|  |  |  |
| --- | --- | --- |
|  | Изображение выглядит как зарисовка, рисунок, символ, эмблема  Автоматически созданное описание |  |
| МИНОБРНАУКИ РОССИИ | | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | |
| **Институт комплексной безопасности и специального приборостроения**  Кафедра КБ-14 «Технологии искусственного интеллекта и анализ данных»  Дисциплина «Введение в машинное обучение» | | |

**Отчёт**

«Практическая работа №5»

Выполнил:

Студент 3 курса

Группа БСБО-06-22

Шифр 22Б1240

Белявцева Екатерина Александровна

Москва, 2024

# Задание

### Часть 1

Создайте папку Очистка данных для сохранения всех файлов практической работы. Запустите Deductor, сохраните новый проект в этой папке в файле. Регулярно сохраняйте результаты работы.

## Задание 1. Восстановление пропущенных данных.

Порядок выполнения работы:

1. Подготовьте в Excel таблицу значений функции sin(x), где аргумент x изменяется от 0 до 10 с шагом 0,05.
2. Удалите несколько значений функции (не аргумента), так чтобы были пропущенные данные.
3. Сохраните полученную таблицу в текстовом формате в Вашей папке.
4. Загрузите эти данные в Deductor. Назовите ветвь Восстановление значений sin(х). Задайте отображение представлений Таблица и Диаграмма.
5. В открывшемся окне выберите пункт Парциальная обработка (в разделе Очистка данных). На втором ее шаге выберите переключатель Аппроксимация. На последующих шагах не задавайте никакой обработки данных. Отобразите представления Таблица и Диаграмма.
6. Сравните обработанные данные с необработанными.

## Задание 2. Редактирование аномалий.

Порядок выполнения работы:

1. Скопируйте в свою папку демонстрационный файл Anketa.txt. Загрузите его данные в тот же проект с помощью Мастера импорта. Назовите ветвь Редактирование аномалий Anketa. Отобразите таблицу и точечную диаграмму для поля Сумма кредита.
2. Для созданной ветви проведите Парциальную обработку. При этом:

* в ее окне Восстановление пропущенных данных установите переключатель Отключить;
* в окне Редактирование аномальных значений выберите поле Количество, установите флажок Редактирование аномальных значений, в списке Степень подавления выберите пункт Малая.
* Сглаживание не задавайте. Выполните редактирование аномалий, нажав кнопки Пуск и Далее. Отобразите снова таблицу и точечную диаграмму для поля Сумма кредита.

1. Сравните обработанные и необработанные данные. Выведите в правой нижней части окна панель детализации. Выясните с ее помощью, на какие значения были заменены аномальные.
2. Для той же ветви Редактирование аномалий Anketa проведите еще одно редактирование аномалий, выбрав большую степень их подавления. Дайте узлам сценария понятные названия. Сравните новые величины для аномальных значений. Сделайте выводы.

## Задание 3. Сглаживание и очистка от шумов.

1. Скопируйте в свою папку демонстрационный файл dynamics\_website.txt. В том же проекте создайте еще одну ветвь сценария Сглаживание шумов website, импортировав данные этого файла. Отобразите при этом диаграмму (линии).
2. Проведите для этого узла трижды сглаживание и очистку от шумов всеми тремя способами (по очереди), которые предлагает обработчик Парциальная обработка (шаг Спектральная обработка). Для каждого из преобразований отобразите диаграмму (линии). Укажите настройки преобразований:

* для опции Сглаживание оставьте величину полосы пропускания равной 50;
* для Вычитания шума установите среднюю степень вычитания;
* для Вейвлет-преобразования оставьте значения глубины 3 и порядка 6.
* Дайте полученным трем узлам сценария осмысленные названия.
* Сравните с помощью диаграмм исходные данные и полученные после каждого из преобразований. Сделайте выводы.

## Задание для самостоятельной работы.

1. Скопируйте в свою папку демонстрационный файл Trade.txt. Загрузите в тот же проект данные файла. Назовите ветвь Очистка данных trade. Отобразите диаграмму (линии).
2. Проведите парциальную обработку данных, применив к узлу сразу в один прием все три типа обработки: аппроксимацию данных, подавление аномалий большой степени и сглаживание данных с полосой пропуска 50. Отобразите диаграмму (линии). Сравните исходные и преобразованные данные.
3. Измените полученный узел. Замените сглаживание данных на вейвлет-преобразование с глубиной разложения 2 и порядком вейвлета 6. Сравните результаты.

Сохраните файл проекта.

### Часть 2

Требуется распределить регионы на кластеры по демографическим данным.

Порядок выполнения работы:

1. Создайте папку Кластеризация. Скопируйте в нее и импортируйте в Deductor данные из демонстрационного файла Population.txt. Сохраните проект в файле.

2. Проведите редактирование аномалий и удаление шумов для полей Удельный вес городского населения, Изменение численности населения, Среднегодовая численность населения занятых в экономике, Среднедушевой денежный доход.

3. Выберите Мастер обработки Кластеризация.

4. На втором шаге Мастера укажите поля Численность населения и Регион информационными столбцами, а 4 обработанных (сглаженных) поля – входными.

5. На следующем шаге определите все множество данных как обучающее.

6. Далее задайте фиксированное количество кластеров, равное пяти.

7. Выберите из списка визуализаторов способы отображения данных Профили кластеров и Куб.

8. Для настройки визуализатора Куб выберите рассматриваемые свойства как факты, а номер кластера и регион – как измерения. Затем укажите отображение среднего значения по всем фактам рассматриваемой группы.

9. Проанализируйте полученные результаты:

* Рассмотрите таблицу Профили кластеров. Задайте сортировку кластеров по убыванию поддержки (количества записей, попавших в кластер). Поля упорядочьте по убыванию значимости (степени зависимости результата от данного поля) для столбца Итого (т.е. для всех данных).

Для каждого поля отобразите только три параметра – Значимость, Доверительный интервал и Среднее значение поля.

Определите кластеры, где самым значимым параметром является среднедушевой доход (определите его среднюю величину). Проанализируйте другие значимые параметры для этих кластеров. Выясните, какие регионы попали в эти кластеры, отобразив данные по кластеру.

Найдите кластеры регионов с положительным и отрицательным

значимым приростом населения. Выясните другие значимые их

характеристики.

* Добавьте визуализатор Многомерная диаграмма: задайте для него отображение регионов, номеров кластеров и (по очереди) какого-либо из исследуемых параметров – например, среднедушевого дохода и т.д. Отобразите легенду.
* Возможно, более удобным для анализа может оказаться визуализатор Диаграмма размещения – он позволяет вывести больше параметров за счет вариаций цвета, размера и формы точек диаграммы.
* Используйте для анализа данных вкладку Куб, выведя на ней кросс-диаграмму.

10. Проведите для тех же исходных данных кластеризацию на 4 кластера. Отобразите необходимые визуализаторы и проведите аналогичный анализ получившихся кластеров.

11. Выберите тот вариант кластеризации, который допускает более удачную интерпретацию. Переименуйте для него каждый кластер в зависимости от его особенностей.

12. Скопируйте узел удачной (на Ваш взгляд) кластеризации и переместите его непосредственно к импортированным данным (без их сглаживания). Проведите эту кластеризацию на несглаженных данных и сравните результаты группировки в обоих случаях.

Сохраните файл проекта.

### Ча**сть** 3

Имеется база данных коммерческих банков с показателями деятельности за текущий период. Необходимо провести их кластеризацию, т.е. выделить однородные группы банков.

Порядок выполнения работы:

1. Создайте папку Карты Кохонена. Скопируйте в нее и импортируйте в Deductor данные из файла Banks.txt. Сохраните проект в файле.

2. Запустите Мастер обработки и выберите обработчик Карта Кохонена.

* На втором шаге Мастера укажите выходным столбцом поле Прибыль, а поля Филиалы, Сумма активов, Собственные активы, Банковские активы, Средства в банке – входными.
* Оставьте значения по умолчанию для обучающего и тестового множеств, настроек параметров карты Кохонена, за исключением числа кластеров: задайте фиксированное число кластеров 4.

3. После завершения процесса обучения в списке визуализаторов выберите Карту Кохонена и Что–если.

* В Мастере настройки карты Кохонена укажите поля для отображения: все входные и выходные столбцы и кластеры. Способ раскраски – цветная палитра. Задайте отображение границ кластеров.

4. На полученной карте Кластеры все данные будут отображены в виде четырех областей разного цвета.

* Выберите кластер меньшего размера. Проанализируйте аналогичные по расположению области других карт: выясните размеры различных активов для банков этого кластера. (Цветовая шкала числовых величин расположена под каждой картой: синие цвета соответствуют наименьшим значениям, красные оттенки – наибольшим.)
* Откройте панель данных внизу окна и включите фильтрацию по кластеру. Выясните, какие банки попали в рассматриваемый кластер.
* Затем выберите самый большой кластер и проанализируйте финансовые показатели его банков на остальных картах. Аналогично – для остальных кластеров.

5. Изучите распределение кластеров на одной из карт с финансовыми показателями – например, Сумма активов. Добавьте визуализатор Многомерная диаграмма для этого же показателя (оси: номер кластера, банк, сумма активов). Сравните оба вида визуализации.

6. С помощью карты Филиалы проверьте, что число филиалов не сказалось на результатах кластеризации.

* Добавьте визуализацию Профилей кластеров. Оставьте в этой таблице параметры Значимость, Доверительный интервал и Среднее. Проверьте значимость поля Филиалы для всех кластеров – соответствует ли она карте Кохонена?

7. С помощью карты Прибыль проверьте, имеется ли связь прибыльности банков и их группировки.

* Создайте еще один узел сценария – аналогичную Карту Кохонена для тех же данных, указав вместо входного параметра Филиалы входной параметр Прибыль. Дайте различные названия обоим узлам сценария.
* Визуализируйте таблицу Профили кластеров. Проверьте значимость поля Прибыль для разных кластеров.
* Сравните оба результата кластеризации на картах Кохонена.

Сохраните файл проекта.

# Ход работы

### Часть 1

Была создана папка «Очистка данных», туда же был сохранён проект из Deuctor.

Изображение выглядит как текст, снимок экрана, Мультимедийное программное обеспечение, линия

Автоматически созданное описание

Рисунок 1. Создание папки «Очистка данных» и сохранение туда проекта

## Задание 1

1-2. Была подготовлена таблица в Excel со значениями функции sin(x), где аргумент x изменяется от 0 до 10 с шагом 0,05. Также были удалены значения в ячейках 7, 16, 24, 30 и 35.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 2. Итоговый вариант таблицы sin(x)

3. Далее полученная таблица была сохранена в текстовый формат в папку «Очистка данных».

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рисунок 3. Сохранение таблицы в текстовом формате

Изображение выглядит как линия, текст, График, диаграмма

Автоматически созданное описание4. После этого файл sin(x) был открыт в Deductor Studio. Ветвь была названа «Восстановление значений sin(х)». Также были заданы отображения представлений в виде таблицы и диаграммы.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Автоматически созданное описание

Рисунок 4. Загрузка данных в Deductor

5. В новой версии Deductor вместо одного обработчика «Парциальная обработка» появилось несколько модулей: «Качество данных», «Заполнение пропусков», «Редактирование выбросов» и «Спектральная обработка»

В этом пункте будет использоваться обработчик «Заполнение пропусков». Вместо аппроксимации будет использоваться «Среднее значение». Представление будет в виде таблицы и диаграммы.

Изображение выглядит как текст, График, линия, Параллельный

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Рисунок 5. Данные после обработчика «Заполнение пропусков» со средним значением

6. При сравнении данных до и после обработки можно увидеть, что пропущенные значения заменились средними значениями по всей таблице.

## Задание 2

1. В папку «Очистка данных» скопирован демонстрационный файл Anketa.txt, а после загружен в проект с помощью Мастера импорта. Ветвь названа «Редактирование аномалий Anketa». Также отображение представлено в виде таблицы и точечной диаграммы для поля «Сумма кредита».

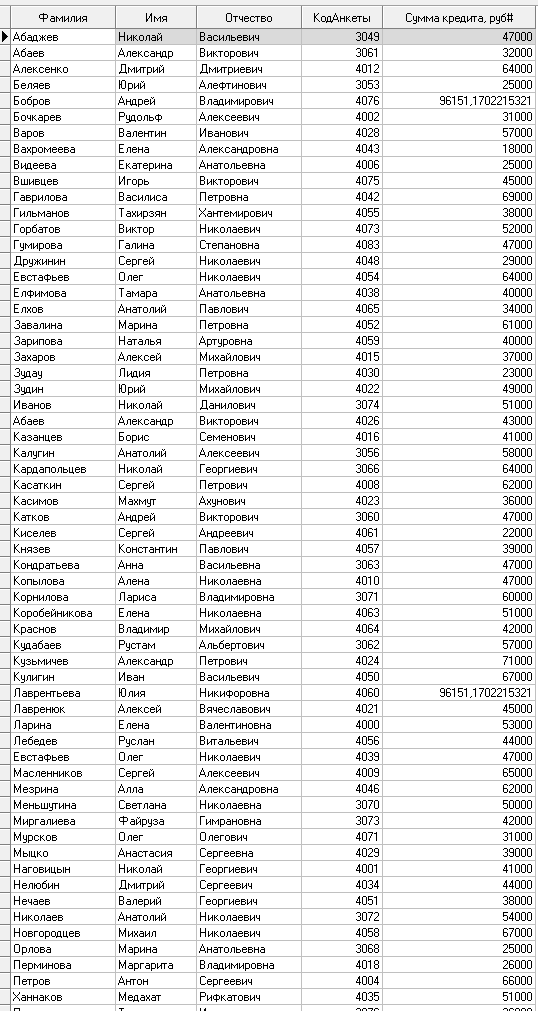
Изображение выглядит как текст, снимок экрана, число, Параллельный

Автоматически созданное описание Изображение выглядит как текст, линия, снимок экрана, Параллельный

Автоматически созданное описание

Рисунок 6. Данные файла Anketa.txt

2. Так как «Парциальная обработка» больше не существует, будет проведено «Редактирование выбросов» при этом будет использоваться метод обработки экстремальных значений в виде «Заменять наиболее вероятным», а метод обработки выбросов – «Ограничивать». Все поля, кроме суммы кредитов, остаются неизменными.

 Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Рисунок 7. Данные файла Anketa.txt после обрботки «Редактирование выбросов»

3. При редактировании выбросов аномальные значения заменились на более вероятные. В частности, все значения, которые были больше 100.000 заменились на 96151,1702215321.

4. Далее было проведено ещё одно редактирование аномальных значений, теперь была задана меньшая мера отклонения от среднего значения. После проведённого редактирования данные представлены в узле под названием «Редактирование выбросов и экстремальных значений (большая степень подавления)». В результате максимальное значение данных стало 71000, а остальные экстремальные значения данных заменились на 40062,5.

Изображение выглядит как текст, меню, документ, число

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описание

Рисунок 8. Данные после большей степени подавления

## Задание 3

1. Демонстрационный файл dynamics\_website.txt был скопирован в папку «Очистка данных».

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рисунок 9. Папка «Очистка данных» на данном этапе

После этого в Deductor Studio в том же проекте была создана ветвь сценария «Сглаживание шумов website». Туда же был импортирован файл dynamics\_website.txt. Данные при этом были отображены в виде диаграммы (линии).

Изображение выглядит как линия, График, диаграмма, текст

Автоматически созданное описание

Рисунок 10. Данные dynamics\_website.txt до спектральной обработки

2. Далее было проведено 3 разных спектральных обработки:

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 11. Данные dynamics\_website.txt после сглаживания в виде Фурье-преобразования с полосой пропускания равной 50. Название узла – Спектральная обработка (сглаживание)

Изображение выглядит как линия, График, текст, Параллельный

Автоматически созданное описание

Рисунок 12. Данные dynamics\_website.txt после сглаживания в виде Вейвлейт-преобразования со значением глубины 3 и порядка 6. Название узла – Спектральная обработка (Вейвлейт-преобазование)

Изображение выглядит как линия, График, текст, диаграмма

Автоматически созданное описание

Рисунок 13. Данные dynamics\_website.txt после вычитания шума в виде Фурье-преобразования со средней степенью вычитания. Название узла – Спектральная обработка (вычитание шума)

По рисункам 11-13 можно сделать следующие выводы:

1. При сглаживании в виде Фурье-преобразования с полосой пропускания равной 50 убираются резкие колебания данных, которые имеют высокую частоту. Основные тренды и медленные изменения остаются. Сами данные становятся более гладкими и менее детализированными.
2. При сглаживании в виде Вейвлейт-преобразования со значением глубины 3 и порядка 6 сохраняются крупные структуры и тренды. Детали в данных с высокой временной или пространственной частотой (мелкие изменения) сглаживаются.
3. При вычитании шума в виде Фурье-преобразования со средней степенью вычитания высокочастотные шумы в данных значительно ослабляются. Основная структура данных остается почти неизменной.

## Задание для самостоятельной работы

1. Скопируйте в свою папку демонстрационный файл Trade.txt. Загрузите в тот же проект данные файла. Назовите ветвь Очистка данных trade. Отобразите диаграмму (линии).
2. Проведите парциальную обработку данных, применив к узлу сразу в один прием все три типа обработки: аппроксимацию данных, подавление аномалий большой степени и сглаживание данных с полосой пропуска 50. Отобразите диаграмму (линии). Сравните исходные и преобразованные данные.
3. Измените полученный узел. Замените сглаживание данных на вейвлет-преобразование с. Сравните результаты.

Сохраните файл проекта.

1. Демонстрационный файл Trade.txt был скопирован в папку «Очистка данных».

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рисунок 14. Папка «Очистка данных» на данном этапе

Так как в файле Trade.txt присутствует столбец с датой в нестандартном формате, был написан код на Python, чтобы преобразовать её. Дата изначально была в формате год-номер\_месяца, после преобразования она стала иметь вид год-месяц-01. Число месяца нам не важно, поэтому было взято 1 число каждого месяца. Также данные в столбце «Количество» преобразуем его в тип данных int, иначе в Deductor Studio не получится его импортировать.

Листинг 1. Код для преобразования даты

|  |
| --- |
| import pandas as pd  file\_path = 'Trade.txt' df = pd.read\_csv(file\_path, delimiter='\t', encoding='cp1251')  def convert\_date(row):  year, month = row.split('-')  month\_number = int(month[1:])  return f'01.{month\_number:02d}.{year}'  df['Дата'] = df['Дата'].apply(convert\_date) df['Количество'] = df['Количество'].astype(int)  output\_file = 'Trade.txt' df.to\_csv(output\_file, sep='\t', index=False, encoding='cp1251') |

После этого в Deductor Studio в том же проекте была создана ветвь сценария «Очистка данных trade». Туда же был импортирован файл Trade.txt. Данные при этом были отображены в виде диаграммы (линии).

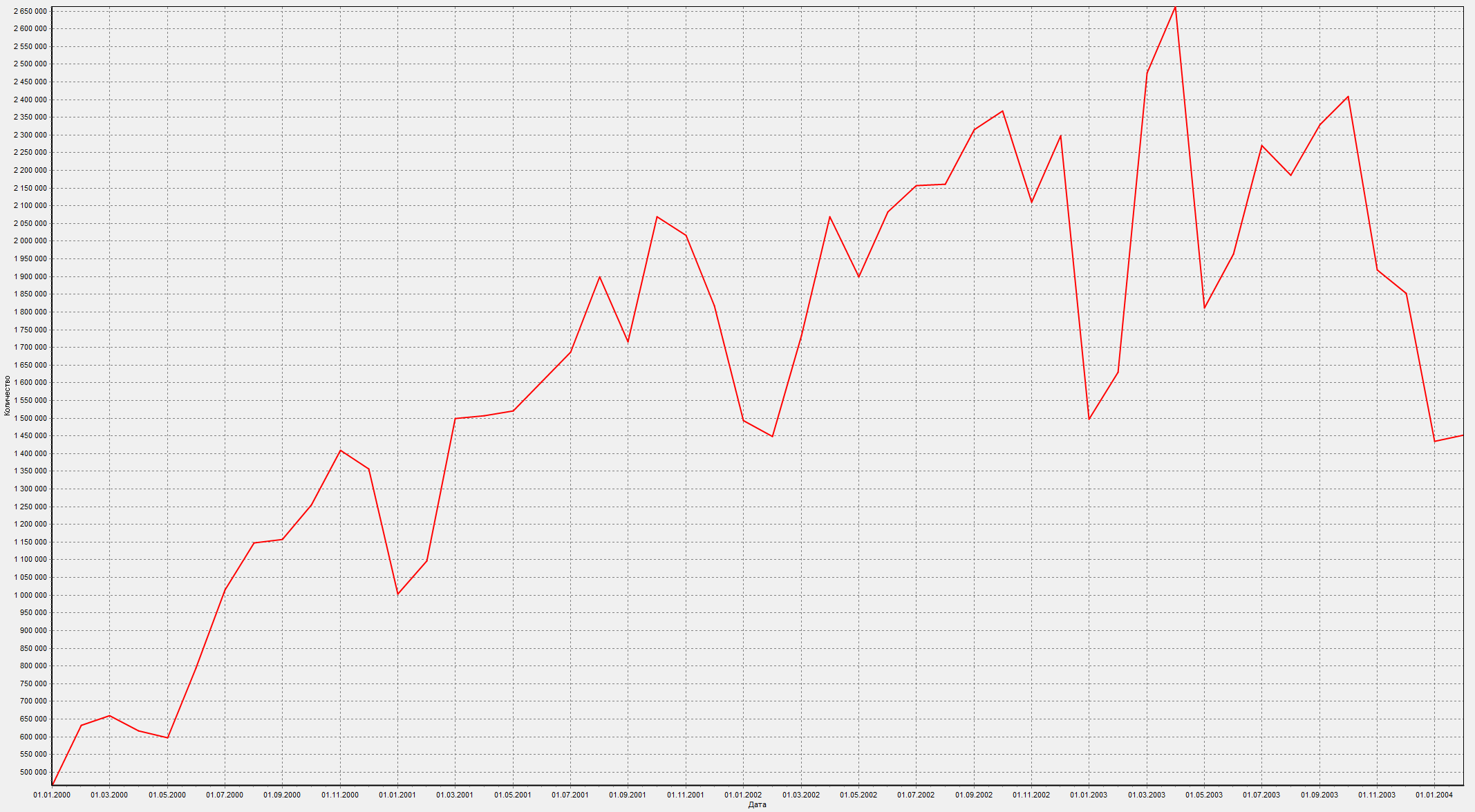


Рисунок 15. Данные файла Trade.txt

2. Далее для этих данных был проведён ряд последовательных процедур. В начале редактирование выбросов и экстремальных значений большой степени.

Изображение выглядит как текст, линия, График, Параллельный

Автоматически созданное описание

Рисунок 16. Данные файла Trade.txt после редактирования выбросов большой степени

Далее была проведена спектральная обработка, точнее сглаживание данных с полосой пропуска 50.

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 17. Данные файла Trade.txt после спектральной обработки

Данные стали более стабильными, ушли аномалии и сгладились углы.

3. Теперь был изменён полученный узел, сглаживание данных было заменено на Вейвлейт-преобазование с глубиной разложения 2 и порядком Вейвлета 6.

Изображение выглядит как линия, текст, График, диаграмма

Автоматически созданное описание

Рисунок 18. Данные файла Trade.txt после изменения вида спектральной обработки

При изменении вида спектральной обработки на Вейвлейт-преобазование с глубиной разложения 2 и порядком Вейвлета 6 данные стали более сглаженными, при этом не нарушилась их структура.

### Часть 2

1. Для выполнения этого задания была создана папка «Кластеризация», куда был скопирован файл population.txt. Также был создан проект в Deductor Studio и также размещён в этой папке.

Изображение выглядит как текст, снимок экрана, Мультимедийное программное обеспечение, программное обеспечение

Автоматически созданное описание

Рисунок 19. Папка «Кластеризация» на данном этапе

После, файл population.txt был импортирован в проект в Deductor. Данные были отображены в виде диаграммы (линий) для полей «Удельный вес городского населения», «Изменение численности населения», «Среднегодовая численность населения занятых в экономике», «Среднедушевой денежный доход».

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Рисунок 20. Диаграмма данных из файла population.txt до изменений

2. Далее было проведено редактирование аномалий и удаление шумов для полей «Удельный вес городского населения», «Изменение численности населения», «Среднегодовая численность населения занятых в экономике», «Среднедушевой денежный доход».

Изображение выглядит как текст, График, линия, диаграмма

Автоматически созданное описание

Рисунок 21. Диаграмма данных из файла population.txt после редактирования аномалий

Для удаления шумов было использование Вейвлейт-преобазование с глубиной разложения 2 и порядком Вейвлета 6.

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 22. Данные после Вейвлейт-преобазования

3. После проведённых операций по предподготовке данных в Мастере обработки был выбран метод EM кластеризации.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 23. Выбор кластеризации в Мастере обработки

4. На втором шаге все столбцы кроме 4 обработанных были указаны как информационные, а оставшиеся 4 остались входными.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 24. Распределение назначения столбцов для кластеризации

5. Далее на следующем шаге всё множество данных было определено как обучающее.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 25. Определение обучающей части данных

6. На следующем шаге было задано фиксированное количество кластеров, равное 5.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 26. Определение фиксированного числа кластеров

7. Далее были выбраны способы отображения данных в виде «Профиля кластеров» и «Куба».

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 27. Выбор режима отображения

8. Следующее окно – настройка визуализатора «Куб».

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 28. Настройка визуализации «Куб»

9. Далее будет проведён анализ полученных результатов.

Изображение выглядит как текст, снимок экрана, число, диаграмма

Автоматически созданное описание

Рисунок 29. Профили кластеров

Вначале нужно задать сортировку кластеров по убыванию поддержки (количества записей, попавших в кластер). Поля упорядочить по убыванию значимости (степени зависимости результата от данного поля) для столбца Итого (т.е. для всех данных).

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Рисунок 30. Параметры сортировки

Далее надо для каждого поля отобразите только три параметра – Значимость, Доверительный интервал и Среднее значение поля.

Изображение выглядит как текст, снимок экрана, Шрифт, дисплей

Автоматически созданное описание

Рисунок 31. Настройка отображения параметров

После этого надо определить кластеры, где самым значимым параметром является среднедушевой доход. При анализе кластеров (см. рис. 32) можно выделить, что 0 кластер с самым значимым параметром среднедушевого дохода. Также в этом кластере находятся самые значимые параметры удельного веса городского населения, среднегодовая численность населения занятых в экономике и изменения численности населения.

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

Рисунок 32. Поле кластера с самым значимым параметром среднедушевого дохода

Для отображения данных по кластеру обратимся к детализации. Это поможет определить, какие регионы попали в этот кластер. Это оказались Пензенская и Самарская области.

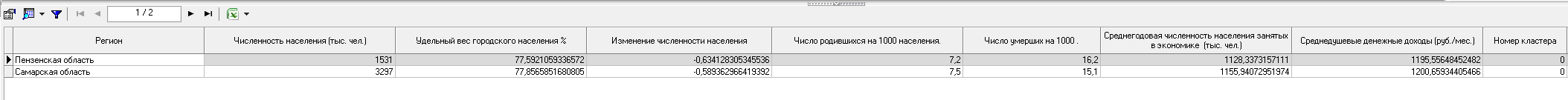


Рисунок 33. Регионы, попавшие в кластер 0

После этого был добавлен визуализатор «Многомерная диаграмма» с отображением регионов, номеров кластеров и (по очереди) исследуемые параметры. Далее на рисунках 34-37 будут представлены многомерные диаграммы для полей «Удельный вес городского населения», «Изменение численности населения», «Среднегодовая численность населения занятых в экономике» и «Среднедушевой денежный доход» соответственно. Легенды находятся под многомерными диаграммами.

Изображение выглядит как текст, Прямоугольник, Параллельный, линия

Автоматически созданное описание

Рисунок 34. Многомерная диаграмма для удельного веса городского населения в процентах

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Рисунок 35. Многомерная диаграмма для изменения численности населения

Изображение выглядит как текст, Параллельный, диаграмма

Автоматически созданное описание

Рисунок 36. Многомерная диаграмма для среднегодовой численности населения занятых в экономике

Изображение выглядит как текст, снимок экрана, Параллельный

Автоматически созданное описание

Рисунок 37. Многомерная диаграмма для среднедушевого денежного дохода

Следующим будет добавлен визуализатор «Диаграмма размещения». На рисунке 38 будет представлена диаграмма размещения для удельного веса городского населения в процентах, для всех остальных параметров будут похожие диаграммы, поэтому дальше они не будут представлены.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок 38. Диаграмма размещения для удельного веса городского населения в процентах

10. Теперь проведём такой же анализ для обработанных данных для 4 кластеров вместо 5. Для отображения и анализа будут использоваться «Профили кластеров» и «Диаграмма размещения».

Изображение выглядит как текст, снимок экрана, число, диаграмма

Автоматически созданное описание

Рисунок 37. Профили кластеров для новых 4 кластеров

Самый значимый параметр для среднедушевого дохода оказался снова в 0 кластере (см. рис. 37), однако на этот раз сюда попали Псковская область и Чукотский автономный округ.

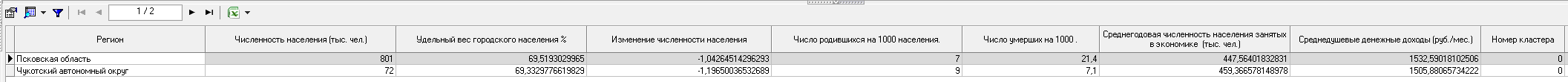


Рисунок 38. Регионы, попавшие в кластер 0

В этом же кластере находятся самые значимые параметры удельного веса городского населения, среднегодовая численность населения занятых в экономике и изменения численности населения.

После составления диаграммы размещения для последней кластеризации (см. рис. 39), состоящей их 4 кластеров, вернёмся к прошлой кластеризации, где были использованы 5 кластеров, и составим такую же диаграмму для неё (см. рис. 40).

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 39. Диаграмма размещения для удельного веса городского населения в процентах. Для большей наглядности форма объекта – номер кластера, цвет – среднегодовая численность населения занятых в экономике, а размер – среднедушевые денежные расходы.

Изображение выглядит как текст, снимок экрана, диаграмма, программное обеспечение

Автоматически созданное описание

Рисунок 40. Диаграмма размещения для прошлой кластеризации

11-12. После проведённого анализа, можно выделить кластеризацию с 4 кластерами, поэтому копируем и переносим этот узел к необработанным данным. Для импортированных данных кластеризация выглядит похожим образом.

Самый значимый параметр для среднедушевого дохода оказался снова в 0 кластере (см. рис. 37), однако теперь сюда попали республика Ингушетия и Чеченская республика.

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

Рисунок 41. Профили кластеров для необработанных данных

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рисунок 42. Диаграмма размещения для кластеризации на необработанных данных

### Часть 3

1. Для выполнения этого задания была создана папка «Карты Кохонена», куда был скопирован файл Banks.txt. Также был создан проект в Deductor Studio и также размещён в этой папке. После, файл Banks.txt был импортирован в проект в Deductor.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 43. Папка «Карты Кохонена» на данном этапе

2. Далее был запущен Мастер обработки и выбран обработчик «Карта Кохонена».

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 44. Выбор «Карты Кохонена» в Мастере обработки

На втором шаге поле Прибыль было указано выходным столбцом, а поля Филиалы, Сумма активов, Собственные активы, Банковские активы, Средства банке – входными.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рисунок 45. Распределение назначения столбцов для «Карты Кохонена»

Далее значения были оставлены по умолчанию для обучающего и тестового множеств, настроек параметров карты Кохонена, за исключением числа кластеров: они были заданы фиксированным числом кластеров 4.

3. После завершения процесса обучения в списке визуализаторов была выбрана «Карта Кохонена» и «Что–если».

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 46. Выбор режима отображения

После этого в следующем окне были применены настройки для карты Кохонена: все входные и выходные столбцы и кластеры, способ раскраски – цветная палитра, также включено отображение границ кластеров.

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Автоматически созданное описание

Рисунок 47. Настройка отображения карты Кохонена

Далее будет проанализированы области с кластером меньшего размера. В данном случае это кластер 1.

Изображение выглядит как линия, Красочность, Цвет электрик, График

Автоматически созданное описание

Рисунок 48. Карты Кохонена

После определения наименьшего кластера (в данном случае номер 1), будут выведены данные, которые попали в этот кластер.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Рисунок 49. Банки, попавшие в 1 кластер

Далее был выбран самый большой кластер, и для него были проанализированы финансовые показатели (они оказались примерно одинаковыми в этом кластере).

Изображение выглядит как Красочность, снимок экрана, Цвет Majorelle blue, Цвет электрик

Автоматически созданное описание

Рисунок 50. Самый большой кластер 0

5. После этого было изучено распределение кластеров на одной из карт с финансовыми показателями – «Сумма активов». Добавлен визуализатор «Многомерная диаграмма» для этого же показателя (оси: номер кластера, банк, сумма активов).

Изображение выглядит как текст, рукописный текст

Автоматически созданное описание

Рисунок 51. Многомерная диаграмма для показателя «Сумма активов»

Изображение выглядит как снимок экрана, текст, Красочность, Мультимедийное программное обеспечение

Автоматически созданное описание

Рисунок 52. Карта Кохонена для показателя «Сумма активов»

Как на многомерной диаграмме, так и на карте Кохонена отчётливо и наглядно видно, как расположены суммы активов в разных кластерах. Также в многомерной диаграмме можно оценить какому банку принадлежат те или иные суммы активов.

6. С помощью карты «Филиалы» будет проверено, что число филиалов не сказалось на результатах кластеризации. Для этого была добавлена визуализация «Профилей кластеров». Значимость для показателя «Филиалы» имеет небольшие отклонения при проверке на соответствие с картой Кохонена, но они в пределах нормы. Следовательно, можно сделать вывод, что количество филиалов не сказалось на результатах кластеризации.

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

Рисунок 53. Профили кластеров для показателя «Филиалы»

Изображение выглядит как снимок экрана, текст, Мультимедийное программное обеспечение, Графическое программное обеспечение

Автоматически созданное описание

Рисунок 54. Карта Кохонена для показателя «Филиалы»

7..

• Создайте еще один узел сценария – аналогичную Карту Кохонена для тех же данных, указав. Дайте различные названия обоим узлам сценария.

• Визуализируйте таблицу Профили кластеров. Проверьте значимость поля Прибыль для разных кластеров .

• Сравните оба результата кластеризации на картах Кохонена.

Далее с помощью карты «Прибыль» будет проверено, имеется ли связь прибыльности банков и их группировки. Будет создан ещё один узел сценария, то есть просто скопирован уже имеющийся, и изменены некоторые параметры: вместо входного параметра «Филиалы» будет входной параметр «Прибыль». Первый узел будет называться «Карта Кохонена (Филиалы)», второй – «Карта Кохонена (Прибыль)».

Изображение выглядит как текст, снимок экрана, диаграмма, число

Автоматически созданное описание

Рисунок 53. Профили кластеров для показателя «Прибыль»

Изображение выглядит как снимок экрана, Красочность, текст

Автоматически созданное описание

Рисунок 54. Карта Кохонена для показателя «Прибыль»

При сравнении обоих результатов кластеризации можно заметить, что данные примерно одинаковой цветовой гаммы. Исходя из карт Кохонена можно сказать, что связи банков и их группировкой неболшая.

Изображение выглядит как текст, снимок экрана, Красочность, Графика

Автоматически созданное описание

Рисунок 55. Карта Кохонена для показателя «Прибыль» при первой кластеризации