МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет компьютерных технологий и прикладной математики

Кафедра вычислительных технологий

**ОТЧЕТ О ПРОХОЖДЕНИИ УЧЕБНОЙ ПРАКТИКИ   
ПРОИЗВОДСТВЕННАЯ ПРАКТИКА**

**(НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА)**

по направлению подготовки

02.03.02 Фундаментальная информатика и информационные технологии

период с 01.04.2024 г. по 28.04.2024 г.

Выполнил студент \_\_\_46\_\_\_ гр. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_Баева Д.Н.

*(подпись) (Ф.И.О. студента)*

Руководитель практики (**производственная практика (научно-исследовательская работа)**)

\_\_\_\_\_\_к.т.н., доцент\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Руденко О.В.

*(ученое звание, должность) (подпись) (Ф.И.О)*

Оценка по итогам защиты практики: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_28\_\_» \_\_\_\_апреля\_\_\_\_\_\_ 2024 г.

Краснодар

2024 г.

ФГБОУ ВО «КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет компьютерных технологий и прикладной математики

Кафедра вычислительных технологий

**ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ, ВЫПОЛНЯЕМОЕ В ПЕРИОД**

**ПРОВЕДЕНИЯ   
ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ**

**(НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА)**

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_Баева Диана Николаевна\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(фамилия, имя, отчество полностью)*

Направление подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Место прохождения практики \_\_\_ФГБОУ ВО «Кубанский государственный университет»\_\_\_\_\_

Срок прохождения практики с 01.04.2024 г. по 28.04.2024 г

Цель практики – формирование навыков самостоятельного осуществления научно-исследовательской работы, направленной на решение профессиональных задач; развитие профессиональных знаний в области прикладной математики и информатики, закрепление полученных теоретических знаний по дисциплинам направления и специальным дисциплинам программы бакалавриата Фундаментальная информатика информационные технологии, овладение необходимыми профессиональными компетенциями по избранному направлению подготовки.

Формирование компетенций, регламентируемых ФГОС ВО:

| **Код компе**  **тенции** | **Содержание компетенции (или её части)** | **Планируемые результаты при прохождении практики** |
| --- | --- | --- |
| УК-1 | Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | Приобретение навыков поиска и анализа информации, умения применять системный подход в решении задач |
| УК-2 | Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений | Получение навыков определения задач в рамках поставленной цели, умение анализировать действующие правовые нормы, имеющиеся ресурсы и ограничения для выбора оптимальных способов решения задач |
| УК-3 | Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде | Развитие навыков социального взаимодействия |
| УК-4 | Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(ых) языке(ах) | Развитие навыков деловой коммуникации на государственном языке Российской Федерации и иностранных языках, в устной и письменной формах |
| УК-6 | Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни | Развитие навыков управления временем, умения выстраивать траекторию саморазвития на основе принципов образования в течение всей жизни |
| ОПК-1 | Способен применять фундаментальные знания, полученные в области математических и (или) естественных наук, и использовать их в профессиональной деятельности | Умение применять знания математики и естественных наук на практике, в том числе для анализа и решения задач, возникающих в профессиональной деятельности |
| ОПК-2 | Способен применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности | Умение использовать современное программное обеспечение и компьютерные методы для решения задач профессиональной деятельности |
| ПК-1 | Способен понимать и применять в научно-исследовательской и прикладной деятельности современный математический аппарат, основные законы естествознания, современные языки программирования и программное обеспечение; операционные системы и сетевые технологии | Умение применять знания математики, естествознания, языков программирования и программного обеспечения, операционных систем и сетевых технологий для решения задач профессиональной деятельности |
| ПК-2 | Способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности | Умение проводить локальные исследования под научным руководством в конкретной области профессиональной деятельности с использованием существующих методов |

Перечень вопросов (заданий, поручений) для прохождения практики

1. Изучить понятие больших данных, кластеризации и основные методы кластеризации
2. Исследовать типы нейронных сетей, которые подходят для кластеризации
3. Рассмотреть применение нейронной сети для решения задачи кластеризации

Ознакомлен \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*подпись студента расшифровка подписи (ФИО)*

Руководитель производственной практики

(научно-исследовательская работа) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Руденко О.В.

**План-график выполнения работ**:

|  |  |  |  |
| --- | --- | --- | --- |
| № | Этапы работы (виды деятельности) при прохождении практики | Сроки | Отметка руководителя практики о выполнении (подпись) |
| 1 | Получение задания. Инструктаж по технике безопасности.\* | 01.04.2024 |  |
| 2 | Постановка ТЗ (технического задания). | 02.04.2024 |  |
| 3 | Описание понятия больших данных, кластеризации и основных методов кластеризации | 03.04.2024-10.04.2024 |  |
| 4 | Определение типа нейронных сетей, которые подходят для кластеризации | 11.04.2024-18.04.2024 |  |
| 5 | Определение применения нейронных сетей для решения задачи кластеризации | 19.04.2024-21.04.2024 |  |
| 6 | Оформление результатов проведенного исследования и их согласование с руководителем (составление отчета о прохождении учебной практики). | 22.04.2024 |  |
| 7 | Защита отчета. | 28.04.2024 |  |

Ознакомлен \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*подпись студента расшифровка подписи (ФИО)*

«01» апреля 2024 г.

Руководитель производственной практики

(научно-исследовательская работа) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Руденко О.В.

\*Инструктаж по ознакомлению с требованиями охраны труда, техники безопасности, а также правилами внутреннего трудового распорядка.

**ОЦЕНОЧНЫЙ ЛИСТ**

**результатов прохождения   
ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ**

**(НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА)**

**по направлению подготовки**

**02.03.02 Фундаментальная информатика и информационные технологии**

Фамилия И.О студента Баева Д.Н.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Курс \_\_\_\_\_4\_\_\_\_\_

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | ОБЩАЯ ОЦЕНКА  (отмечается руководителем практики) | Оценка | | | |
| 5 | 4 | 3 | 2 |
|  | Уровень подготовленности студента к прохождению практики |  |  |  |  |
|  | Умение правильно определять и эффективно решать основные задачи |  |  |  |  |
|  | Степень самостоятельности при выполнении задания по практике |  |  |  |  |
|  | Оценка трудовой дисциплины |  |  |  |  |
|  | Соответствие программе практики работ, выполняемых студентом в ходе прохождении практики |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | СФОРМИРОВАННЫЕ В РЕЗУЛЬТАТЕ УЧЕБНОЙ ПРАКТИКИ КОМПЕТЕНЦИИ  (отмечается руководителем практики от университета) | Оценка | | | |
| 5 | 4 | 3 | 2 |
|  | УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач |  |  |  |  |
|  | УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений |  |  |  |  |
|  | УК-3 Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде |  |  |  |  |
|  | УК-4 Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(ых) языке(ах) |  |  |  |  |
|  | УК-6 Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни |  |  |  |  |
|  | ОПК-1 Способен применять фундаментальные знания, полученные в области математических и (или) естественных наук, и использовать их в профессиональной деятельности |  |  |  |  |
|  | ОПК-2 Способен применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности |  |  |  |  |
|  | ПК-1 Способен понимать и применять в научно-исследовательской и прикладной деятельности современный математический аппарат, основные законы естествознания, современные языки программирования и программное обеспечение; операционные системы и сетевые технологии |  |  |  |  |
|  | ПК-2 Способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности |  |  |  |  |

Руководитель производственной практики

(научно-исследовательская работа) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Руденко О.В.

Сведения о прохождении инструктажа по ознакомлению с требованиями охраны труда, технике безопасности, пожарной безопасности, а также правилами внутреннего трудового распорядка

Предприятие Федеральное государственное бюджетное образовательное учреждение высшего образования «Кубанский государственный университет»

Факультет компьютерных технологий и прикладной математики

Кафедра вычислительных технологий

|  |  |
| --- | --- |
| Студент | Баева Диана Николаевна, 21 год |
|  | (ФИО, возраст) |

Дата 01 апреля 2024 г.

**1.     Инструктаж по требованиям охраны труда**

|  |  |
| --- | --- |
| Провел | к. т. н., доцент Руденко О.В. |
|  | (должность, ФИО сотрудника, проводившего инструктаж, подпись) |

|  |  |
| --- | --- |
| Прослушал | Баева Д.Н. |
|  | (ФИО, подпись студента) |

**2.     Инструктаж по технике безопасности**

|  |  |
| --- | --- |
| Провел | к. т. н., доцент Руденко О.В. |
|  | (должность, ФИО сотрудника, проводившего инструктаж, подпись) |

|  |  |
| --- | --- |
| Прослушал | Баева Д.Н. |
|  | (ФИО, подпись студента) |

**3.     Инструктаж по пожарной безопасности**

|  |  |
| --- | --- |
| Провел | к. т. н., доцент Руденко О.В. |
|  | (должность, ФИО сотрудника, проводившего инструктаж, подпись) |

|  |  |
| --- | --- |
| Прослушал | Баева Д.Н. |
|  | (ФИО, подпись студента) |

**4. Инструктаж по правилам внутреннего трудового распорядка**

|  |  |
| --- | --- |
| Провел | к. т. н., доцент Руденко О.В. |
|  | (должность, ФИО сотрудника, проводившего инструктаж, подпись) |

|  |  |
| --- | --- |
| Прослушал | Баева Д.Н. |
|  | (ФИО, подпись студента) |

**Отзыв руководителя**

**О ПРОХОЖДЕНИИ ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ**

**(НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА)**

|  |  |
| --- | --- |
| Студента | Баевой Дианы Николаевны |
|  | *(фамилия, имя, отчество полностью)* |

Направление подготовки 02.03.02. Фундаментальная информатика и информационные технологии

В процессе проведения производственной практики (научно-исследовательской работы) студент провёл анализ выбранной предметной области, собрал необходимые данные для исследования. Также были проведены вычислительные эксперименты, связанные с тестированием разработанного программного продукта. Были проведены все необходимые мероприятия, связанные с выделением целей и задач НИР, а также с подготовкой её аналитической части. В начале исследования была обоснована актуальность выбранной темы. Все пункты индивидуального плана-графика выполнены в полной мере.

В ходе работы поставленное индивидуальное задание по теме “Применение нейронной сети для решения задачи кластеризации” выполнено в полном объеме, студент проявил высокий уровень самостоятельности, и творческий подход к его выполнению. Представлен оформленный текст собранного материала. Программа НИР полностью соответствует отчёту. Работа заслуживает оценки «\_\_\_\_\_\_\_\_\_\_\_\_\_\_».

Руководитель производственной практики

(научно-исследовательская работа) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Руденко О.В.

**СОДЕРЖАНИЕ**

[Введение 9](#_Toc121998006)

[1 Общие сведения о больших данных 12](#_Toc121998007)

[1.1 Понятие и основные характеристики больших данных 12](#_Toc121998008)

[1.2 Понятие задачи кластеризации и общая постановка 13](#_Toc121998009)

[1.3 Методы кластеризации и их классификация 17](#_Toc121998010)

[2 Нейронная сеть для кластеризации](#_Toc121998012) 19

[2.1 Понятие нейронной сети 19](#_Toc121998013)

[2.2 Архитектура искусственных нейронных сетей 21](#_Toc121998014)

[2.3 Классификация нейронных сетей 25](#_Toc121998015) [2.3.1 Нечёткие нейронные сети 29](#_Toc74084975)

2.4 Применение нейронных сетей для решения задачи кластеризации 30

[Заключение 33](#_Toc121998020)

[Список использованных источников 34](#_Toc121998021)

**ВВЕДЕНИЕ**

Современный мир цифровых технологий ежедневно порождает огромное количество данных. С ростом объёмов информации возникли новые проблемы, связанные с её хранением и обработкой. Стремительно увеличивается и число источников данных – социальные сети, камеры видеонаблюдения, видеорегистраторы и т. д. Стоит отметить, что помимо количества поступающей информации существенно изменился и её характер. Основной объем данных представляет собой неструктурированный, непрерывный массив без какой-либо предварительной обработки. Одним из способов решения этой проблемы и выступает кластерный анализ. Он помогает выполнить разделение больших объемов данных на более мелкие группы на основе похожих признаков. Это позволяет упростить их анализ, выявлять структуры и обнаруживать новые закономерности.

Актуальность данной работы обусловлена тем, что нейронные сети решают задачу кластеризации на сложных и многомерных данных, таких как текст, изображения, числовые ряды и другие типы, а глубокие нейронные сети способны выявлять более сложные зависимости. Кроме этого они имеют возможность автоматически извлекать признаки из данных, что значительно упрощает процесс кластеризации и позволяет учитывать разнообразные характеристики объектов.

Основная цель работы – исследование применения нейронной сети для решения задачи кластеризации, т. е. изучение того, как нейронные сети могут облегчить задачу разбиения данных на кластеры и эффективность такого подхода.

Для реализации поставленной цели предполагается решить следующие задачи:

* изучить теорию машинного обучения и анализа данных для исследования задачи кластеризации и способов её решения, а также теорию искусственных нейронных сетей для выбора подходящего вида и архитектуры сети;
* разработать программу, выполняющую разбиение на кластеры входных данных с помощью нейронной сети;
* провести ряд экспериментов для исследования зависимости времени и качества работы алгоритма от размера входных данных и их количества;

Объектом исследования в данной работе является способность нейронных сетей, в частности, отдельного их вида – нечетких нейронных сетей – выполнять кластеризацию данных.

Предметом исследования является алгоритм кластеризации данных с использованием нейронных сетей.

Информационной базой исследования являются результаты работы программы. Методологическая основа исследования включает в себя эмпирический метод (многократный запуск программы на разных конфигурациях нейронной сети и получение результатов её работы), сбор статистической информации (подсчёт среднего времени работы алгоритма и качества разбиения на кластеры), анализ полученных зависимостей, синтез выводов.

Научная новизна работы заключается в том, что исследуется новый подход в кластеризации, так как ранее в большинстве случаев она выполнялась с использованием традиционных методов, таких как иерархическая кластеризация, алгоритм k-средних, блочные модели или самоорганизующиеся карты Кохонена.

Теоретическая и практическая значимость заключается в расширении понимания применения нейронных сетей для задачи кластеризации. Исследование вносит важный вклад в область машинного обучения показывая, как архитектуры нейронных сетей наилучшим образом подходят для разных типов задачи кластеризации, что может привести к новым методам и подходам. Разработанный в работе алгоритм может использоваться для повышения точности и эффективности кластеризации в реальных бизнес-процессах в различных предметных областях.

1. **Общие сведения о больших данных и кластеризации**
   1. **Понятие и основные характеристики больших данных**

Большие данные (Big Data) — группа технологий и методов обработки разноформатных данных большого размера, в распределенных информационных системах, для экономичного извлечения ценности, путем их быстрого захвата, обработки и анализа. При работе с данными такого объема традиционные инструменты не способны осуществить необходимые манипуляции за приемлемое время.

Данный термин используется в сферах, где постоянно происходит увеличение скорости потока данных в организационных процессах: экономике, банковской сфере, маркетинге, веб-аналитике, медицине и т. д.

Необходимо определить какими характеристиками должны обладать данные, чтобы отнести их к категории больших данных. Согласно последним исследованиям, большие данные имеют следующие параметры:

1. Объем (volume). Большими считаются данные, объем которых превышает сто терабайт.
2. Скорость (velocity). Важна как скорость создания данных, так и скорость обработки.
3. Разнообразие источников и форм хранения данных (variety). Данные содержат неструктурированную информацию, необходима возможность одновременной обработки разных типов структур данных.
4. Достоверность (veracity). Любой анализ будет совершенно бесполезен, если данные окажутся недостоверными, поэтому они имеют внутреннюю ценность.

Следует отметить, что большие данные поступают непрекращающимся потоком, вследствие чего очень часто плохо структурированы, имеют большое количество пропусков и нуждаются в предварительной обработке.

Можно выделить следующие основные способы предварительной обработки данных:

1. Очистка данных (Data Cleaning). Обнаружение и исправление ошибок, пропущенных значений, дубликатов и несогласованностей в данных.
2. Уменьшение размерности (Dimensionality Reduction). Снижение размерности данных для сокращения числа переменных, определения структуры взаимосвязей между ними и устранения мультиколлинеарности.
3. Предварительная нормировка и масштабирование данных (Feature Selection and Transformation). Используется при работе с нейронными сетями, где для модели выбираются наиболее информативные признаки, так как ошибки, обусловленные переменными, будут сильнее влиять на обучение сети.
4. Кластеризация (Clustering). Разбиение выборки данных на группы схожих объектов позволяет упростить дальнейшую обработку и принятие решений, применяя к каждому кластеру свой метод анализа.

Наибольший интерес в данной работе представляет именно кластерный анализ. Кластеризация является в том числе элементом именно анализа данных, а не только этапом предварительной обработки.

* 1. **Понятие задачи кластеризации и общая постановка**

Задача кластеризации схожа с задачей классификации и является её логическим продолжением. Сам термин впервые был введен американским психологом Робертом Трионом в 1939 в его работах, посвященных психонометрике и индивидуальным различиям человека.

Кластеризация (кластерный анализ) — это разделение исследуемого множества объектов на группы «похожих» объектов, называемых кластерами. Синонимами термина «кластер» (англ. Cluster — сгусток, пучок, группа) являются термины класс, сгущение. Выявление скрытых закономерностей в заданном наборе данных в большинстве случаев выполнимо только на основе решения задачи кластеризации. Как было сказано выше, задача кластеризации является логическим продолжением задачи классификации, однако существует аспект, принципиально различающий их. Решением задачи классификации является отнесение каждого из объектов к одному из заранее определенных классов. В задаче кластеризации происходит отнесение объекта к одному из заранее неопределенных классов. Разбиение объектов по кластерам осуществляется при их одновременном формировании. Недостатком классификации является то, что любой новый объект, предъявляемый модели, обязательно будет отнесен к одному из классов, даже если в действительности ни к одному из заданных классов не относится. В случае кластеризации «нетипичные» объекты сформируют новый кластер, что позволит обнаружить их.

Характеристиками кластера можно назвать два признака: внутренняя однородность и внешняя изолированность. Отдельно важно отметить роль содержательной интерпретации каждого кластера. Каждому кластеру необходимо присвоить содержательное название, отражающее суть его объектов. Для этого необходимо выявить, признаки, объединяющие объекты. Это может потребовать статистического анализа свойств объекта кластера.

В качестве основных целей кластерного анализа выделяют:

1) Понимание данных при выявлении кластерной структуры. После получения кластеров возможно построение модели для анализа данных каждого кластера.

2) Обнаружение новизны. Выделяются нетипичные объекты, которые не попали ни в один из кластеров.

3) Сжатие данных. Сокращение объема исходных данных путём выделения наиболее типичного представителя каждого кластера.

При решении задачи кластеризации различными методами, полученные кластеры могут быть пересекающимися (overlapping) и непересекающимися или эксклюзивными (non-overlapping, exclusive). Пример кластерного разбиения на пересекающиеся и непересекающиеся кластеры приведен на рисунке 1.

Изображение выглядит как линия, диаграмма, шаблон, дизайн

Автоматически созданное описание

Рисунок 1 – Непересекающиеся и пересекающиеся кластеры.

Следует отметить, что в результате применения различных методов кластеризации полученные кластеры могут быть разной формы, размеров и типа. Некоторые методы особенно чувствительны к шумам или выбросам, а другие – менее.

Кластеризация является описательной процедурой, не делающей никаких статистические выводов, однако дающая возможность провести исследователю развернутый анализ данных и выявить скрытую структуру в них. Вследствие этого нормой является и получение неодинаковых результатов, являющихся особенностью того или иного алгоритма.

Кластерный анализ в общем виде состоит из следующих этапов:

- отбор выборки данных (объектов) для анализа;

- определение множества переменных, по которым будет происходить оценивание объектов в выборке, при необходимости – нормализация значений переменных;

- вычисление значения меры сходства (расстояния) между объектами;

- применение метода кластеризации;

- содержательная интерпретация кластеров, состоящая в изучении свойств объектов, попавших в каждый кластер;

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до тех пор, пока не получен оптимальный результат.

Математическая постановка задачи кластеризации выглядит следующим образом.

Пусть – множество описаний объектов. Задана функция расстояния между объектами – мера близости между объектами, называемая расстоянием. Имеется конечная обучающая выборка объектов Требуется разбить выборку на подмножества, называемые кластерами так, чтобы каждый кластер состоял из объектов, близких по метрике а объекты разных кластеров по выбранной метрике существенно отличались. Это означает, что если расстояние меньше некоторого значения, то объекты считаются близкими и помещаются в один кластер. В противном случае – объекты отличны друг от друга и их помещают в разные кластеры.

Как и в случае с классификацией, множество описаний объектов зачастую называют входными данными. Между процессами классификации и кластеризации существует ещё одно, но очень важное отличие – в задаче классификации используется конечное множество имён классов , которое отсутствует в задаче кластеризации.

Существует множество метрик, но наиболее популярными являются следующие:

* 1. Расстояние Минковского. Вычисляется по формуле .
  2. Евклидово расстояние – частный случай расстояния Минковского: , где и – точки в мерном пространстве.
  3. Манхэттенское расстояние (расстояние городских кварталов) – расстояние, которое можно пройти, двигаясь параллельно осям координат, как на Манхэттене или других городах с прямоугольной продольно-поперечной планировкой улиц: .
  4. Расстояние Чебышева: .
  5. Квадрат евклидового расстояния: .

Данные расстояния применяются, когда объекты можно представить как точки в мерном пространстве. В случаях, когда такое представление не является возможным, применяется метрика коэффициент корреляции Пирсона. Важно отметить, что результат кластеризации существенно зависит от метрики, выбор которой, как правило, субъективен и определяется экспертом.

* 1. **Методы кластеризации и их классификация**

В настоящее время существует большое число алгоритмом кластерного анализа. Общепринятой и единой классификации методов не существует, однако выделяют ряд методов, которые можно классифицировать по различным признакам – иерархические и плоские, четкие и нечеткие алгоритмы.

Иерархические алгоритмы являются наиболее распространенными алгоритмами кластерного анализа, их также называют алгоритмами таксономии. Они строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений. На выходе получается дерево кластеров – дендограмма, корнем которого является вся выборка, а листья – наиболее мелкие кластеры. Данные алгоритмы также делятся по принципу действия на агломеративные (AGNES, Agglomerative Nesting) и итеративные (дивизимные, DIANA, Divisive Analysis). При использовании агломеративных алгоритмов на начальном этапе все объекты определяются в отдельные кластеры, и в процессе работы наиболее схожие по признакам кластеры объединяются в один. Итеративные алгоритмы являются логической противоположностью агломеративным методам. На начальном этапе все объекты объединяются в один кластер и в процессе работы происходит разделение по признаку их отдаленности. На каждом шаге используется матрица расстояний, что при большом количестве объектов, подлежащих кластеризации, приводит к росту вычислительных ресурсов и времени.

Плоские алгоритмы не учитывают иерархическую структуру кластеров. Они стремятся найти оптимальное разбиение объектов на заданное количество кластеров, т. е. обычно требуется указать заранее, какое количество кластеров алгоритм должен найти. Объекты внутри одного кластера должны быть более похожи друг на друга, чем на объекты из других кластеров.

Четкие или непересекающиеся алгоритмы каждому объекту в наборе данных ставят в соответствие номер кластера, т. е. каждый объект принадлежит только одному кластеру. Основные характеристики данной группы методов: внутренняя однородность – объекты внутри одного кластера схожи между собой и внешняя изолированность – кластеры разделены друг от друга.

Нечеткие или пересекающиеся алгоритмы каждому объекту ставят в соответствие набор вещественных значений, которые показывают степень отношения объекта к кластерам, т. е. каждый объект относится к каждому кластеру с некоторой вероятностью. Такие методы позволяют объектам принадлежать нескольким кластерам одновременно с разной степенью принадлежности.

1. **Нейронная сеть для кластеризации**
   1. **Понятие нейронной сети**

Нейронные сети (искусственные нейронные сети, ИНС, НС, нейросети) представляют собой технологию, которая зародилась на стыке таких дисциплин как математика, статистика, физика, нейрофизиология, информатика.

Первое упоминание встречается в работе Уоррена МакКаллока и Уолтера Питца в 1943 г. в статье о логическом исчислении идеи и нервной активности, в которой ученые предложили идею использования нейронных сетей в качестве вычислительных машин. В 1957 г. нейрофизиолог Френк Розенблатт разработал первую нейронную сеть, а в 2010 г. их впервые стали использовать для машинного обучения.

Нейронная сеть – параллельно распределенная структура обработки информации, которая состоит из отдельных элементов (нейронов), соединенных между собой связями. Искусственный нейрон имитирует свойства биологического нейрона. На его вход поступает некоторое множество сигналов, каждый из которых является выходом другого нейрона. Каждый вход умножается на соответствующий вес, аналогичный синаптической силе, и все произведения суммируются, определяя уровень активации нейрона. Таким образом, в состав нейрона включены следующие элементы: набор синапсов, сумматор, а также функция активации. Также еще выделяют пороговый элемент. Его значение сигнализирует об уменьшении или увеличении значения входного сигнала, подаваемого в качестве аргумента в функцию активации. Её также называют передаточной функцией или нелинейным преобразователем. Он реализует нелинейную функцию одного аргумента – выхода сумматора. Можно сказать, что в целом, нейрон реализует скалярную функцию векторного аргумента. На рисунке 2 представлена модель нейрона.

Изображение выглядит как текст, диаграмма, Шрифт, линия

Автоматически созданное описание

Рисунок 2 – Модель нейрона.

Математически функционирование нейрона *k* описывается парой уравнений:

(1)

(2)

Функция активации характеризует зависимость сигнала на выходе нейрона от суммы сигналов на его входах. В общем случае, входной сигнал, весовые коэффициенты и значения смещения могут принимать вещественные значения. Выходной сигнал определяется видом функции активации и может быть как действительным, так и целым. Для некоторых алгоритмов обучения необходимо, чтобы активационная функция была непрерывно дифференцируема на всей числовой оси.

Современные исследования выделяют четыре основных вида функции активации:

1. пороговая (функция Хевисайда);
2. сигмоидальная (логическая), имеющая наибольшее распространение;
3. линейная;
4. гиперболический тангенс.

Очевидно, что существуют и другие функции, однако вышеперечисленные является наиболее частно встречающимися и использующимися.

Важным аспектом в контексте нейронных сетей является понятие обучения сети. В широком смысле слова обучение заключается в нахождении верных коэффициентов связи между нейронами, а также в обобщении данных и выявлении сложных зависимостей между входными и выходными сигналами. В большинстве случаев этот процесс заключается в «прогонке» по итерациям, где на каждой итерации происходит изменения значений синаптических весов и смещений.

Нельзя не отметить, что искусственная нейронная сеть игнорируют многие свойства своего биологического двойника. Например, она не учитывает задержки во времени, которые воздействуют на динамику системы, т. е. входные сигналы сразу же порождают выходной сигнал.

* 1. **Архитектура искусственных нейронных сетей**

Первая разработанная человеком нейронная сеть – персептрон Розенблатта – имеет простейшую структуру. Данная сеть предназначена для классификации линейно-разделимых сигналов, т. е. все исходные объекты входного набора данных разделяются некоторой гиперплоскостью на две области.

За прошедшее время было открыто и изучено множество архитектур нейронных сетей для решения различных задач. Существует большое число видов нейронных сетей, однако выделяют следующие три основополагающих вида сетей.

Первый вид – однослойная сеть прямого распространения (однослойный персептрон). Является одной из самых простых форм нейронных сетей. Имеется один входной слой и один выходной слой, скрытые слои отсутствуют. Сигнал идет строго в одном направлении – от нейронов на входном слое к нейронам на выходном. Обучение может осуществляться с помощью методов оптимизации (например, градиентный спуск), путем настраивания весов между входными и выходными слоями. Также часто используется метод обратного распространения ошибки (Back Propagation): сперва вычисляется градиент функции потерь по параметрам сети, а затем обновляются веса. Однослойные сети в большинстве случаев используются как базовая модель для более сложных архитектур или для решения простых задач кластеризации, регрессии. На рисунке 3 представлена архитектура такой сети.

Изображение выглядит как линия, диаграмма, зарисовка, рисунок

Автоматически созданное описание

Рисунок 3 – Однослойная сеть прямого распространения.

Следующий вид, логично вытекающий из предыдущего – многослойные сети прямого распространения (многослойные персептроны, Multilayer Perceptrons, MLP). Можно сказать, что такие сети представляют собой обобщение однослойного персептрона. В структуре сети, помимо входного и выходного слоев, также присутствуют несколько скрытых слоев вычислительных нейронов (так называемые, скрытые нейроны). Они позволяют обучаться решению сложных задач, последовательно извлекая наиболее важные признаки их входного вектора. Входной сигнал распространяется в прямом направлении слева направо от слоя к слою. В отличие от однослойной сети, использование многослойной сети, позволяет находить и выделять статистические зависимости более высокого порядка именно за счет присутствия скрытых нейронов. Многослойный персептрон имеет два отличительных признака:

1. Каждый нейрон имеет нелинейную функцию активации. Она должна быть гладкой, т. е. всюду дифференцируемой.
2. Сеть обладает высокой степенью связности. Это реализуется через синаптические соединения (веса).

На рисунке 4 представлен многослойный персептрон с двумя скрытыми слоями.

Изображение выглядит как диаграмма, линия, зарисовка, оригами

Автоматически созданное описание

Рисунок 4 – Многослойный персептрон прямого распространения.

Показанная на рисунке сеть является полносвязной (Fully Connected Neural Networks). Это означает, что каждый нейрон любого слоя связан со всеми нейронами предыдущего слоя.

Такую сеть можно интерпретировать как модель вход-выход, в которой веса и пороговые значения являются свободными параметрами модели. Она может моделировать функцию практически любой степени сложности, причем число слоев и нейронов в каждом слое определяют сложность функции.

Основными достоинствами многослойного персептрона является простота в использовании, гарантированное получение ответа после работы сети и широко применяемые алгоритмы обучения. С другой стороны, существует спорные моменты, связанные с проектирование таких сетей. Например, трудно установить, сколько скрытых слоев необходимо для решения данной задачи, сколько следует выбрать нейронов в каждом слое, какой размер обучающей выборки необходим и т. д.

Третий вид – рекуррентные нейронные сети (Recurrent Neural Networks. RNN). В отличие от сетей прямого распространения, они обязательно содержат одну или более обратную связь. Ещё одной отличительной особенностью RNN является наличие внутренней памяти, позволяющей им запоминать предыдущие состояния входных данных и использовать эту информацию для прогнозирования следующих значений. Структура рекуррентных нейронных сетей аналогична структуре многослойного персептрона: входной слой, скрытый и выходной. Внутренняя память располагается именно в скрытом слое.

Рекуррентные сети представляют собой отдельную широкую область в разделе нейронных сетей. Они используются в задачах прогнозировании временных рядов, обработке естественного языка, генерации текста и других задачах, где важным условием является контекст и последовательность данных. Существует три вида рекуррентных нейронных сетей:

* простые рекуррентные нейронные сети (Simple RNNs);
* применение рекуррентные сети с долгой краткосрочной памятью (Long Short-Term Memory, LSTM);
* гейтовые рекуррентные нейронные сети (Gated Recurrent Units,GRU).

На рисунке 5 представлена простая рекуррентная нейронная сеть.

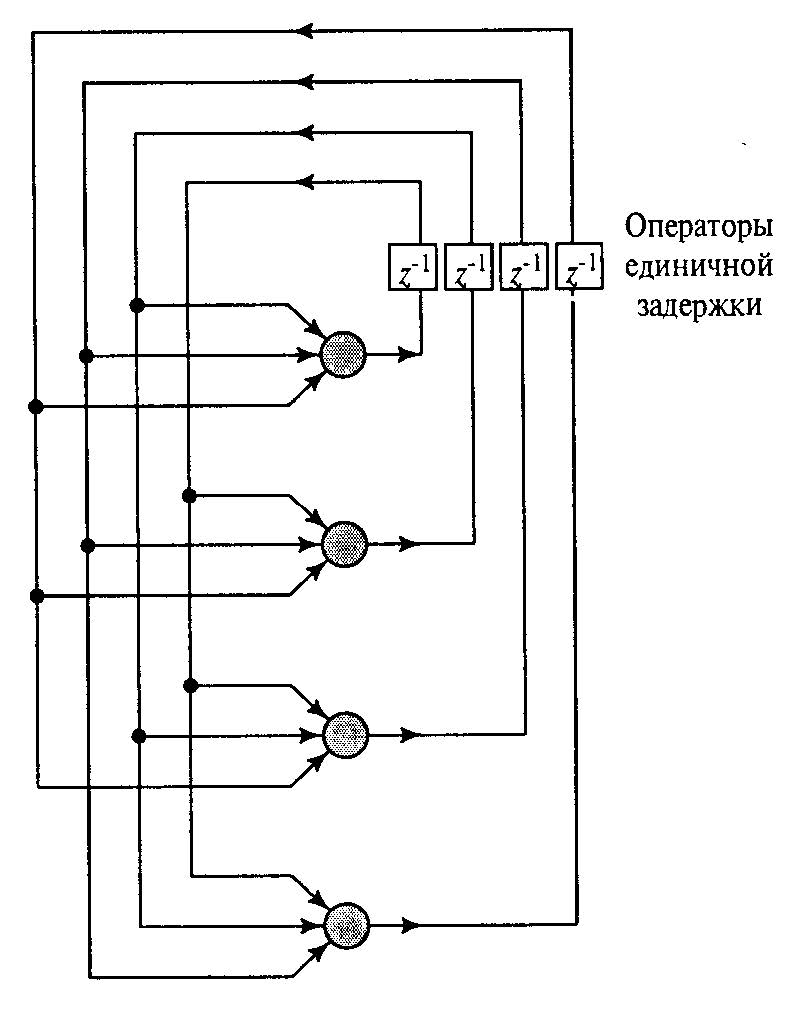


Рисунок 5 – Рекуррентная сеть.

Выбор структуры искусственной нейронной сети это отдельная задача, включающая в себя выбор подходящей для конкретной задачи архитектуры (топологии) сети и функции активации каждого нейрона. После процесса обучения сети, анализа полученных результатов и вычисления значения функции ошибки, выполняется корректировка параметров сети с целью минимизации последней. Если требуемой точности достигнуть не удалось, то архитектура сети изменяется и вышеописанные шаги повторяются вновь.

* 1. **Классификация нейронных сетей**

Классификацию нейронных сетей можно производить в зависимости от различных признаков. На рисунке 6 представлены некоторые признаки.

Изображение выглядит как текст, диаграмма, План, Параллельный

Автоматически созданное описание

Рисунок 6 – Классификация нейронных сетей.

По организации обучения разделяют обучение нейронных сетей с учителем (supervised learning) и без учителя (unsupervised learning).

Первый случай предполагает, что имеются входные сигналы и эталонные реакции на них. Другими словами, существует внешняя среда, которая предоставляет обучающие примеры (значения входов и соответствующим им значения выходов) на этапе обучения или оценивает правильность функционирования сети и в соответствии с определенными критериями изменяет состояние сети. В процессе обучения происходит целенаправленная модификация синаптических весов, связей для достижения максимального соответствия. Сигнал ошибки и обучающий вектор также задействованы в процессе корректировки параметров. Все это делается для того, чтобы сеть находила ответы, наиболее близкие к правильным. Обучение с учителем также называют контролируемым или наблюдаемым обучением.

Обучение без учителя используется, когда не для всех данных обучающей выборки известны правильные ответы. В этом случае, нейронная сеть формирует выходное пространство решений только на основе входных данных. Предпринимаются попытки определить коэффициенты сети для определения классов образцов и дальнейшего их распределения по выделенным классам. В частности, используется для решения задачи кластеризации. Иногда данный способ обучения называют обучением на основе самоорганизации, самоорганизующимся или неконтролируемым (ненаблюдаемым).

В некоторых источниках выделяют также смешанное (гибридное) обучение. В нем часть весов определяется посредством обучения с учителем, а другая часть получается с помощью алгоритмов самообучения.

По характеру настройки весов различают сети с фиксированными связами и с динамическими связями. В первом случае, весовые коэффициенты сети выбираются сразу, исходя из условий задачи. При этом должно выполняться следующее равенство:

(3)

где

– весовые коэффициенты сети,

– время работы сети.

Для сетей с динамическими связями настройка синаптических связей происходит в процессе обучения, т. е.

(4)

По способу обучения разделяют обучение по входам и выходам. При обучении по входам обучающий пример представляет собой только вектор входных сигналов, а при обучении по выходам в него входит и вектор выходных сигналов, который соответствует входному вектору. При данном способе обучения подразумевается использование пар входных и соответствующих выходных данных для предсказания целевых значений. В то время как похожее обучении с учителем включает в себя использование целевых значений в процессе обучения, а обучение без учителя основано на обнаружении закономерностей в данных без явных целевых значений.

По структуре нейронные сети можно разделить на полносвязные, многослойные (слоистые) и слабосвязные (с локальными связями). На рисунках 7–9 представлены такие сети.

Изображение выглядит как зарисовка, диаграмма, рисунок, оригами

Автоматически созданное описание

Рисунок 7 – Полносвязная нейронная сеть.

Изображение выглядит как линия, диаграмма, круг, зарисовка

Автоматически созданное описание

Рисунок 8 – Многослойная сеть с последовательными связями.

Изображение выглядит как линия, диаграмма, оригами

Автоматически созданное описание

Рисунок 9 – Слабосвязные сети.

В свою очередь многослойные нейронные сети (в некоторых источниках упоминаются как неполносвязные – описываемые неполносвязынми ориентированными графами) также имеют свое собственно деление. Они подразделяются на однослойные и многослойные персептроны, с прямыми, перекрестными и обратными связами. Наиболее часто встречаются сети прямого распространения. Кроме того, персептроны различают на персептроны с регулярными и случайными связями.

* + 1. **Нечёткие нейронные сети**

Отдельное место в общей классификации нейронных сетей занимают такие сети, которые располагаются на стыке ИНС и нечёткой логики. Они решают особую задачу, зависящую от специфики применения и методов обучения.

Нечёткая нейронная сеть (ННС, Fuzzy Neural Networks, FNN) – многослойная нейронная сеть, в которой слои выполняют функции элементов системы нечёткого вывода. Характерной особенностью таких сетей является возможность использовать для расчета выходного сигнала нечёткие правил вывода. Нейроны, их структура и параметры, аналогичны обычным ИНС.

Другое название ННС — гибридные сети. Они были разработаны для объединения в себе достоинств нейронных сетей и систем нечёткого вывода. Так, они позволяют создавать и представлять модели систем в форме правил нечётких продукций (ПНП), обладающих свойством наглядности и простотой содержательной интерпретации. В то же время, для построения ПНП используются методы нейронных сетей, что более удобно и менее вычислительно затратно для разработки. Можно отметить, что такие сети схожи «человеческому» способу мышления нечёткими правилами принятия решений. Нечёткие нейронные сети – это комбинация лингвистического стиля рассуждений нечётких систем со структурой и обучением нейронной сети. Они позволяют учесть неопределенность и нечёткость в данных, часто встречающихся в реальных практических задачах.

В зависимости от специфики решаемой задачи определяется архитектура и вид нечёткой сети. Для отдельных типов задач разработаны оптимальные конфигурации сетей, которые позволяют быстро найти требуемой решение. Нечёткие нейронные сети по структуре схожи с многослойными сетями, однако у первых скрытые слои соответствуют этапам работы нечёткой системы (введение нечёткости, определение множества нечётких правил, приведение к четкости). Значения входных сигналов, выходных и весов представляют собой вещественные числа на отрезке [0, 1].

Обучение гибридных сетей, так же как и классических сетей, проводится либо по алгоритму с учителем, который основан на минимизации целевой функции, либо по алгоритму самоорганизации.

Нечёткие нейронные сети не являются самым распространённым и широко использующимся типом нейронных сетей, но они остаются незаменимым инструментом в области моделирования неопределенности и обработки нечётких данных.

* 1. **Применение нейронных сетей для решения задачи кластеризации**

Как было сказано выше, искусственные нейронные сети это мощный инструмент для обработки данных. Задача кластеризации также позволяет облегчить такую обработку, поэтому особый интерес представляет совместное использование названных инструментов.

Применение нейронных сетей в задаче кластеризации имеет несколько преимуществ. В первую очередь, нейронные сети способны обрабатывать большие объемы данных и находить сложные зависимости между объектами. Благодаря этому возможно получить более точные и качественные результаты кластерного анализа.

Помимо этого, нейронные сети могут автоматически извлекать признаки из данных, позволяющее обнаруживать скрытые закономерности и структуры в данных. Это облегчает обработку данных со сложной структурой.

В настоящее время одним из наиболее популярных методов кластеризации, который позволяет не только формировать многомерную кластерную структуру, но и эффективно визуализировать ее, являются самоорганизующиеся карты Кохонена (SOM – self-organizing map), или карты Кохонена. В основе функционирования SOM лежит технология обучения без учителя для специального вида нейронных сетей, называемых сетями Кохонена, с последующей визуализацией кластеров с помощью проекции многомерных данных на плоскость с сохранением топологического подобия (проекции Саммона). Сеть Кохонена — это однослойная сеть, построенная из нейронов типа WTA (Winner Takes All — победитель получает все). Она относятся к самоорганизующимся нейронным сетям. С их помощью производится кластеризация объектов, описываемых количественными характеристиками.

Также широко используется алгоритм адаптивной резонансной теории (ART). Его суть заключается в идеи создании адаптивной системы, способной гибко адаптироваться к изменениям в данных и обнаруживать новые кластеры. В ART используется два вида нейронов: нейроны-кандидаты, которые подразумевают под собой потенциальные кластеры и нейроны-победители, которые представляют собой текущие кластеры. В процессе обучения сети первый тип нейронов соревнуется со вторым, и если нейрон-кандидат побеждает, то он становится нейроном-победителем, представляющим собой новый кластер.

Часто встречается в различных источниках и алгоритм гибридной нейронной сети (Hybrid Neural Network, HNN). Это комбинация применения нейронных сетей и методов кластеризации. Нейронная сеть выполняет предварительную обработку данных, извлечение данных, а далее работает какой-либо метод кластерного анализа, выполняющий непосредственное разбиение данных на кластеры. Такой подход позволяет использовать преимущества нейронных сетей в извлечении сложных признаков и применять более эффективные методы кластеризации для разделения данных.

Конечно же, вышеперечисленные методы являются лишь часть подобных алгоритмов. Выбор того или иного алгоритма зависит от требований данных, их размера и характеристик.

Одним из недостатков использования нейронных сетей в кластеризации является высокая вычислительная сложность, требующая больших ресурсов и времени, особенно при обработке больших объемов данных. Помимо этого, результаты работы сети могут быть сложными для интерпретации из-за их сложной структуры, что вызывает затруднение в понимании полученных кластеров. Результаты кластеризации могут сильно зависеть от исходных данных и выбранных параметров модели, вследствие чего возникает необходимость в тщательной настройке и проверке.

**ЗАКЛЮЧЕНИЕ**

В заключение данной работы мы рассмотрели понятие больших данных и их характеристик, а также ввели основные понятия, связанные с кластерным анализом. Была рассмотрена общая классификация существующих методов кластеризации.

Особое внимание было уделено изучению нейронных сетей, их архитектуре и классификации. Помимо этого, был рассмотрен особый вид – нечёткие нейронные сети. Было выполнено исследование какие типы нейронных сетей можно использовать для кластеризации.

В ходе исследования было рассмотрено применение нейронных сетей для решения задачи кластеризации. В частности, какие алгоритмы кластеризации используют нейронные сети, а после были выделены преимущества и ограничения такого подхода.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

* 1. **Мандель, И. Д.** Кластерный анализ / И. Д. Мандель. — Москва : Финансы и статистика, 1988. — 176 с. — ISBN 5-279-00050-7. — Текст: непосредственный.
  2. **Калан, Р.** Основные концепции нейронных сетей / Р. Калан. — Москва : Издательский дом «Вильямс», 2001. — 290 с. — ISBN 5-8459-0210-X. — Текст: непосредственный.
  3. What Is Data Preprocessing & What Are The Steps Involved? – Текст: электронный // Tobias Geisler Mesevage — 2021. — URL: https://monkeylearn.com/blog/data-preprocessing/ (дата обращения 02.04.2024).
  4. Нейронные сети в кластеризации данных: алгоритмы и примеры применения // Давид Б. — 2021. — URL: https://nauchniestati.ru/spravka/nejronnye-seti-i-reshenie-klasterizaczii-dannyh// (дата обращения 20.04.2024).
  5. **Хайкин, С.** Нейронные сети: полный курс, 2-е издание / С. Хайкин; [пер. с англ. Н. Куссуль, А. Шелестовой]. – Москва : Издательский дом «Вильямс», 2016. – 1104 с. — ISBN 5-8459-0890-6. — Текст: электронный. — URL: https://studizba.com/files/show/djvu/1762-1-haykin-s--neyronnye-seti.html (дата обращения: 12.04.2024).