УДК 519.6

ПРИМЕНЕНИЕ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ДЛЯ КОМПЛЕКСНОЗНАЧНЫХ ДАННЫХ

И.А. Седых
Липецкий государственный технический университет
Россия, 398055, Липецк, Московская ул., 30
E-mail: sedykh-irina@yandex.ru

К.Н. Макаров
Липецкий государственный технический университет
Россия, 398055, Липецк, Московская ул., 30
E-mail: kirik0-1@yandex.ru

Ключевые слова: комплекснозначные данные, кластерный анализ, нечеткая кластеризация, кластер, метод нечетких с-средних, метод нечетких с-эллипсоидов.

Аннотация: В данной работе рассматриваются некоторые методы нечеткой кластеризации. Даны их определения, реализованы алгоритмы на примере нечеткой кластеризации комплекснозначных данных. Представлены графики результатов, а также сделан вывод о проделанной работе.

1. Введение

Комплекснозначные данные используются в различных сферах деятельности, например, в робототехнике, космонавтике. При этом актуальным является изучение, моделирование, анализ и прогнозирование поведения сложных систем.

Кластерный анализ – математическая процедура, позволяющая на основе схожести количественных значений нескольких признаков, свойственных каждому объекту (например, испытуемому) какого–либо множества, сгруппировать эти объекты в определенные классы, или кластеры. Осуществляется путем вычисления расстояния между этими объектами [1]

Кластерный анализ – это обобщенное название достаточно большого набора алгоритмов, используемых при классификации данных. В ряде изданий используются и такие синонимы кластерного анализа, как классификация и разбиение.

В данной статье приводятся примеры нечеткой кластеризация комплекснозначных данных, которые состоят из двух столбцов комплексных чисел.

2. Нечеткая кластеризация

Нечеткая кластеризация – метод кластеризации, при котором точки данных могут принадлежать более чем к одной группе («кластеру»). Кластеризация делит точки данных на группы и направлена на поиск закономерностей или сходства между элементами в наборе; элементы в кластерах должны быть как можно более похожи друг на друга и как можно более непохожи на элементы в других группах. С точки зрения вычислений гораздо проще создать нечеткие границы, чем привязать один кластер к одной точке [2].

В работе рассмотрены два популярных метода кластеризации – нечетких с-средних и нечетких с-эллипсоидов.

2.1. Метод нечетких с-средних

Этот метод, разработанный Даном в 1973 году и улучшенный Бездеком в 1981 году, часто используется в распознавании образов [3]. Строго говоря, этот алгоритм работает путем присвоения степени принадлежности каждой точке данных, соответствующей центру кластера, на основе расстояния между кластером и точкой данных [4]. Чем ближе данные к центру кластера, тем больше их степень принадлежности к конкретному центру кластера. Очевидно, что суммирование степеней принадлежности каждой точки данных должно быть равно единице [5].

В таблице 1 показана схема алгоритма нечеткой кластеризации методом с-средних.

Таблица 1. Алгоритм нечеткой кластеризации методом с-средних.

Шаг	Базовый алгоритм нечетких с-средних
1	Установить параметры алгоритма: C - количество кластеров; m - экспоненциальный вес; ε- параметр остановки алгоритма.
2	Случайным образом сгенерировать матрицу нечеткого разбиения $U = [u_{ij}]; u_{ij} \in [0,1];\ i \in 1, \dots, N; j \in 1, \dots, C.$
3	Рассчитать центры кластеров: $c_j = \frac{\sum_{i=1}^{N} (u_{ij})^m x_i}{\sum_{i=1}^{N} (u_{ij})^m}.$
4	Рассчитать расстояния между объектами из $X = x_1, \dots, x_N$ и центрами кластеров: $D_{ij} = \sqrt{\|x_i - c_j\|^2}.$
5	Пересчитать элементы матрицы нечеткого разбиения: $u_{ij} = \frac{1}{\left({D_{ij}}^2 \sum_{k=1}^{C} \frac{1}{{D_{ik}}^2} \right)^{\frac{1}{m-1}}}$, если $D_{ij} > 0$; $u_{ik} = \begin{cases} 1, k = j \\ 0, k \neq j \end{cases}, k = 1, \dots, K,$ если $D_{ij} = 0.$
6	Проверить условие $\|U - U^*\|^2 \leq \varepsilon$,где U^*- матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 7, иначе - к шагу 3.
7	Конец.

Цель состоит в том, чтобы найти такие положения для центров кластеров, так чтобы расстояние между каждым объектом и связанным с ним центром кластера было минимальным.

2.2. Метод нечетких с-эллипсоидов

Нечеткие с-эллипсоиды представляют собой выпуклые комбинации точек и линий (или гиперплоскостей) [6].

Алгоритм кластеризации методом нечетких с-эллипсоидов отличается от метода нечетких с-средних нахождением расстояния от точек до центров кластеров. В с-эллипсоидов используется ковариационная матрица с собственными значениями и векторами. В таблице 2 показана схема алгоритма нечеткой кластеризации методом с-эллипсоидов.

Таблица 2. Алгоритм нечеткой кластеризации методом с- эллипсоидов

Шаг	Алгоритм нечетких с-эллипсоидов
1	Установить параметры алгоритма: C - количество кластеров; m - экспоненциальный вес; ε- параметр остановки алгоритма.
2	Случайным образом сгенерировать матрицу нечеткого разбиения $U = [u_{ij}]; u_{ij} \in [0,1];\ i \in 1, \dots, N; j \in 1, \dots, C.$
3	Рассчитать центры кластеров: $c_j = \frac{\sum_{i=1}^{N} (u_{ij})^m x_i}{\sum_{i=1}^{N} (u_{ij})^m}.$
4	Вычислить матрицу ковариации для j-ого кластера: $A_j = \frac{\sum_{i=1}^{N} (u_{ij})^m \cdot (x_i - c_j)^T \cdot (x_i - c_j)}{\sum_{i=1}^{N} (u_{ij})^m}.$
5	Рассчитать расстояния между объектами из $X = x_1, \dots, x_N$и центрами кластеров: $D_{ij} = \sqrt{\|x_i - c_j\|^2 - \alpha \cdot \sum_{s=1}^{r} [{S_{js}}^T (x_i - c_j)]^2}.$ где $\|x_i - c_j\|^2$- евклидово расстояние, $r -$ количество собственных векторов, S_{js} - s-ый собственный вектор ковариационной матрицы A_j кластера j. Параметр $\alpha = 1 - \frac{\lambda_2}{\lambda_1}$, где $\lambda_1, \lambda_2 - \ max$ и min собственное значение матрицы A_j.
6	Пересчитать элементы матрицы нечеткого разбиения: $u_{ij} = \frac{1}{\left({D_{ij}}^2 \sum_{k=1}^{C} \frac{1}{{D_{ik}}^2}\right)^{\frac{1}{m-1}}}$, если $D_{ij} > 0;$ $u_{ik} = \begin{cases} 1, k = j \\ 0, k \neq j \end{cases}, k = 1, \dots, K,$ если $D_{ij} = 0.$
7	Проверить условие $\|U - U^*\|^2 \leq \varepsilon$,где U^*- матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 8, иначе - к шагу 3.
8	Конец.

3. Кластеризации комплекснозначных данных

В таблица 3 приведен фрагмент исходной выборки данных.

Таблица 3. Фрагмент исходной выборки данных.

Переменные	X	Y
1	4.016+1.403i	0.454+1.493i
2	0.406+2.004i	4.296+0.878i
3	0.747+0.719i	2.465+0.476i
4	7.113+1.118i	4.867+0.371i
5	4.139+1.412i	4.94+0.46i
6	0.761+0.372i	4.828+2.622i
7	2.265+0.254i	7.532+1.08i
8	6.547+0.663i	6.11+0.391i
9	2.017+1.566i	2.244+0.174i
10	5.103+0.458i	7.784+0.657i
…		

Проведем кластеризацию исходных данных приведенными выше методами. Для наглядного примера кластеризации ниже представлены графики данных после кластеризации (рис. 3-5).

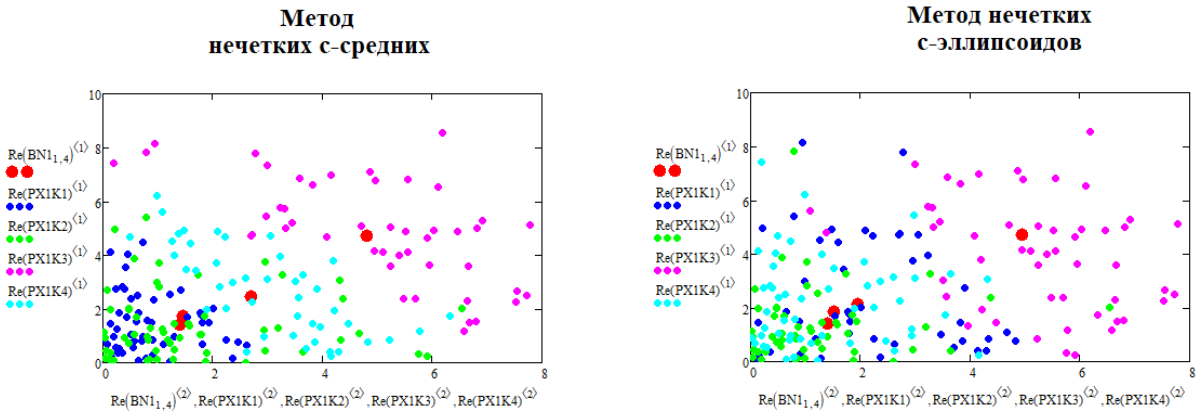

Рис. 3. Результаты кластеризации в проекции на действительные оси.

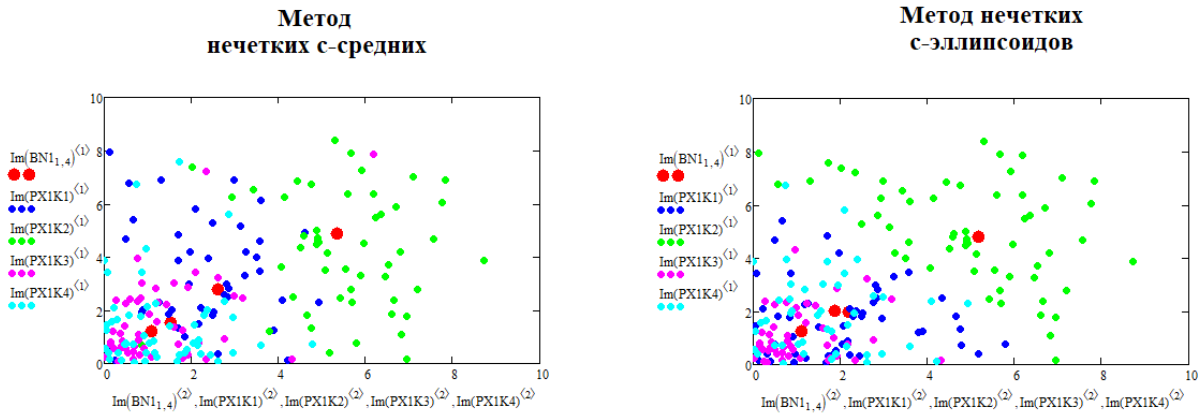

Рис. 4. Результаты кластеризации в проекции на мнимые оси.

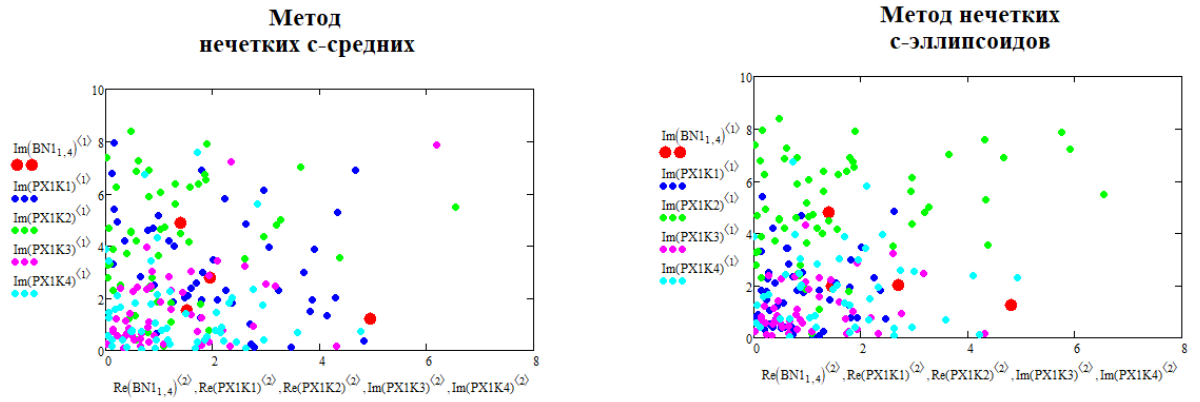

Рис. 5. Результаты кластеризации в проекции на действительную и мнимую ось.

Так как исходные данные задаются в пространстве C^2, то на графиках показаны проекции на оси, где PX1Kj – проекция для j-го кластера, а BN1 – центры кластеров (таблица 2).

Таблица 2. Центры кластеров.

	Нечетких с-средний		Нечеткий с-эллипсоидов	
Номер кластера	**X**	**Y**	**X**	**Y**
1	1.739+1.989i	1.457+2.189i	2.125+2.771i	1.945+2.605i
2	1.401+4.806i	1.4+5.161i	1.402+4.88i	1.4+5.355i
3	4.709+1.247i	4.81+1.104i	4.718+1.214i	4.936+1.087i
4	2.471+1.998i	2.696+1.861i	1.869+1.519i	1.524+1.532i

4. Заключение

В работе рассмотрены комплекснозначные данные, методы нечеткой кластеризации. На примере показана нечеткая кластеризация комплекснозначных данных. Графически представлены результаты кластеризации. Как видно из графиков и таблицы с центрами кластеров, результаты, полученные двумя рассмотренными методами, практически не отличаются друг от друга. Однако, по рис. 3-5 можно заметить, что сами кластеры различаются. Более однородные кластеры в данном примере дал метод нечетких с-средних. В дальнейшем планируется рассмотрение других нечётких методов кластеризации, в частности, Густафсона-Кесселя.

Список литературы

1. Баюк Д.А., Баюк О.А., Берзин Д.В., и др. Практическое применение методов кластеризации, классификации и аппроксимации на основе нейронных сетей. М.: Прометей, 2020. 448 с.
2. Бессмертный И.А., Нугуманова А.Б., Платонов А.В. Интеллектуальные системы. М.: Юрайт, 2023. 243 с. https://urait.ru/bcode/511999 (дата обращения: 24.05.2023).
3. Воронов М.В., Пименов В.И., Небаев И.А. Системы искусственного интеллекта. М.: Юрайт, 2023. 256 с. https://urait.ru/bcode/519916 (дата обращения: 24.05.2023).
4. Трегубов В.Н., Каткова М.А. Информационное пространство логистического кластера: теория и методология формирования на основе облачных технологий. М.: Юрайт, 2023. 495 с. https://urait.ru/bcode/530657 (дата обращения: 24.05.2023).
5. Arthur D., Vassilivitskii S. k-means++: the advantages of careful seedings // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007. P. 1027-1035.
6. Bradley P.S., Fayyad U.M. Refining initial points for k-means clustering // San Francisco: Proceedings of the Fifteenth International Conference of Machine Learning, 1998. 99 p.