УДК 004.8

# О РЕАЛИЗАЦИИ МЕТОДОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ

Е. М. БОРЧИК, М. В. АЛЕКСЕЙКОВ, Д. В. МИКАЛУЦКИЙ
Белорусско-Российский университет
Могилев, Беларусь

Пусть задано множество наблюдений $X = \{x_1, \ldots, x_m\}$, $x_i \in R^n$, которое необходимо разделить на непересекающиеся подмножества (кластеры). Для решения данной задачи используются методы кластерного анализа, такие как K-Means, Tree и Fuzzy Relation Clustering и др. [1].

Ранее были рассмотрены вопросы совместного использования методов кластерного анализа многомерных данных для уточнения результатов кластеризации данных несколькими методами с учетом положительных сторон и особенностей работы самих методов и обобщения результатов кластеризации несколькими методами [1].

Для практического применения при построении кластеров многомерных данных в рамках ПТКИ Belsim2#.random [1] реализован метод Fuzzy Relation Clustering (FRC). Методы K-Means Clustering, Tree Clustering ранее были доступны в ПО STATISTICA. В данный момент методы нечеткого кластерного анализа рассматриваются в рамках изучения курса «Интеллектуальные информационные системы». Студенты применяют встроенные библиотеки языка Python для изучения работы методов Fuzzy C-Means (метод, родственный K-Means), Fuzzy Relation Clustering.

В ходе решения задачи о разделении пациентов на кластеры по схожести заболеваний разработан программный модуль (ПМ) на Python с использованием библиотек pandas, numpy, skfuzzy и sys, реализующий метод кластерного анализа Fuzzy C-Means, позволяющий группировать в кластеры объекты, между которыми есть последовательность «близких» друг к другу элементов, что также соответствует интуитивному представлению о группировке. Исходные данные считываются из файла формата *.csv. Файл данного формата может быть открыт для просмотра и заполнения в Excel, как показано на рис. 1 для файла (base.csv).

| | Пациент 1 | Пациент 2 | Пациент 3 | Пациент 4 | Пациент 5 |
|---|---|---|---|---|---|
| Гемофилия | 0 | 1 | 0 | 1 | 0 |
| Кистозный фиброз | 1 | 0 | 1 | 0 | 1 |
| ДЦП | 0 | 1 | 0 | 0 | 1 |
| Синдром Дауна | 0 | 1 | 0 | 1 | 1 |

Рис. 1. Пример формата ввода входных данных о пациентах

Выходные данные с номерами построенных кластеров в транспонированном виде записываются в файл 'outputF.csv'. Нумерация кластеров начинается с 0. Разработанное ПО и файл исходных данных base.csv размещаются в одном каталоге. Файл 'outputF.csv' создается по итогам работы ПМ (рис. 2).

После запуска из командной строки ПМ на исполнение в ходе работы

алгоритма первым шагом запрашивается имя файла с исходными данными – например, base3_2.csv, вводится с клавиатуры планируемое для построения количество кластеров, задается порог нечеткости Альфа (рис. 3). Итоговые данные записываются в 'outputF.csv'.

| | Гемофилия | Кистозный фиброз | ДЦП | Синдром Дауна | Cluster |
|---|---|---|---|---|---|
| Пациент 1 | 0 | 1 | 0 | 0 | 1 |
| Пациент 2 | 1 | 0 | 1 | 1 | 0 |
| Пациент 3 | 0 | 1 | 0 | 0 | 1 |
| Пациент 4 | 1 | 0 | 0 | 1 | 0 |
| Пациент 5 | 0 | 1 | 1 | 1 | 2 |

Рис. 2. Выходные данные с итогами разделения на кластеры

```
Введите название файла : base3_2.csv
Введите количество кластеров: 3
Введите альфу: 0.65
Название выходного файла:  outputF.csv
Нажмите любую клавишу для завершения РАБОТЫ!!!
```

Рис. 3. Шаги ввода исходных данных в ПМ, итоги работы ПМ

При этом заголовки строк и столбцов в файле с исходными данными (base3_2.csv) могут быть переименованы с учетом исследуемых данных, количества элементов и их размерности. Например, при решении задачи о разделении на кластеры четырех межпозвонковых расстояний десяти пациентов файл преобразован, как показано на рис. 4. Данные в base3_2.csv разделяются символом «;», дробные части десятичных чисел отделяются символом «.» . В файле 'outputF.csv' аналогично – данные разделяются символом «,», дробные части десятичных чисел отделяются символом «.» (рис. 5).

base3_2.csv – Блокнот

Файл Правка Формат Вид Справка

```
;a1;a2;a3;a4;a5;a6;a7;a8;a9;a10
1 позвонок;0.679558011;0.552486188;0.613259669
2 позвонок;0.73480663;0.580110497;0.662983425;
3 позвонок;0.91160221;0.679558011;0.861878453;
4 позвонок;0.690607735;0.629834254;0.729281768
```

Рис. 4. Пример ввода исходных данных в формате *.csv

outputF.csv – Блокнот

Файл Правка Формат Вид Справка

```
,1 позвонок,2 позвонок,3 позвонок,4 позвонок,Cluster
a1,0.679558011,0.73480663,0.91160221,0.690607735,0
a2,0.552486188,0.580110497,0.679558011,0.629834254,1
a3,0.613259669,0.662983425,0.861878453,0.729281768,1
a4,0.596685083,0.563535912,0.585635359,0.806629834,1
a5,0.569060773,0.679558011,0.762430939,0.580110497,1
```

Рис. 5. Кластеризованные данные в формате *.csv

В шагах работы ПМ осуществляется проверка данных на пустоту после транспонирования, выдается рекомендация проверки входного файла и его формата в случае ввода некорректных исходных данных.

Файл 'outputF.csv' может быть открыт в Excel с учетом разделительного знака «,», что обеспечивает простоту и удобство работы с данными.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. **Якимов, А. И.** Интеллектуальный анализ данных для имитационного моделирования производственных систем: монография / А. И. Якимов, Е. А. Якимов, Е. М. Борчик. – Могилев : Белорус.-Рос. ун-т, 2021. – 184 с.