РАЗМИТА КЛЪСТЕРИЗАЦИЯ НА ДАННИ
FUZZY CLUSTERING OF DATA

Цветанка Ковачева

Abstract. In the paper is considered the fuzzy c-means method, which is one of the most commonly used methods for multidimensional data analysis. Some different types of data using the square of the Euclidean metrics are clustered and the results are compared with those of hard c-means clustering. The basic mathematical characteristics of the resulting clusters are calculated. The results of clustering in case of border data and outliers from the cluster centers are discussed. Calculations are made in Matlab.

Key words: *cluster analysis, fuzzy clustering, membership degree*

1. Въведение

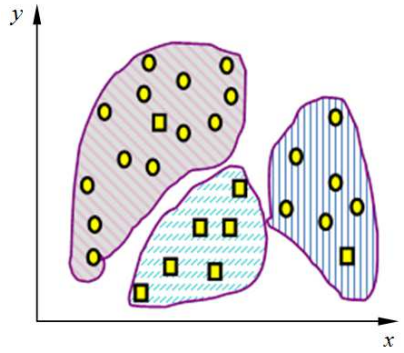
Клъстерният анализ (*cluster analysis*) е общ подход за статистически анализ на многомерни данни [3,5,6]. Неговата цел е да намери оптимално групиране на данните, при което елементите в даден клъстер са подобни. За измерване на близостта (сходството) се използват различни метрики: Евклидова, диагонална, Маханобисова и др.

В зависимост от това дали броят на клъстерите е предварително зададен се използват се използват *две основни клъстерни техники*:

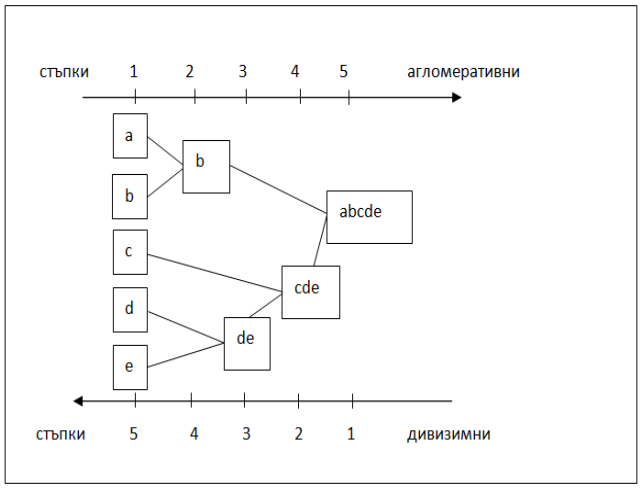
- *йерархична* – броят на клъстерите не е известен (Фиг.1):
  - *агломеративна* - bottom up (AGNES – AGglomerative NESting);
  - *дивизимна* - the top down (DIANA – DIvisive ANAlysis).
- *итерационна* – броят на клъстерите е зададен – метод на *c*-средните, метод на размитите *c*-средни и др. (Фиг.2).

При клъстеризацията на данни се използват следните *два методи*:

- *конвенционални (неразмити, твърди)* – изходното множество от обекти се разделя на непресичащи се подмножества (клъстери) (Фиг.3). Всеки обект от множеството принадлежи само на един клъстер.
- *размити* – един и същ обект принадлежи едновременно на няколко (или на всички) клъстери, но с различна степен (Фиг.4).

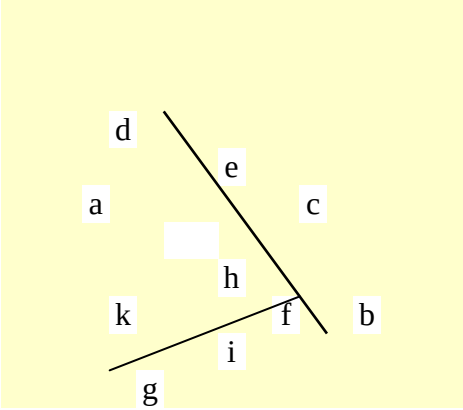


Фиг.1. Йерархичен подход

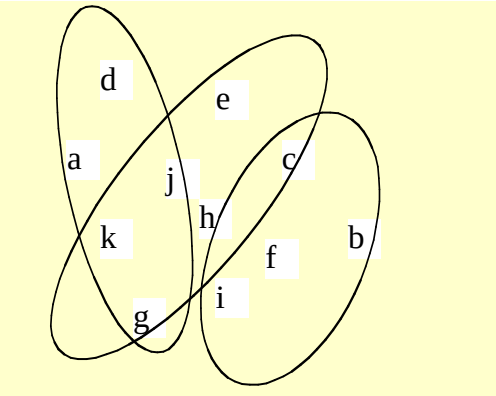


Фиг.2. Итерационен подход

В много случаи размитата клъстеризация се явява по-естествена от твърдата, например за обекти, които се явяват на границата на клъстерите.



Фиг.3. Неразмити клъстери



Фиг.4. Размити клъстери

В статията се разглежда метода на размитите c-средни, като разширение на метода на c-средните с цел намиране на оптимално разделяне на множество S, в случаите когато има данни на границата на два или повече клъстери, а и такива които са твърде отдалечени от центровете им и тяхното влияние върху характеристиките на клъстерите. Това дава възможност да се видят някои от предимствата на размитите методи пред конвенционалните.

Разгледани са два примера – единия със структурирани данни, а другия е с нормално разпределени генерирани данни. Данните са клъстеризирани по метода на c-средни и размити c-средни. Получените резултати са представени графично. Изчислени са основните характеристики на клъстерите.

2. Метод на c-средни и размити c-средни [1,2,4,7,8]

2.1. Входно множество от данни

Нека множеството $S=(s_1,s_2,...,s_i,...,s_m)$ съдържа m данни (обекти, наблюдения). Всеки обект се характеризира с n признаци (свойства)

$$s_i=(a_{i1},a_{i2},...,a_{ij},...,s_{in}) \qquad (1)$$

където - $a_{ij}\in\Re$, $i=\overline{1,m}$.

Данните могат да се представят като точки в $\Re^n$.

Множеството S се описва със следната матрица на данните

$$X=(a_{ij})_{m,n} \qquad (2)$$

в редовете на която са описанията на обектите, а в стълбовете – на техните признаци.

2.2. Твърди и размити разделяния

Твърдо разделяне на множеството S на c непресичащи се подмножества (клъстери) $\pi_1,\pi_2,...,\pi_c$, за което

$$\bigcup_{i=1}^{c}\pi_i=S;\ \pi_i\bigcap\pi_j=\varnothing,\ i\neq j;$$

$$\varnothing\subset\pi_i\subset S,\ i=\overline{1,c} \qquad (3)$$

Разделянето $\Pi(S)=\{\pi_1,\pi_2,...,\pi_c\}$ на множеството S е удобно да се представи с матрицата

$$U=(\mu_{ik})_{c,m} \qquad (4)$$

където – μ_{ik} - степен на принадлежност на k-тия обект към i-тия клъстер.

Матрицата U има следните свойства:

$$\mu_{ik}\in\{0,1\},\ i=\overline{1,c},\ k=\overline{1,m} \qquad (5)$$

$$\sum_{i=1}^{c}\mu_{ik}=1,\ k=\overline{1,m} \qquad (6)$$

$$0<\sum_{k=1}^{m}\mu_{ik}<m,\ i=\overline{1,c} \qquad (7)$$

Пространството от всички възможни матрици на разделяне за множеството S се дефинира по следния начин

$$M_{hc}=\left\{\begin{array}{l}U\in\Re^{c\times m}\,\Big|\,\mu_{ik}\in\{0,1\},\ \forall i,k;\ \sum_{i=1}^{c}\mu_{ik}=1,\ \forall k;\\ 0<\sum_{k=1}^{m}\mu_{ik}<m,\ \forall i\end{array}\right\} \qquad (8)$$

При *размитото разделяне* на множеството S матрицата (4) притежава следните свойства:

$$\mu_{ik}\in[0,1],\ i=\overline{1,c},\ k=\overline{1,m} \qquad (9)$$

$$\sum_{i=1}^{c}\mu_{ik}=1,\ k=\overline{1,m} \qquad (10)$$

$$0 < \sum_{k=1}^{m} \mu_{ik} < m,\ i = \overline{1,c} \qquad (11)$$

От (9) се вижда, че при размитото разделяне степента на принадлежност на всеки обект към даден размит клъстер е число между 0 и 1, а не само 0 или 1, както при твърдото разделяне (5).

От уравненията (6) и (10) следва, че общата степен на принадлежност на всеки обект S_i от S е равна на 1.

Пространството от всички възможни матрици на размитото разделяне за множеството S се дефинира по следния начин

$$M_{fc} = \left\{ \begin{array}{l} U \in \Re^{c \times m} \left| \mu_{ik} \in [0,1],\ \forall i,k;\ \sum_{i=1}^{c} \mu_{ik} = 1,\ \forall k; \right. \\ 0 < \sum_{k=1}^{m} \mu_{ik} < m,\ \forall i \end{array} \right\} \qquad (12)$$

2.3. Метод на c-средните

При метода на c-средните (k-means, hard c-means, crisp) предварително се определя предварително броя клъстерите c. Отчита се разстоянието на всяка точка на множеството S до избраните центрове на клъстерите, като най-близкото разстояние определя принадлежността на точката към съответния клъстер. Следователно за елементите на двоичната матрица $U \in M_{hc}$ (4) е изпълнено

$$\mu_{ik} = \begin{cases} 1, & \|s_k - v_i\|^2 \le \|s_k - v_j\|^2,\ \forall j \ne i \\ 0, & \textit{в противния случай} \end{cases} \qquad (13)$$

Целевата функция има вида

$$F_h = \arg\min_{\Pi(S)} \sum_{i=1}^{c} \sum_{s_k \in \pi_i} \|s_k - v_i\|^2 \qquad (14)$$

където – $\|.\|^2$ - квадрат на Евклидовата норма;

- $v = (v_1\ v_2\ \ldots\ v_i\ \ldots\ v_c) \in M_{hc}$, $v_i \in \Re^n$ - вектор на клъстерните центрове;

- $v_i = \dfrac{\sum_{s_k \in \pi_i} s_k}{|\pi_i|}$ (15)

- $|\pi_i|$ - брой на обектите на $i-$тия клъстер.

Клъстеризацията на данните се свежда до решаване на следната *оптимизационна задача*: търси се матрица $U \in M_{hc}$, минимизираща стойността на критерия (14). Дискретният характер на твърдото разделяне води до трудности при намиране на оптимална клъстеризация, поради негладкостта на целевата функция.

Алгоритъмът на c-средните е следния:

1 стъпка – определяне на основните параметри – избор на желания брой клъстери c, задаване на максимален брой на итерациите N, задаване на параметър $\varepsilon > 0$ за прекратяване на алгоритъма.

2 стъпка – инициализация – избор на клъстерни центрове v_i, $i = \overline{1,c}$.

3 стъпка – класификация – изследване на всяка точка от зададеното множество S и причисляването му към клъстер, чиито център е най-близко до нея.

4 стъпка – изчисляване на нови клъстерни центрове v_i^*, $i = \overline{1,c}$ по формула (15).

5 стъпка – критерии за сходимост – при достигане на максималния брой на итерациите N или изпълнение на условието $\|v - v^*\| < \varepsilon$ при две последователни стъпки на итерацията - изход от алгоритъма, в противен случай $v = v^*$ и се повтаря стъпка 3.

2.4. Метод на размитите c-средни

Методът на размитите c-средните (fuzzy c-means) е един от най-често използваните клъстерни методи и е въведен от Бездек през 1981, като разширение на метода на c-средните. След предварително определяне на броя на клъстерите c, входното множество от данни S се разделя на c размити клъстери. Данните се причисляват към клъстерите със съответна степен на принадлежност изменяща се в интервала $[0,1]$. Целевата функция има вида

$$F_f = \arg\min_{v} \sum_{i=1}^{c} \sum_{k=1}^{m} \mu_{ik}^{w} \|s_k - v_i\|^2 \qquad (16)$$

където - $v = (v_1\ v_2\ \ldots\ v_i\ \ldots\ v_c) \in M_{fc}$, $v_i \in \Re^n$ - вектор на клъстерните центрове;

- $w \in \Re$, $w \ge 1$ - експоненциално тегло, което определя степента на размитост на получените клъстери и влияе на стойностите на степените на принадлежност.

От необходимото условие за екстремум на функцията F_f частните й производни относно μ_{ki} и v_i са равни на нула. От получените уравнения и условие (10) се изчисляват степените на принадлежност и центровете на клъстерите

$$\mu_{ik} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|s_i - v_j\|}{\|s_i - v_k\|}\right)^{\frac{2}{w-1}}} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{w-1}}} \quad (17)$$

$$v_i = \frac{\sum_{k=1}^{m}\mu_{ik}^{w} s_k}{\sum_{k=1}^{m}\mu_{ik}^{w}} \quad (18)$$

Намирането на матрица $U \in M_{fc}$ с минимална стойност на критерия (16) е *задача на нелинейната оптимизация*. Най-често използван метод за нейното решаване е разглеждания метод на размитите c-средните, в основата на който е метода на неопределените множители на Лагранж. Той позволява да се намери локален оптимум, затова изпълнението на алгоритъма от различни начални точки може да доведе до различни резултати.

Алгоритъмът на размитите c-средни е следния:

Стъпка 1. Определяне на основните параметри: $0 < c < m$ - брой на клъстерите, $w > 1$ - експоненциално тегло, задаване на максимален брой на итерациите N, $\varepsilon > 0$ – параметър за прекратяване на алгоритъма.

Стъпка 2. Инициализация - съставяне по случаен начин на матрицата U, която да удовлетворява условията (9) – (11).

Стъпка 3. Пресмятат се размитите центрове v_i, $i = \overline{1,c}$ на клъстерите по формула (18).

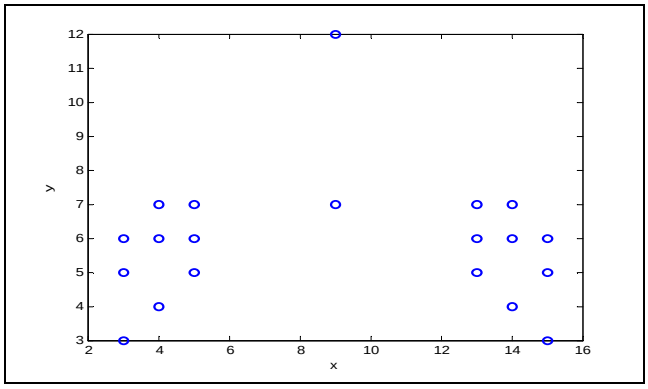
Стъпка 4. Преизчисляват се елементите на случайно избраната матрица на размито разделяне $U^* \in M_{fc}$ по формула (17).

Стъпка 5. Критерии за сходимост - при достигане на максималния брой на итерациите N или изпълнение на условието $\|U - U^*\| < \varepsilon$ при две последователни стъпки на итерацията - изход от алгоритъма, в противен случай $U = U^*$ и се повтаря отново стъпка 3.

3. Експериментален анализ

Разглеждат се два примера, в които при изчисленията с Matlab се приема, че максималният брой на итерациите е 100 и критерия за спиране на итерационния процес е $\varepsilon = 0{,}01$. При клъстеризиране на данните се използва квадрат на Евклидовата метрика.

Пример 1. Разглежда се множество от 20 структурирани данни в $\Re^2$: $S = \{$ (3,3), (4,4), (5,5), (3,5), (3,6), (4,6), (5,6), (4,7), (5,7), (9,7), (9,12), (15,3), (14,4), (13,5), (15,5), (13,6), (14,6), (15,6), (13,7), (14,7) $\}$ (Фиг.5). От фигурата се вижда, че данните може да се групират в 2 клъстера. Точките (9,7) и (9,12) са избрани така, че да са равно отдалечени от центровете им.

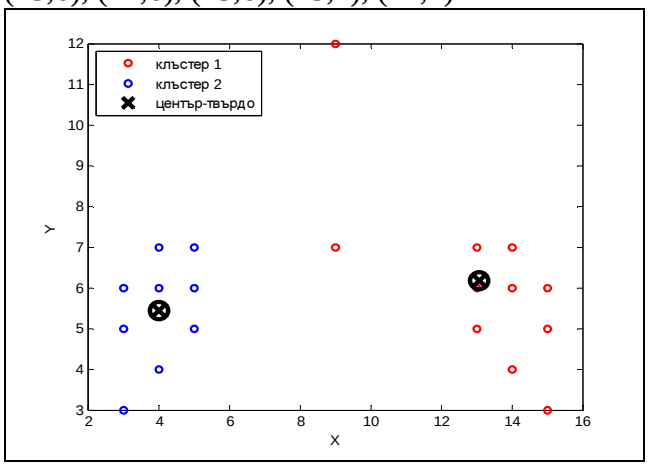


Фиг.5. Входни данни

По метод на c-средните данните се разделят на следните два клъстера (Фиг.6):

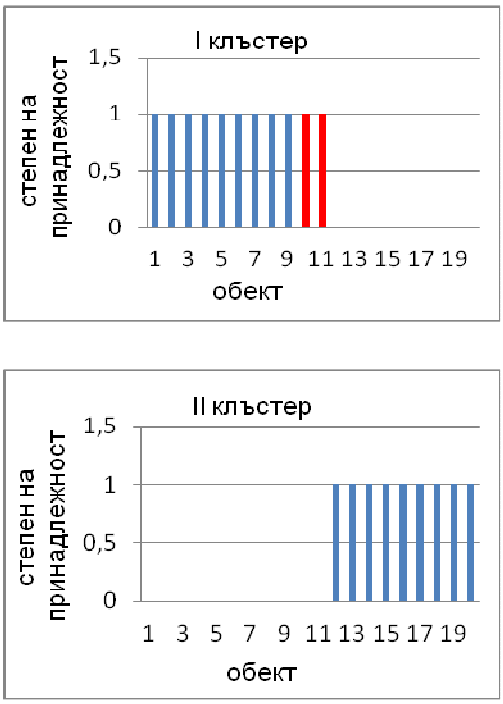
I: (3,3), (4,4), (5,5), (3,5), (3,6), (4,6), (5,6), (4,7), (5,7), (9,7), (9,12)

II: (9,12), (15,3), (14,4), (13,5), (15,5), (13,6), (14,6), (15,6), (13,7), (14,7)



Фиг.6. Клъстеризирани данни

Центровете на клъстерите са определени по формула (15). Степените на принадлежност на данните към съответния клъстер са представени графично на Фиг.7.



Фиг. 7. Изменение на степените на принадлежност на твърдите клъстери

Направена е клъстеризация на данните, като се изключат двете гранични точки (9,7) и (9,12) и само най-отдалечената точка (9,12). Получените основни характеристики на двата клъстери в трите случая са дадени в Табл.1.

Табл.1

N	*характеристика*	*I кл.*	*II кл.*
1	център c	(4.91,6.18)	(14,5.44)
2	дисперсия D	3,96	0,35
3	средно-квадратично отклонение s	1,99	0,59
4	радиус R	7,11	2,64

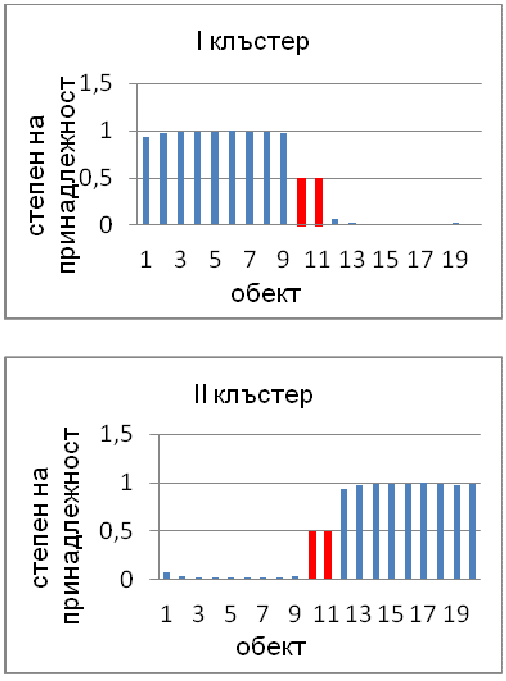
I кл. без точки (9,7), (9,12)	*I кл.* без точка (9,12)	*II кл.* без точка (9,12)
(4,5.44)	(4,5.44)	(13.5,5.6)
0,35	1.78	0.53
0,59	1.33	0.73
2,64	5.24	3.00

Граничните точки (9,7) и (9,12) се причисляват към първи клъстер и изместват неговия център. Наличието на отдалечената точка (9,12) води до значително увеличение на радиуса на първия клъстер. Ако се изключат точките (9,7) и (9,12), се изменя само първия клъстерния център и от симетрията на данните, характеристиките на клъстерите са еднакви. Изключването само на най-отдалечената точка (9,12) измества двата центъра, като граничната точка (9,7) се причислява към втория клъстер.

Точки, които се намират на разстояние по-голямо от $3s$ за даден клъстер, могат да не се причисляват към клъстера и да се разгледат като outliers. Такава точка е (9,12).

Размитата клъстеризация на c-средните се прилага за трите варианта на данните. При разделяне на два клъстери, степените на принадлежност на данните изменящи се в интервала [0,1] са представени графично на Фиг.8.

Двете гранични точки (9,7) и (9,12) се причисляват към първи клас и имат степени на принадлежност $0,5$. Това показва че правилно отразяват позицията си по средата между двата клъстера и могат да принадлежат към единия или другия клъстер. Аналогично точката (9,12) е също е на разстояние по-голямо от $3s$ и може да не се причислява към клъстера.



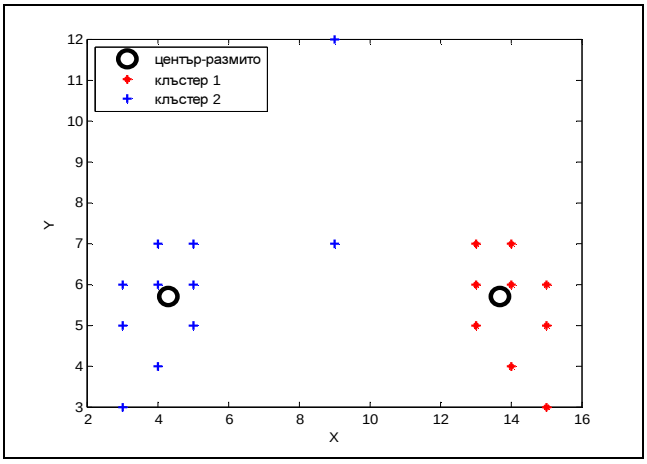
Фиг.8. Изменение на степените на принадлежност на размитите клъстери

Основните характеристики на получените два клъстера са дадени в Табл.2.

Табл.2

N	*характеристика*	*I кл.*	*II кл.*
1	център c	(4.29,5.69)	(13.71,5.69)
2	дисперсия D	1.02	0.44
3	средно-квадратично отклонение s	1.01	0.66
4	радиус R	3.94	2.78

I кл. без точки (9,7), (9,12)	*II кл.* без точки (9,7), (9,12)	*I кл.* без точка (9,12)	*II кл.* без точка (9,12)
(4.01,5.46)	(13.99,5.46)	(4.15,5.5)	(13.18,5.5)
0.25	0.25	0.27	0.24
0.5	0.5	0.52	0.49
2.31	2.31	2.38	2.39

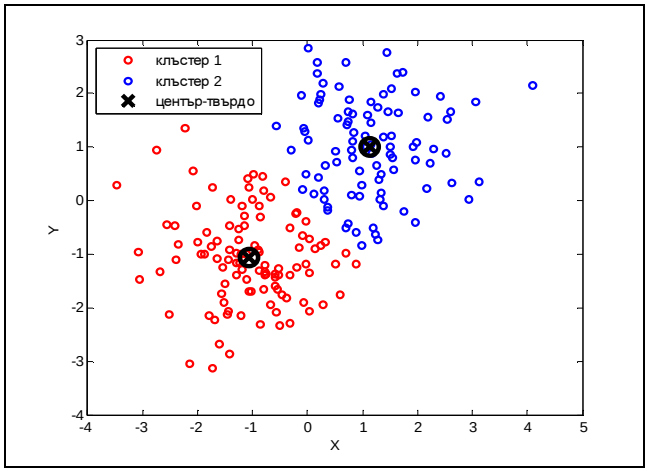


Фиг.9. Размита клъстеризация

От таблицата се вижда, че при размитата клъстеризация изместването на центровете на клъстерите, към които се причисляват граничните точки е по-малко, отколкото при твърдата клъстеризация при която те се причисляват само към даден клъстер.

Пример 2. Генерира се множество S състоящо се от 200 нормално разпределени случайни данни. Данните се клъстеризират по двата метода в два клъстера с използване на квадрат на Евклидовата метрика.

По метод на c-средните получените клъстери (Фиг.10) са с обеми $n_1 = 99$ и $n_2 = 101$ и характеристиките им са дадени в Табл.3. Клъстерните центрове са изчислени по формула (15). От първи клъстер има 13 точки, а от втори – 18 точки, които са на разстояние по-голямо от $3s$ за съответния клъстер и може да не се причисляват към тях.



Фиг.10. Твърдо клъстеризиране

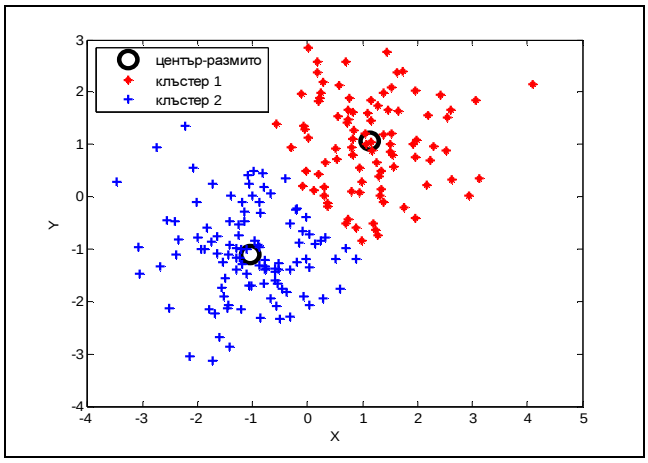
Табл.3

N	*характеристика*	*I кл.*	*II кл.*
1	център c	(-1.02,-1.03)	(0.99,1.14)
2	дисперсия D	0.38	0.36
3	средно-квадратично отклонение s	0.62	0.60
4	радиус R	3.20	2.63

Степените на принадлежност и клъстерните центрове са изчислени по формула (17) и (18). По метод на размитите c-средните данните се разделят на два клъстера (Фиг.11) с равни обеми, като характеристиките им са дадени в Табл.4.

Табл.4

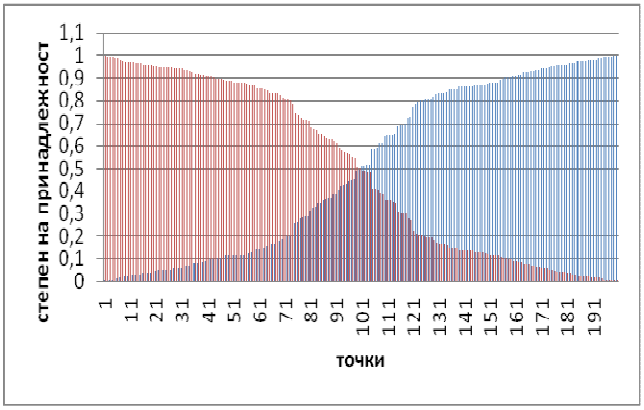
N	*характеристика*	*I кл.*	*II кл.*
1	център c	(-1.08,-0.99)	(1.04,1.13)
2	дисперсия D	0.18	0.15
3	средно-квадратично отклонение s	0.42	0.39
4	радиус R	1.86	1.64



Фиг.11. Размито клъстеризиране

Степените на принадлежност и клъстерните центрове са изчислени по формула (17) и (18). Степените на принадлежност на данните са представени графично на Фиг.12. Няма гранични данни, които да са точно на едно и също разстояние от двата центъра. От всеки клъстер има по 7 точки, които са на разстояние по-голямо от 3*s* за съответния клъстер и може да не се причисляват към тях.

Степените на принадлежност на данните са представени графично на Фиг.12. Няма гранични данни, които да са точно на едно и също разстояние от двата центъра. От всеки клъстер има по 7 точки, които са на разстояние по-голямо от 3*s* за съответния клъстер и може да не се причисляват към тях.



Фиг.12. Изменение на степените на принадлежност при размитите клъстери

4. Заключение

От анализа на направените изчисления на разгледаните множества от данни в статията следва, че при размитата клъстеризация наличието на гранични и твърде отдалечени данни изместват по-малко центъра на съответния клъстер, понеже имат по-малка степен на принадлежност. Това позволява по определен критерий по-лесно да се определят отдалечените точки.

За в бъдеще работата ще бъде насочена към определяне и изследване на гранични и твърде отдалечени данни.

Литература:

1. Bezdek J. *Fuzzy mathematics in pattern classification* Cornell university, Ithaca, 1973

2. Jain A., P. Murfy, P. Flynn *Data clustering: A Review*, ACM Computing Serveys, v.31,3,1999, 264 - 323

3. Кофман А. *Введение в теорию нечётких множеств*, М., Радио и связь, 1982

4. Kovacheva Ts. *Cluster analysis*, Information technologies and Control, Sofia, 3, 2004, 24-29

5. Мандель *Кластерный анализ*, Москва, Финансы и статистика, 1988

6. Hartigan J., M. Wong *A K-Means Clustering Algorithm*, Applied Statistics, 28, 1, 1979, 100-108

7. Höppner, F., F. Klawonn, R. Kruse, T. Runkler *Fuzzy Cluster Analysis*, John Wiley and Sons, 1999

За контакти:

9010 Варна, ул. “Студентска”1
Технически университет - Варна
секция “Математика”, ДМФЕО

гл.ас д-р Цветанка П. Ковачева
e-mail: tsveta_kovacheva@tu-varna.bg