УДК 004.85

РЕШЕНИЕ ЗАДАЧИ КЛАСТЕРИЗАЦИИ НА МНОЖЕСТВЕ ЧИСЛОВЫХ И СИМВОЛЬНЫХ ДАННЫХ, ГРАФИЧЕСКИХ ОБЪЕКТАХ

Хлебородова Лидия Дмитриевна

студент

Научный руководитель: **Осипов Геннадий Сергеевич**

д.т.н., зав. кафедрой информатики

ФГБОУ ВО «Сахалинский государственный университет»

Аннотация: Разработан программный комплекс, предназначенный для решения задачи кластеризации. Исследованы и практически реализованы процедуры разбиения на кластеры объектов, характеризующихся разнородной информацией: текстовыми и числовыми данными, а также фото и графическими цифровыми примитивами.

Ключевые слова: кластеризация, методы машинного обучения.

SOLVING THE CLUSTERING PROBLEM ON A SET OF NUMERIC AND SYMBOLIC DATA, GRAPHIC OBJECTS

Khleborodova Lydia Dmitrievna

Abstract: A software package designed to solve the clustering problem has been developed. Procedures for clustering objects characterized by heterogeneous information: text and numerical data, as well as photo and graphic digital primitives have been studied and practically implemented.

Key words: clustering, machine learning methods.

Цель работы – разработка программного комплекса, позволяющего решать задачу кластеризации объектов, характеристиками которых могут быть числовые величины, текстовая информация, фото и иные графические объекты.

В качестве программной оболочки был выбран современный пакет символьной математики и компьютерной алгебры Wolfram Mathematica [1].

1. Кластеризация объектов числового типа

Проведем кластеризацию объектов, информация о которых представлена в [2]. В данном случае все параметры объектов являются действительными числами. На *рисунке 1* представлен фрагмент данных, которые использовались для решения задачи кластеризации.

	A	B	C	D	E	F	G
1	area A	perimeter P	compact	length of	width of	asymme	length of
2	15.26	14.84	0.871	5.763	3.312	2.221	5.22
3	14.88	14.57	0.8811	5.554	3.333	1.018	4.956
4	14.29	14.09	0.905	5.291	3.337	2.699	4.825
5	13.84	13.94	0.8955	5.324	3.379	2.259	4.805
6	16.14	14.99	0.9034	5.658	3.562	1.355	5.175
7	14.38	14.21	0.8951	5.386	3.312	2.462	4.956
8	14.69	14.49	0.8799	5.563	3.259	3.586	5.219
9	14.11	14.1	0.8911	5.42	3.302	2.7	5
10	16.63	15.46	0.8747	6.053	3.465	2.04	5.877
11	16.44	15.25	0.888	5.884	3.505	1.969	5.533
12	15.26	14.85	0.8696	5.714	3.242	4.543	5.314
13	14.03	14.16	0.8796	5.438	3.201	1.717	5.001
14	13.89	14.02	0.888	5.439	3.199	3.986	4.738
15	13.78	14.06	0.8759	5.479	3.156	3.136	4.872
16	13.74	14.05	0.8744	5.482	3.114	2.932	4.825
17	14.59	14.28	0.8993	5.351	3.333	4.185	4.781

Рис. 1. Фрагмент исходного массива данных

Кластеризация осуществлялась с помощью процедуры FindCluster, параметрами которой являются наименование набора данных и количество кластеров, на которое будет разбито исходное множество. Результат выводится в виде списка кластеров и перечисления соответствующих им элементов (см. рисунок 2).

```
FindClusters[x, 3] (*Разделяем на 3 кластера*)
[найти кластеры]

{{ {15.26, 14.84, 0.871, 5.763, 3.312, 2.221, 5.22}, {
  {14.29, 14.09, 0.905, 5.291, 3.337, 2.699, 4.825},
```

Рис. 2. Разбиение на кластеры

Рис. 3. Последовательное перечисление номеров кластеров

```
c = ClusterClassify(x, 3, Method → "KMeans")
[классифицировать по клас... [метод]

ClassifierFunction [
Input type: Mixed (number: 7)
Classes: 1, 2, 3
Method: KMeans
Number of training examples: 207
Data not in notebook. Store now
]

Классифицируем новый объект

c[{"13.2", "13.66", "0.8883", "5.236", "3.232", "8.315", "5.056"}, "Probabilities"]
<| 1 → 0.00267564, 2 → 6.17945 × 10-16, 3 → 0.997324 |>
```

В данном случае с вероятностью 99,7% новый объект относится к третьему кластеру.

МЦНП «Новая наука»

Рис. 5. Информация об используемом методе

2. Кластеризация при смешанных данных

Решим задачу кластеризации автомобилей. Информация взята из [3, 4] и представляет собой смешанные числовые и символьные данные (см. рисунок 6).

	A	B	C	D	E	F
1	x1	x2	x3	x4	x5	x6
2	vhigh	vhigh	2	2	small	low
3	vhigh	vhigh	2	2	small	med
4	vhigh	vhigh	2	2	small	high
5	vhigh	vhigh	2	2	med	low
6	vhigh	vhigh	2	2	med	med
7	vhigh	vhigh	2	2	med	high
8	vhigh	vhigh	2	2	big	low
9	vhigh	vhigh	2	2	big	med
10	vhigh	vhigh	2	2	big	high
11	vhigh	vhigh	2	4	small	low
12	vhigh	vhigh	2	4	small	med
13	vhigh	vhigh	2	4	small	high
14	vhigh	vhigh	2	4	med	low
15	vhigh	vhigh	2	4	med	med
16	vhigh	vhigh	2	4	med	high
17	vhigh	vhigh	2	4	big	low

Рис. 6. Фрагмент массива исходных данных

Результаты разбиение объектов на кластеры представлены на *рисунке 7*.

[найти кластеры](#)

```
ClusteringComponents[x, 4, 1]
```

кластерные компоненты

{1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1,

Рис. 7. Результаты кластеризации

На рисунке 8 представлены операторы, характеризующие используемый метод кластеризации и отнесение нового объекта к соответствующему кластеру.

классифицировать по клас... метод

Input type: Mixed (number: 6)
Classes: 1, 2, 3, 4
Method: KMedoids
Number of training examples: 1725

информация

The k-medoids clustering algorithm iteratively minimize the distance between points labeled to be in a cluster and a point belonging to the cluster designated as the center of it (medoid).

Классифицируем новый объект

 $\langle |1 \rightarrow 0.0761576, 2 \rightarrow 0.275987, 3 \rightarrow 0.636986, 4 \rightarrow 0.0108697| \rangle$

Рис. 8. Информация о методе и кластеризация нового объекта

Создадим массив цветowych объектов (см. рисунок 9) и проведем их разбиение, например, на 4 кластера.

Рис. 9. Результат разбиения на классы цветовых объектов

Рис. 10. Обработка нового цветового объекта

Рис. 11 Пример кластеризации фотографий

В данном случае новый объект отнесен к кластеру номер 3.

Основные результаты

Разработана программа, которая позволяет решать задачу кластеризации объектов разного типа. Программное средство реализовано на платформе Wolfram Mathematica, представляющее собой современный пакет символьной математики.

Список литературы

1. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: <https://www.wolfram.com/language/elementary-introduction/2nd-ed/> (Дата обращения 29.10.2021).
2. Seeds Data Set. URL: <https://archive.ics.uci.edu/ml/datasets/seeds> (Дата обращения 04.02.2022).

3. Car Evaluation Data Set.

4. URL: <https://archive.ics.uci.edu/ml/datasets/car+evaluation> (Дата обращения 04.02.2022).

5. Хлебородова Л.Д. Сравнительный анализ методов машинного обучения для решения задачи классификации в среде Wolfram Mathematica / Л.Д. Хлебородова, Г.С. Осипов // Международный форум молодых исследователей: Сборник статей Международной научно-практической конференции, Петрозаводск, 09 декабря 2021 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука». 2021. – С. 214-221. DOI: 10.46916/13122021-6-978-5-00174-401-6.

© Л.Д. Хлебородова, 2022