ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

В. В. Малахова, О. В. Малахов

Луганский государственный университет им. В. И. Даля, Россия
E-mail: malakhova_viktoriya84@mail.ru, oleg_home1@mail.ru

В работе рассматривается процедура проведения анализа неструктурированных статистических данных с использованием алгоритмов искусственного интеллекта. Последовательность действий подробно описана с разбиением на следующие этапы: сбор информации, формирование исходных данных, выбор модели машинного обучения и метода кластеризации данных, обучение модели кластеризации данных, выполнение анализа данных с использованием алгоритмов искусственного интеллекта.

USING ARTIFICIAL INTELLIGENCE ALGORITHMS TO ANALYZE UNSTRUCTURED STATISTICAL DATA

V. V. Malakhova, O. V. Malakhov

The paper considers the procedure for analyzing unstructured statistical data using artificial intelligence algorithms. The sequence of actions is described in detail, divided into the following stages: collecting information, generating source data, choosing a machine learning model and a data clustering method, training a data clustering model, performing data analysis using artificial intelligence algorithms.

Введение. В современном мире искусственный интеллект занял прочное положение в поисковых системах, активно применяется при обработке больших объемов неструктурированной информации, для управления технологическими процессами и подвижными объектами без участия человека [1].

Владение технологиями искусственного интеллекта, либо не владение ими, в настоящее время стало тем критерием, по которому целые страны и народы либо перейдут в новую эпоху развития Человечества, либо так и останутся на задворках цивилизации [2].

Постановка задачи. Разработать и описать процедуру проведения анализа неструктурированных статистических данных с использованием алгоритмов искусственного интеллекта.

Этап 1 – Сбор исходных данных. В качестве типичного примера источника первичной информации используем справочник «Россия в цифрах» [3] – официальное издание Росстата. Статистические данные представлены таблицей, заголовок и начальная часть которой показаны в таблице. Поля таблицы содержат текстовую информацию и цифровые данные в формате вещественных чисел.

Исходная форма представления статистических данных

	Площадь территории, тыс. км^2	Численность населения, тыс. человек	Среднегодовая численность занятых, тыс. человек	Среднедушевые денежные доходы в месяц, руб.	Среднедушевые денежные расходы в месяц, руб.	Среднемесячная номинальная начисленная заработная плата работников организаций, руб.	Валовый региональный продукт (в текущих основных ценах), млрд. руб.	Основные фонды в экономике (по полной учетной стоимости), млрд. руб.	Объем отгруженных товаров собственного производства, выполненных работ и услуг собственными силами, млн. руб.			Продукция сельского хозяйства, млн. руб.	Ввод в действие общей площади жилых помещений, тыс. м^2	Оборот розничной торговли, млрд. руб.	Сальдированный финансовый результат (прибыль минус убыток) в экономике, млн. руб.	Индекс потребительских цен, процентов	Инвестиции в основной капитал, млрд.руб.
									Добыча полезных ископаемых	Обрабатывающие производства	Производство и распределение электроэнергии, газа и воды						
Белгородская область	27,1	1552,9	698,1	30024	26998	26873	686,4	1290	88757	568628	27415	226544	1350,1	298,7	213778	104,4	143,8
Брянская область	34,9	1220,5	522,2	25606	25030	22819	269,9	627	280	171421	16421	78312	665,1	219,9	16516	106,1	68,3

Этап 2 – Формирование исходных данных. Производим сканирование страниц издания и распознавание присутствующей в составе полученных изображений текстовой информации. Первые строки полученного текста, содержащего информацию о социально-экономических показателях с разбивкой по областям, приведены в листинге 1.

```
Белгородская область| 27,1| 1552,9| 698,1| 30024| 26998| 26873| 686,4|
Брянская область    | 34,9| 1220,5| 522,2| 25606| 25030| 22819| 269,9|
```

Листинг 1. Исходные данные, приведенные к текстовому формату

Выполняем преобразование полученного текста в формат CSV – текстовую форму представления электронных таблиц [4]. В документе формата CSV каждая строка таблицы представлена строкой текстового файла с разделителями. В процессе преобразования запятая, символ десятичного разделителя целой и дробной части вещественных чисел, подлежит замене на символ «точка». В качестве разделителя используем символ «точка с запятой» (листинг 2).

```
Белгородская область;27.1;1552.9;698.1;30024;26998;26873;686.4;
  Брянская область;34.9;1220.5;522.2;25606;25030;22819;269.9;
```

Листинг 2. Представление исходных данных в формате CSV – кода

Для обработки информации средствами языка программирования *Python* используем библиотеку математических функций с открытым исходным кодом *numpy* [5]. Загружаем исходный CSV файл с образованием многомерного массива X (листинг 3).

```
import numpy as np
X = np.loadtxt('data_1.csv', delimiter=';')
```

Листинг 3. Загрузка исходного CSV файла
в массив средствами библиотеки *numpy*

В качестве двумерного тестового набора данных для проведения анализа

выберем:

- среднедушевые денежные доходы в месяц (руб./чел.) A;
- отношение валового регионального продукта (млн. руб.) к численности населения региона (тыс. человек) B.

Визуализация тестового набора данных, выполненная средствами *Python* с использованием графической библиотеки *matplotlib.pyplot*, приведена на рис. 1.

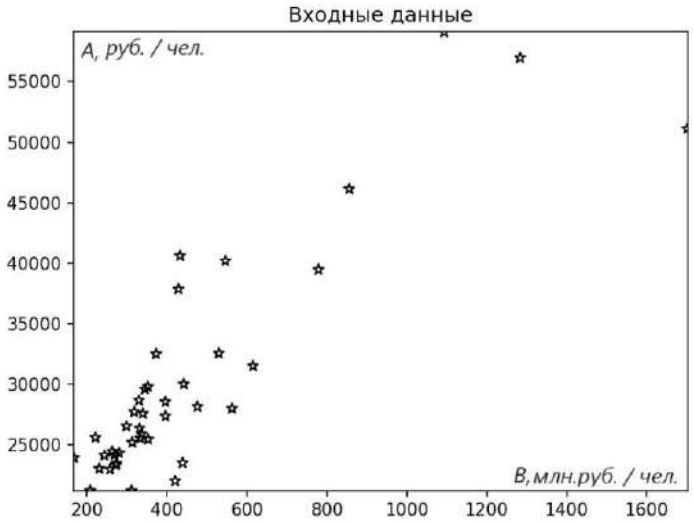

Рис. 1. Визуализация исходных данных

Этап 3 – Выбор модели машинного обучения и метода кластеризации данных. В общем случае обучение алгоритмов классификации производится по тренировочным данным, с которыми ассоциированы определенные метки (маркеры). После прохождения этапа обучения алгоритм искусственного интеллекта может классифицировать немаркированные, но однотипные с тестовыми данные, относя их к той или иной группе (кластеру).

При известном заранее количестве кластеров N можно воспользоваться методом *k*-средних (*k-means*) [6]. Основная идея метода заключается в циклическом обновлении центров тяжести кластеров (центроидов). Итеративный процесс продолжается до тех пор, пока при выполнении очередной итерации все центроиды не сдвинутся более, чем на некоторую заданную величину погрешности, т.е. с определенной точностью займут свои оптимальные положения. Выбор положения центроида каждого кластера направлен на минимизацию инерции или суммы квадратов внутри кластера:

$$\sum_{i=0}^{N} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right),$$

где μ_j – положение центроида в кластере C;

x_i – образцы в кластере K_j.

Рассмотрим более сложный механизм формирования модели машинного обучения, относящейся к категории искусственного интеллекта «обучение без учителя». Построим модель машинного обучения без привлечения маркирован-

ных тренировочных данных. Предоставим возможность искусственному интеллекту самому распределить представленные данные по категориям, которые пока нам неизвестны. Полагаем, что набор исходных данных генерируется под влиянием неявных факторов, управляющих их распределением.

В рамках проводимого исследования выполняем кластеризацию данных принимая условие, что количество кластеров заранее не известно. Воспользуемся непараметрическим алгоритмом обучения, основанном на методе сдвига среднего (*Mean Shift*) [7]. Данный алгоритм рассматривает все пространство признаков как функцию распределения вероятности. При этом в базовом распределении существуют *K* пиков, соответствующих центрам тяжести такого же количества кластеров. Результатом выполнения алгоритма, реализующего метод сдвига среднего является определение максимального количества кластеров и идентификация их центров.

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)},$$

где $N(x_i)$ – соседство образцов на заданном расстоянии вокруг x_i;

$m(x_i)$ – вектор среднего сдвига, указывающий на максимальную плотность точек.

Этап 4 – Обучение модели кластеризации данных. Одним из основных параметров чувствительности базового процесса при оценке плотности распределения данных в алгоритме сдвига среднего является ширина окна (*bandwidh*). Малая ширина окна способствует появлению большого количества кластеров, излишне широкое окно приводит к слиянию отдельных кластеров в более крупные. Следует учесть, что отклонение ширины окна как в одну, так и в другую сторону от оптимального значения приводит к снижению информативности результатов кластеризации данных. Подключаем модуль *sklearn.cluster* [8] библиотеки *sklearn*. Ширина окна задается величиной параметра *quantile* (листинг 4).

```
from sklearn.cluster import MeanShift, estimate_bandwidth
bandwidth_X = estimate_bandwidth(X, quantile=0.1, n_samples=len(X))
```

Листинг 4. Подключение модуля *sklearn.cluster*,
оценка ширины окна для массива *X*

Производим обучение модели кластеризации с учетом заданной оценки ширины окна (листинг 5).

```
meanshift_model = MeanShift(bandwidth=bandwidth_X, bin_seeding=True)
                    meanshift_model.fit(X)
```

Листинг 5. Обучение модели кластеризации на основе метода сдвига среднего

Этап 5 – Выполнение анализа данных с использованием алгоритмов искусственного интеллекта. Определяем количество кластеров *num_clusters*

для заданной ширины окна и массив координат их центров *cluster_centers* (листинг 6).

```
labels = meanshift_model.labels_
num_clusters = len(np.unique(labels))
cluster_centers = meanshift_model.cluster_centers_
```

Листинг 6. Определение количества
кластеров и расположение их центров

Результат работы программы в текстовом виде представлен на листинге 7.

```
Количество кластеров = 9
    Центры кластеров:
[[  283.60666667 24341.33333333]
 [  356.675      26623.41666667]
 [  395.86       28550.9       ]
 [  546.675      39541.75      ]
 [  505.73333333 32197.66666667]
 [ 1701.5        51124.        ]
 [ 1283.1        56974.        ]
 [ 1093.         59097.        ]
 [  855.8        46135.        ]]
```

Листинг 7. Результат определения количества
кластеров и расположения их центров

Графическое представление результата с использованием библиотеки *matplotlib.pyplot* приведено на рис. 2.

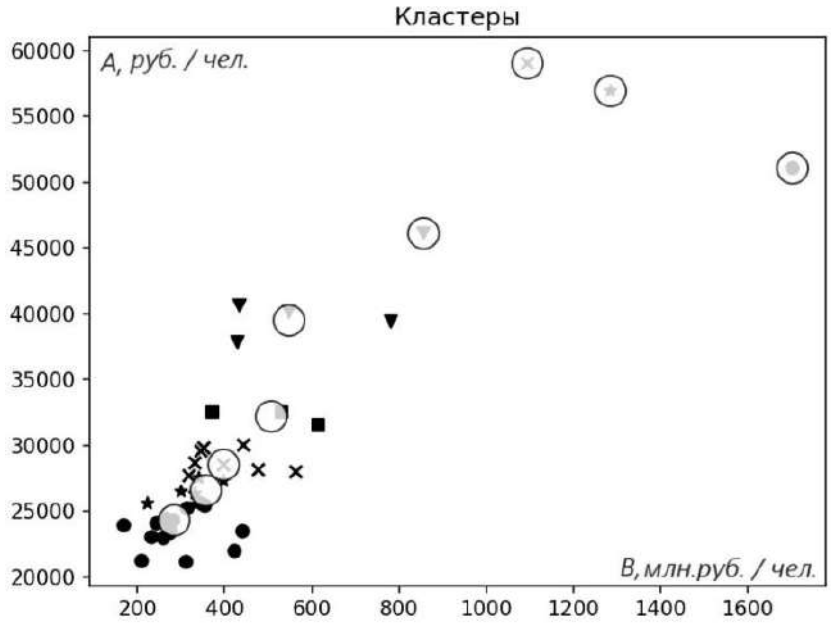

Рис. 2. Графическое представление результата кластеризации данных

Заключение. В работе рассматривается процедура проведения анализа не-

структурированных статистических данных с использованием алгоритмов искусственного интеллекта. Последовательность действий подробно описана с разбиением на следующие этапы: сбор информации, формирование исходных данных, выбор модели машинного обучения и метода кластеризации данных, обучение модели кластеризации данных, выполнение анализа данных с использованием алгоритмов искусственного интеллекта.

Использование алгоритмов искусственного интеллекта при обработке больших массивов неструктурированных данных позволяет производить кластеризацию представленных данных как опираясь на известные ранее критерии, так и способом поиска различных изначально неявных метрик сходства.

СПИСОК ЛИТЕРАТУРЫ

1. *Киселев Б. Н.* Результат интеллектуальной деятельности, созданный искусственным интеллектом, и результат интеллектуальной деятельности, созданный при помощи искусственного интеллекта // Актуальные проблемы науки и практики: Гатчинские чтения-2019 : Сборник науч. трудов по материалам VI Междун. науч.-практич. конф. 2019. Т. 2. С. 123-126.

2. *Гутаревич В. О., Рябко К. А., Рябко Е. В.* Проблемы и направления совершенствования экологических характеристик горно-транспортных машин с дизельной установкой // Вестник Донецкого национального технического университета. 2018. № 1 (11). С. 12-17.

3. Россия в цифрах. 2017 : Краткий статистический сборник / Федеральная служба государственной статистики (Росстат). М., 2017. 511 с. [Электронный ресурс]. URL: http://komitet4.km.duma.gov.ru/upload/site28/rus17.pdf?ysclid=lq9j5gnsus110188126 (дата обращения 10.10.2023).

4. Учебник Python 3. CSV. [Электронный ресурс]. URL: https://learn4kid-python.firebaseapp.com/python_data_structure/python_csv_all/#пример-фаила-csv (дата обращения 24.09.2023).

5. Библиотека математических функций Python numpy. [Электронный ресурс]. URL: https://numpy.org (дата обращения 24.09.2023).

6. Кластеризация K-means. [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/clustering.html#k-means (дата обращения 24.09.2023).

7. *Московкин В. М., Сизъунго М., Голиков Н. А.* Кластерный анализ инновационной активности регионов РФ на основе метода Mean Shift // Оригинальные исследования. 2019. Т. 9. № 6. С. 117-149.

8. Кластеризация Mean Shift. [Электронный ресурс]. URL: https://pythonprogramming.net/mean-shift-from-scratch-python-machine-learning-tutorial/ (дата обращения 24.09.2023).