскольку именно в ней реализована новая методика инициализации весов слоя Кохонена. Детальное описание процессов позволило лучше изучить поведение разрабатываемой системы. С этой целью была разработана диаграмма потоков данных (DFD), которая изображена на рисунке 2.

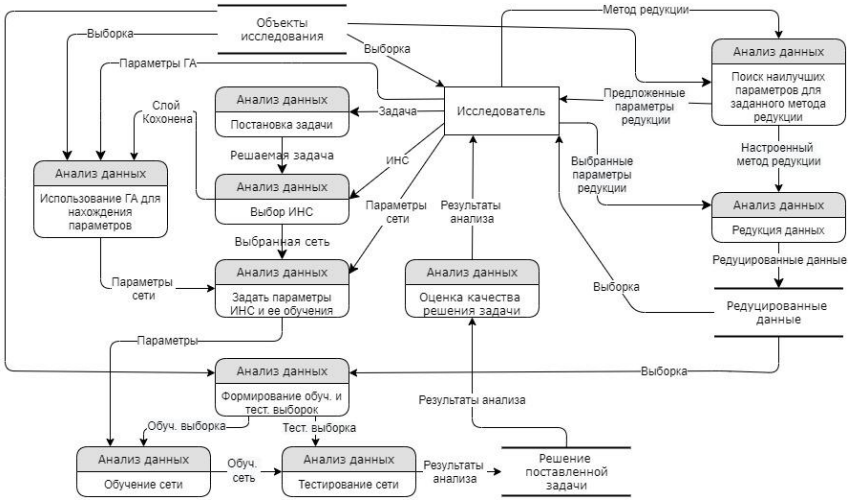

Рисунок 2 – Диаграмма DFD для СППР

Центральным элементом в ней является внешняя сущность – исследователь, которая взаимодействует с разрабатываемой системой. Также используется 3 хранилища:

1) объекты исследования (выборка технических объектов);

2) редуцированные данные (выборка меньшей размерности);

3) результат решения поставленной задачи (числовой вектор номеров классов для каждого объекта).

Далее в главе описаны разработанный автором алгоритм совместного использованию нейронных сетей и генетического алгоритма, приведена методика использования генетического алгоритма для предварительной настройки матрицы весов слоя Кохонена.

При использовании стандартных способов инициализации весов слоя Кохонена качество кластерного анализа может быть ниже требуемого уровня. Разработанная методика решает эту проблему путем применения генетического алгоритма, реализующего основные этапы алгоритма K-средних. Выбор алгоритма K-средних обоснован тем, что имеет схожий принцип работы с алгоритмом слоя Кохонена (таблица 1).

Таблица 1 – Сравнение алгоритмов работы слоя Кохонена и K-средних

Критерий сходства	Слой Кохонена	Алгоритм K-средних
Предварительное определение числа кластеров	Обязательно предварительное определение числа кластеров (число нейронов в выходном слое)	Обязательно предварительное определение числа кластеров
Определение центра кластеров	Координаты случайного объекта из выборки	Координаты случайного объекта из выборки
Отнесение объекта к кластеру	Активация нейрона-победителя в выходном слое на основе расстояний от анализируемого объекта	Основано на выборе минимального расстояния до всех центров кластеров от анализируемого объекта

Сравнение принципов работы алгоритма K-средних с принципами работы генетического алгоритма в качестве алгоритма обучения слоя Кохонена позволило выявить ряд особенностей (таблица 2).

Таблица 2 – Сравнение этапов работы алгоритмов

Этап	Алгоритм K-средних	Генетический алгоритм
Начало работы алгоритма	Определение координат центров кластеров	Формирование начальной популяции
Проверка останова алгоритма	Координаты центров кластеров на текущей итерации не отличаются от соответствующих координат на предыдущей итерации алгоритма	Формирование новой популяции не приводит к значимому улучшению значения функции приспособленности
Изменение центров кластеров	Перерасчет координат происходит по четко определенной формуле	Селекция, скрещивание, мутация и формирование новой популяции

Таким образом, было выявлено, что оба алгоритма имеют 3 общих этапа работы. На каждом этапе генетический алгоритм, в отличие от алгоритма K-средних, оперирует популяциями, которые образуют конечное множество объектов в n-мерном пространстве. Каждая особь популяции - некоторый потенциальный центр кластера. Хромосома - одна из координат потенциального центра кластера в десятичном формате.

В работе графически представлен процесс кластерного анализа с помощью слоя Кохонена, прошедшего предварительную инициализацию весов матрицы по разработанной методике, в виде схемы, приведенной на рисунке 3.

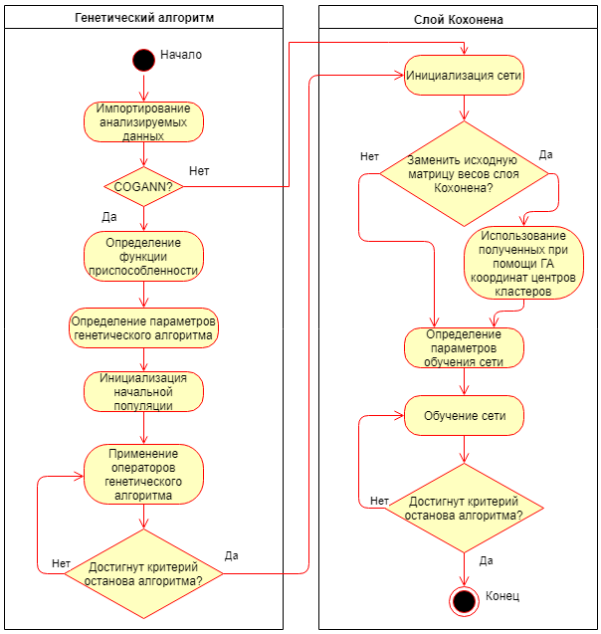

Рисунок 3 - Схема совместного использования слоя Кохонена и генетического алгоритма

На формирование новой популяции оказывает влияние функция приспособленности.

В случае K-средних распространен критерий – минимизация суммы квадратов расстояний от точек до центров кластеров, к которым они относятся. Иными словами, задачу кластерного анализа методом K-средних можно свести к задаче:

$$F = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - c_i)^2 \rightarrow min, \quad (1)$$

где k - число кластеров, S_i - полученные кластеры (подмножества множества всех анализируемых наблюдений), x_j - наблюдения, c_i - центры кластеров.

На основании приведенного критерия минимизации была сформирована функция приспособленности. В данном случае целью генетического алгоритма является минимизация функции F (см. формулу (1)) и нахождение ее глобального экстремума.

В разработанной методике селекция производилась согласно формуле:

$$p_s(ch_i) = \frac{F(ch_i)}{\sum_{i=1}^{N} F(ch_i)},$$

где N – численность популяции, $F(ch_i)$ – значение функции приспособленности хромосомы ch_i, а $p_s(ch_i)$ – вероятность селекции этой хромосомы.

Для оператора мутации был использован принцип Гауссовской мутации, согласно которому оператор мутации добавляет полученное с помощью распределения Гаусса случайно число к каждому элементу родительского вектора.

Математически это можно описать следующим образом:

$$\widetilde{x_i} = x_i + N(\mu, \sigma^2),$$

где $\widetilde{x_i}$ – полученный вектор, x_i – родительский вектор, $N(\mu, \sigma^2)$ – число полученное с помощью распределения Гаусса.

Апробация данной методики произведена при решении задачи кластеризации в различных предметных областях, таких как: сфера транспорта, машиностроение, биология, банковское дело, сельское хозяйство.

В работе детально описана стадия проектирования системы с помощью аппарата сетей Петри. На рисунках 4а-4б приведена сеть Петри, которая содержит 52 вершины. Овалами обозначены позиции, а прямоугольниками – переходы. В начальном состоянии сети лишь одна позиция сети «Выбор в главном меню» содержит метку.

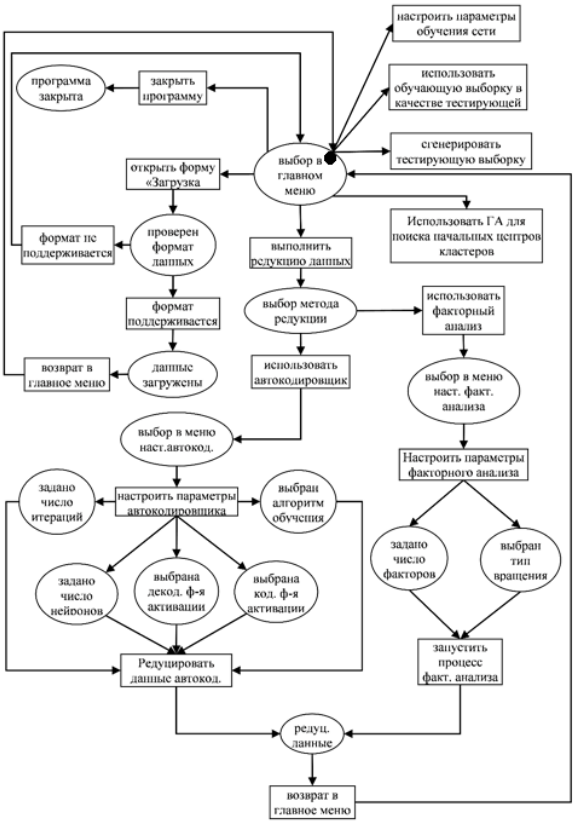

Рисунок 4а – Сеть Петри, описывающая проектируемую систему

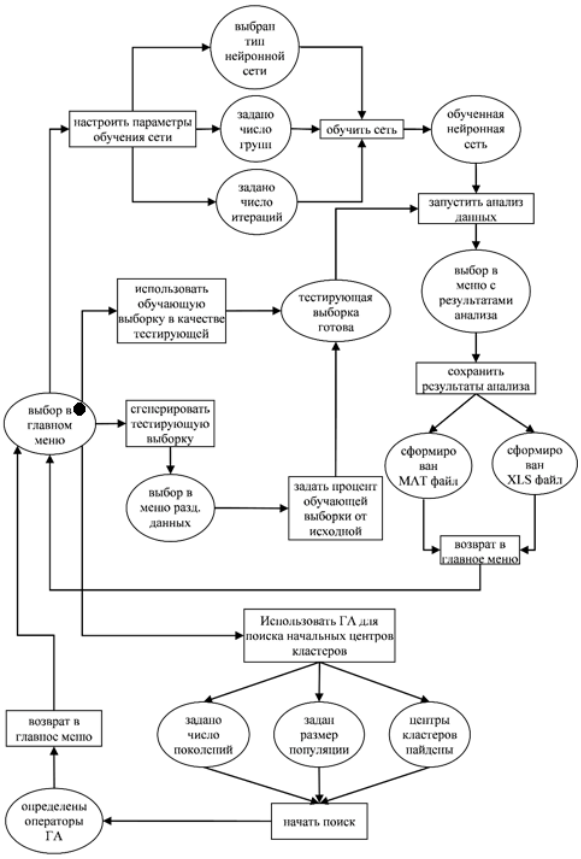

Рисунок 4б – Сеть Петри, описывающая проектируемую систему (продолжение)

Переход может запускаться только в том случае, если он разрешен.

Переход $t_j \in T$ в маркированной сети Петри с маркировкой μ разрешен, если для всех $p_i \in P$:

$$\mu(p_i) \geq \#\left(p_i, I(t_j)\right)$$

В среде CPN Tools версии 4.0.1 был проведен анализ сети Петри, а именно была решена задача достижимости путем моделирования спроектированной сети.

Узлы дерева в среде CPN Tools отображаются в виде (рисунок 5):

Рисунок 5 – Отображение узла дерева в среде CPN Tools

На рисунке 5 указаны: N – числовой номер узла дерева, IN – число входных дуг (ветвей), OUT – число выходных дуг.

Результатом анализа стало дерево достижимости, построенное на основе сети Петри (рисунок 6).

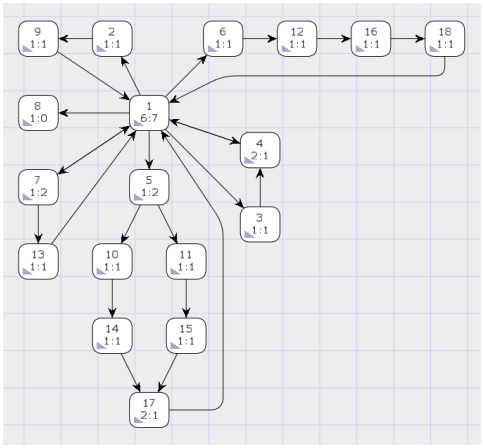

Рисунок 6 – Дерево достижимости в среде CPN Tools

Как показал анализ, в построенной сети Петри отсутствуют мертвые переходы; это означает, что архитектура системы была спроектирована корректно.

В третьей главе дано описание структуры и возможностей реализованной автором системы. Реализация произведена в среде MATLAB в связи с наличием в ней широкого спектра инструментов для анализа данных.

Апробация реализованной системы была проведена путем решения задач классификации и кластеризации технических объектов на примере решения задач анализа дорожно-транспортных происшествий (ДТП). В ходе апробации для объектов ДТП была предложена концептуальная модель СППР (рисунок 7), включающая в себя разработанную автором систему анализа данных.

Разработанная концептуальная модель СППР позволяет наглядно отобразить принцип работы системы. В ней выделены 3 подсистемы: сбора и обработки данных; анализа данных; управления с целью повышения пропускной способности дорожного участка.

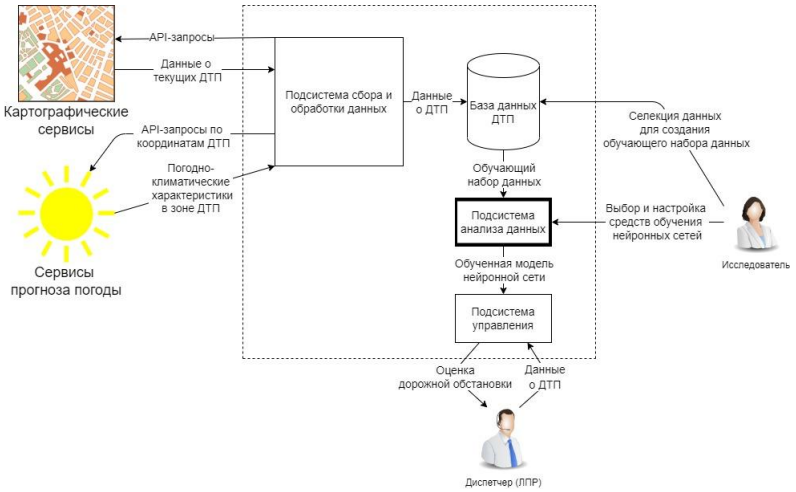

Рисунок 7 – Концептуальная модель СППР для классификации ДТП по их влиянию на пропускную способность дорожного участка

Оценка снижения пропускной способности дорожного участка по географическим и погодно-климатическим характеристикам ДТП на основе методов кластерного анализа

В рамках данного исследования оценка влияния ДТП решалась путем проведения кластерного анализа следующими способами:

- слой Кохонена;
- слой Кохонена, который использует данные после проведения редукции;
- предварительно настроенный слой Кохонена, который использует данные после проведения редукции.

Анализируемые данные являются реальными и были получены из открытого источника[1]. Сбор данных производился с помощью картографического сервиса MapQuest и поисковой системы Bing с 2016 по 2020 гг. Набор данных содержит примерно 3.5 миллионов объектов, каждый из которых описывается 48 признаками, которые подразделяются на 4 группы (территориальные, погодно-климатические, временные и группа средств регулирования дорожного движения). Однако в процессе исследования были выделены 6 географических и погодно-климатических признаков, оказывающих наибольшее влияние на серьезность ДТП: широта (градусы), долгота (градусы), температура (Цельсий), дальность видимости (километры), период дня (светлое время суток/темное время суток), гражданские сумерки (да/нет). К подобному выбору набора признаков, влияющих на серьезность ДТП, пришел Artur Filipowicz[2], основываясь на статистическом подходе.

Поскольку дорожная инфраструктура может быть изменена, то целесообразно оценивать загруженность дорог за наиболее актуальный интервал времени, в частности за последний в выборке год (2019 г.). Также в целях сокращения вычислительных затрат на проведение кластерного анализа из набора данных были взяты объекты, относящиеся к одному субъекту. Целесообразность выбора данного субъекта основывается также на том, что погодно-климатические условия в нем схожи с условиями в большинстве регионов России. Поскольку при составлении выборки сервисы MapQuest и Bing оценивали серьезность ДТП по разным критериям, то корректно сравнивать ДТП, полученные от одного сервиса (в этом исследовании использовался MapQuest). В результате анализируемые данные представляют собой набор из 12094 объектов, где каждое ДТП относится к одному из 2 кластеров:

- ДТП, оказывающие незначительное влияние на снижение пропускной способности дорожного участка;
- ДТП, оказывающие значительное влияние.

Исследование состояло из 2 этапов, на каждом из которых для достоверности полученных результатов произведено 25 экспериментов.

На первом этапе исследования кластерный анализ производился с помощью: слоя Кохонена, как без предварительной нормализации, так и с нормализацией, и после проведения редукции с помощью автокодировщика. Качество проведенной классификации, кластеризации оценивалось: общей долей правильно классифицированных объектов; точностью (доля объектов, истинно принадлежащих данному классу, которые классификатор отнес к этому классу); полнотой (доля объектов ис-

[1] Набор данных «A Countrywide Traffic Accident Dataset». URL: https://kaggle.com/sobhanmoosavi/us-accidents (дата обращения: 24.10.2020).

[2] US Car Accidents Severity Analysis. URL: https://www.kaggle.com/art12400/us-car-accidents-severity-analysis/ (дата обращения: 20.02.2021).

тинно принадлежащих данному классу относительно всех объектов, истинно принадлежащих данному классу); индексами Рэнда: RI (доля объектов, для которых исходное и полученное разбиения согласованы); ARI (мера расстояния между различными разбиениями выборки).

Математически индексы Рэнда можно описать следующим образом:

$$RI = \frac{a+b}{\binom{n}{2}} = \frac{2(a+b)}{n(n-1)}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}$$

В таблице 3 приведены результаты проведения первого этапа анализа.

Таблица 3 – Сравнение результатов первого этапа анализа

Метод кластерного анализа	Среднее знач. RI по экспериментам	Среднее знач. ARI по экспериментам	Среднее время обуч.,с
Слой Кохонена (без нормализации и редукции)	0,5006	0,0459	26,985
Слой Кохонена (нормализация)	0,5067	-0,0048	27,2386
Слой Кохонена (автокодировщик)	0,5023	0,0029	24,7662

Использование редуцированного набора данных позволило сократить издержки времени на обучение слоя Кохонена на 2,2 с, не теряя при этом качество полученных результатов.

На втором этапе исследования был применен генетический алгоритм для нахождения начальной матрицы весов слоя Кохонена.

Результаты экспериментов приведены в таблице 4. В 1 строке таблицы присутствует наилучшее решение задачи кластеризации из первого этапа.

Таблица 4 – Сравнение результатов второго этапа анализа

Метод кластерного анализа	Среднее значение RI по экспериментам	Среднее значение ARI по экспериментам
Слой Кохонена (автокодировщик)	0,5023	0,0029
Слой Кохонена (автокодировщик + генетический алгоритм)	0,5148	0,0086

Получено увеличение средних арифметических значений индексов Рэнда при предварительной настройке матрицы весов Кохонена, что сигнализирует об улучшении качества кластерного анализа.

На рисунке 8 визуализирован результат одного из экспериментов.

Объекты одной группы обозначены синими «*», другой – красными «+».

В результате классификатор, реализованный слоем Кохонена с применением данной методики, будет более точно определять серьезность ДТП по сравнению с классическим его использованием, и, как следствие, более корректно оценивать влияние ДТП на загруженность участка дороги, на котором она произошла.

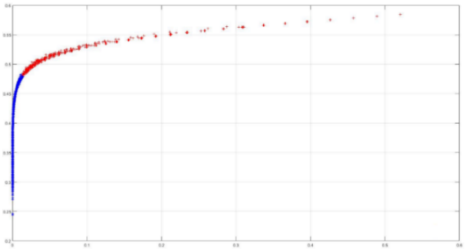

Рисунок 8 – Разделение объектов на кластеры

В большинстве случаев классификатор на основе сетей прямого распространения обеспечивает более высокое качество его работы по сравнению с использованием слоя Кохонена.

Оценка снижения пропускной способности дорожного участка по географическим и погодно-климатическим характеристикам ДТП на основе сетей прямого распространения

Анализ произведен на выборке, описанной в предыдущем исследовании, в которой присутствуют известные уровни серьезности каждого ДТП, образующие целевой вектор, что дает возможность решать задачу классификации.

Для оценки снижения пропускной способности дорожного участка под влиянием ДТП были реализованы классификаторы на основе сетей прямого распространения, которые являются наиболее часто используемыми для решения задачи классификации: слоя софтмакс, сети распознавания образов, многослойного персептрона (сокр. МСП) и каскадной сети прямого распространения сигнала (сокр. КСПРС). Значения признаков были нормализованы с помощью предложенного автором способа использования автокодировщика как альтернативного метода нормализации. Сети обучались по наиболее рекомендуемым в литературе алгоритмам. В обучающую выборку вошли 80% случайно выбранных ДТП из набора данных, в валидационную и тестирующую по 10%.

Первые созданные в рамках данного исследования классификаторы представляли собой сеть с одним скрытым слоем. Краткая информация, содержащая лучшие результаты классификации, приведена в таблице 5. В последующих таблицах полужирным шрифтом выделены наилучшие результаты проведенных экспериментов.

Информация, приведенная в таблице 5, сигнализирует о том, что сети с одним скрытым слоем не могут обеспечить достаточное качество классификации. Увеличение числа нейронов в скрытом слое также заметно не повышает качество классификации.

Таблица 5 – Результаты обучения и классификации по сетям с одним скрытым слоем

Тип сети	Число нейронов	Число итераций обуч.	Время обуч., с	Общая доля прав. классиф. объектов, %	Точность (% по каждому классу)	Полнота (% по каждому классу)
МСП	40	33	2,7368	67,53	71,11; 59,77	79,29; 48,87
КСПРС	**40**	**150**	**9,6492**	**69,85**	**74,19; 61,98**	**78,01; 56,91**

Следующий этап реализации классификатора состоял в обучении нейронных сетей, содержащих в себе два и более скрытых слоя. Краткая информация об их архитектуре, параметрах обучения и результатах классификации, которые были получены при помощи них, приведены в таблице 6.

Таблица 6 – Результаты обучения и классификации по сетям с несколькими скрытыми слоями

Тип сети	Число нейронов	Число итераций обуч.	Время обуч., с	Общая доля прав. классиф. объектов, %	Точность (% по каждому классу)	Полнота (% по каждому классу)
КСПРС	40; 20	35	61,4421	69,81	73,26; 62,82	80; 53,64
	40; 20; 10	49	203,6082	72,63	77,27; 64,96	78,48; 63,35
	40; 20; 10; 5	**54**	**317,1074**	**72,95**	**77,21; 65,69**	**79,33; 62,84**

Соответствующие значения из таблиц 5 и 6 показывают различия в качестве классификации. Таким образом, была найдена модель нейронной сети, позволяющая получать достаточно хорошее качество классификации (по обоим классам) ДТП, производить качественную оценку их влияния на пропускную способность дорожных участков по их географическим и погодно-климатическим признакам.

Использование 4 слоев нейронной сети обеспечивает повышение общей доли правильно классифицированных объектов на 3,1%, а полноты классификации на 1,3%. Имея в распоряжении координаты ДТП, становится возможным использовать не только погодно-климатические признаки, но и другие признаки.

Оценка снижения пропускной способности дорожного участка по использующимся средствам регулирования, а также географическим и погодно-климатическим характеристикам ДТП на основе сетей прямого распространения

В рамках данного исследования по сравнению с предыдущим было произведено количественное изменение набора признаков для описания ДТП: был добавлен набор признаков, описывающих наличие или отсутствие средств регулирования дорожного движения. Таким образом, к используемым ранее 6 признакам были добавлены еще 5, имеющие наибольшее влияние на изменение пропускной способности дорожного участка из представленных в исходной выборке:

1. «Уступи дорогу»;
2. «Перекресток» (наиболее эквивалентные знаки в РФ – «Участок перекрестка», «Пересечение равнозначных дорог» и «Пересечение со второстепенной дорогой»);
3. «Пересечение железнодорожных путей» (наиболее эквивалентные знаки в РФ – «Однопутная железная дорога», «Многопутная железная дорога»);
4. «Движение без остановки запрещено»;
5. «Светофорное регулирование».

Каждый из этих признаков является бинарным (есть / нет).

Для классификации производилось обучение сетей, показавших наилучшие результаты в прошлом исследовании: МСП и КСПРС. Алгоритмы обучения, нормализации и разделения выборки остались без изменений.

Краткая информация о полученных результатах сведена в таблицу 7.

Таблица 7 – Результаты обучения и классификации по сетям с одним скрытыми слоями

Тип сети	Число нейронов	Число итераций обучения	Время обуч., с	Общая доля прав. классиф. объектов, %	Точность (% по каждому классу)	Полнота (% по каждому классу)
МСП	40	41	4,9621	68,44	72; 60,97	79,46; 50,94
КСПРС	**40**	**57**	**7,2908**	**69,18**	**74,08; 60,7**	**76,56; 57,47**

Исходя из результатов классификации, приведенных в таблицах 5 и 7, следует, что качественное изменение набора признаков не позволило существенно повысить качество классификации, реализуемое сетями с одним скрытым слоем. Однако, как показали предыдущие эксперименты (таблицы 5-6) каскадная сеть прямого распространения сигнала обеспечивает наилучшее качество классификации из всех представленных в этой работе типов сетей. Основываясь на этом, целесообразно использовать для классификации данный тип сети. В таблице 8 приведены результаты классификации, полученные при использовании каскадных сетей с несколькими скрытыми слоями.

Таблица 8 – Результаты обучения и классификации по сетям с несколькими скрытыми слоями

Тип сети	Число нейронов	Число итераций обучения	Время обуч., с	Общая доля правильно классиф. объектов, %	Точность (% по каждому классу)	Полнота (% по каждому классу)
КСПРС	**40; 20; 10**	**52**	**347,7558**	**74,98**	**80,21; 67,09**	**78,61; 69,21**
	40; 20; 10; 5	40	308,5396	71,77	77,19; 63,33	76,64; 64,04

Проведенное исследование показало, что при добавлении новой группы признаков точность классификатора повысилась при 3 слоях нейронов на 6%, при 4 – на 3%.

Таким образом, был реализован классификатор, имеющий близкую к 75% общую долю правильно классифицированных объектов в определении влияния ДТП (описываемое только набором признаков, извлекаемых по географическим координатам ДТП) на пропускную способность дорожного участка.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИОННОЙ РАБОТЫ

В диссертационной работе получены следующие результаты.

1. Разработана методика, использующая совместно алгоритм кластеризации К-средних и генетический алгоритм, позволяющая инициализировать матрицу весов слоя Кохонена, обеспечивающая повышение качества решения задачи кластеризации. Данная методика прошла апробацию на различных предметных областях (включая регулирование характеристик дорожных участков). Установлено, что при ее использовании возрастают значения индексов Рэнда.

2. Проведены исследования по кластерному анализу с помощью сети Кохонена на различных предметных областях, в ходе которых сравнивалось качество анализа с использованием данных, нормализованных общепринятыми статистическими подходами, и данных, редуцированных с помощью автокодировщика. При использовании последних значения индексов Рэнда в большинстве случаев выше.

3. Разработана модель функционирования системы «Анализ данных» путем использования математического аппарата сетей Петри, что явилось подтверждением корректности спроектированной архитектуры системы.

4. Основываясь на результатах моделирования, была реализована СППР в виде программного комплекса «Анализ экспериментальных данных на основе нейронных сетей», используемая для построения классификаторов с использованием нейросетевых и эвристических методов.

5. Произведена апробация реализованной системы при классификации и кластеризации технических объектов, в частности было определено, что при классификации ДТП по степени их влияния на пропускную способность дорожного участка в среднем общая доля правильно классифицированных объектов составляет 75%. Дополнительно показана возможность применения разработанной методики при решении задач классификации и кластеризации в других предметных областях (сельское хозяйство, биология, банковское дело).

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях, рекомендованных ВАК:

1. **Харахинов, В. А.** Исследование способов кластеризации деталей машиностроения на основе нейронных сетей / В. А. Харахинов, С. С. Сосинская // Программная инженерия. - 2017. - №4. - С. 170-176.

2. **Харахинов, В. А.** Генетический алгоритм как альтернатива обучения слоя Кохонена / В. А. Харахинов // Информационные технологии. - 2018. - №10. - С. 642-648.

3. **Харахинов, В. А.** Использование сетей Петри при проектировании архитектуры программного продукта для анализа данных с помощью нейронных сетей / В. А. Харахинов, С. С. Сосинская // Научный вестник НГТУ. - 2018. - №4(73). - С. 91-100.

4. **Харахинов, В. А.** Влияние сокращения размерности пространства признаков на результаты классификации листьев различных видов растений / В. А. Харахинов, С. С. Сосинская // Программная инженерия. - 2018. - №2. - С. 82-90.

Свидетельства о регистрации программ для ЭВМ:

5. Свидетельство о государственной регистрации программы для ЭВМ 2017617294 Российская Федерация. Программный комплекс «Анализ экспериментальных данных на основе нейронных сетей» / **В. А. Харахинов**, С. С. Сосинская ; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский национальный исследова-

тельский технический университет». - № 2017614163 ; заявл. 04.05.2017 ; опубл. 04.07.2017. – 1 с.

В изданиях, индексируемых в Scopus:

6. **Kharakhinov, V. A.** Information technology of accounting the impact of data reduction methods on the results of classification of plant leaves / V. A. Kharakhinov, S. S. Sosinskaya, R. S. Dorofeev, A. S. Dorofeev, R. I. Bazhenov // Applied Physics, Information Technologies and Engineering: conference proceeding (Krasnoyarsk, 25-27 September 2019) / Krasnoyarsk Science & Technology City Hall of the Russian Union of Scientific and Engineering Associations. – Krasnoyarsk, 2019. - DOI: 10.1088/1742-6596/1399/3/033006.

В других изданиях:

7. **Харахинов, В. А.** Классификация деталей машиностроительного производства с использованием нейронных сетей / В. А. Харахинов, С. С. Сосинская // Винеровские чтения: материалы VIII Всероссийской молодежной научно-практической конференции, 1-3 июня 2016 г., г. Иркутск. – Иркутск: Изд-во ИРНИТУ, 2016. - С. 19-23.

8. **Харахинов, В. А.** Оценка времени обучения нейронной сети с предварительной редукцией данных в задаче классификации листьев различных видов растений / В. А. Харахинов // Винеровские чтения: материалы XI Всероссийской молодежной научно-практической конференции, 6-7 июня 2019 г., г. Иркутск. – Иркутск: Изд-во ИРНИТУ, 2019.

9. **Харахинов, В. А.** Оценка урожайности картофеля в различных районах Иркутской области с применением методов интеллектуального анализа данных / В. А. Харахинов // Информационные технологии интеллектуальной поддержки принятия решений: сборник трудов VIII Всероссийской научной конференции (с приглашением зарубежных ученых), 6-9 октября 2020 г., г. Уфа. – Уфа: Изд-во УГАТУ, 2021. - Т. 2. –С. 21-27.

10. **Харахинов, В. А.** Нейросетевой подход к оценке пропускной способности дорожного участка по различным характеристикам ДТП / В. А. Харахинов // Молодая наука Сибири. - 2021. - №4(14).

Подписано в печать 21.04.2023 г.
Формат 60х90 1/16. Бумага офсетная.
Печать трафаретная. Усл. печ. л. 1,125
Тираж 100 экз. Заказ № __
Иркутский национальный исследовательский технический университет
664074, г. Иркутск, ул. Лермонтова, 83