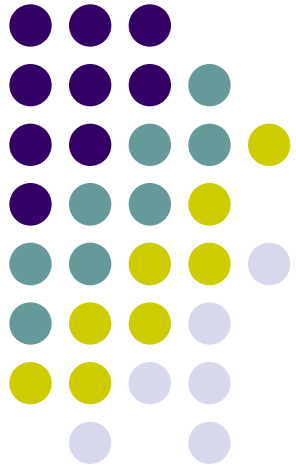
Кластеризация данных нечеткими методами

Выполнила:
ассистент кафедры ПМиИТ,
ЛГПУ
Волкова Елена

Научный руководитель:
д.ф.-м.н., профессор
Блюмин Семен Львович,
к.т.н.,доцент
Шуйкова Инесса Анатольевна



Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров. В метрическом пространстве «схожесть» обычно определяют через расстояние.

Алгоритмы кластеризации оперируют с объектами. Каждому объекту *X* отождествляется *вектор характеристик* $x=(x_1,...,x_d)$. Компоненты x_i, $i=1,..,d$ являются отдельными *характеристиками* объекта. Количество характеристик *d* определяет *размерность* пространства характеристик.

Множество, состоящее из всех векторов характеристик, обозначается $M=(X_1,...,X_n)$, где $X_i=(x_{i1},...,x_{id})$.

Кластер представляет собой подмножество «близких друг к другу» объектов из *M*. Расстояние $D(X_i, X_j)$ между объектами X_i и X_j определяется на основе выбранной метрики в пространстве характеристик.

Существует множество методов кластеризации, которые можно классифицировать на четкие и нечеткие.

Четкие методы кластеризации разбивают исходное множество объектов X на несколько непересекающихся подмножеств. При этом любой объект из X принадлежит только одному кластеру.

Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью принадлежности. Нечеткая кластеризация во многих ситуациях более "естественна", чем четкая, например, для объектов, расположенных на границе кластеров.

Четкая (непересекающаяся) кластеризация – кластеризация, в которой каждый объект X_i из M относится только к одному кластеру.

Нечеткая кластеризация – кластеризация, при которой для каждого X_i из M определяется μ_{ik} - вещественное значение, показывающее степень принадлежности x_i к кластеру y_j.

Развитие и широкое применение нечёткая кластеризация получила благодаря Бездеку и его методу нечетких k-средних (Fuzzy c-means - FCM).

Базовый алгоритм нечетких k-средних

Самый простой алгоритм нечеткой кластеризации – это нечеткий алгоритм k-средних. Этот алгоритм находит компактные кластеры, например, сферической формы.

Нечеткие кластеры опишем матрицей нечеткого разбиения $U = \left[u_{ij} \right]$, $u_{ij} \in [0,1]$ $i = 1,...,N$, $j = 1,...,K$, где i-я строчка содержит степени принадлежности объекта $x = (x_1, x_2, ..., x_N)$ кластерам $(U_1, U_2, ..., U_K)$. Единственным отличием матрицы степеней принадлежности четкого разбиения от нечеткого является то, что элементы матрицы принимают значения из двухэлементного множества $\{0,1\}$, а не из интервала $[0,1]$.

В базовом алгоритме нечетких k-средних расстояние между объектом $x = (x_1, x_2, ..., x_N)$ и центром кластера $c = (c_1, c_2, ..., c_K)$ рассчитывается через стандартную Евклидову норму:

$$D^2 = \|x - c\|^2$$

В результате алгоритмов кластеризации с фиксированной нормой форма всех кластеров получается одинаковой. Алгоритмы кластеризации как бы навязывают данным неприсущую им структуру, что приводит не только к неоптимальным, но иногда и к принципиально неправильным результатам. Для устранения этого недостатка предложено несколько методов, среди которых выделим алгоритм Густафсона-Кесселя (Gustafson-Kessel algorithm).

Алгоритм Густафсона-Кесселя(Gustafson – Kessel, GK)

По отношению к обычному алгоритму k-средних главное изменение состоит во введении в формулу расчета расстояния между векторами масштабирующей матрицы A. В качестве масштабирующей обычно применяется симметричная положительно определенная матрица, т.е. матрица, у которой все собственные значения являются действительными и положительными.

Алгоритм Густафсона-Кесселя использует адаптивную норму для каждого кластера, т.е. для каждого j-го кластера существует своя норм-порождающая матрица A_j. В этом алгоритме при кластеризации оптимизируются не только координаты центров кластеров и матрица нечеткого разбиения, но также и норм-порождающие матрицы для всех кластеров. Это позволяет выделять кластера различной геометрической формы.

GK – простой нечеткий алгоритм кластеризации, позволяющий обнаружить кластеры эллипсоидальной формы. В комбинации с алгоритмом нечетких k – средних он часто используется, чтобы инициализировать другие нечеткие алгоритмы кластеризации.

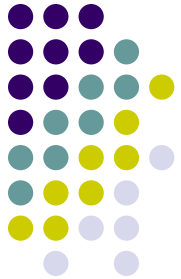
Базовый алгоритм нечетких k-средних	Алгоритм Густафсона-Кесселя
Шаг 1. Установить параметры алгоритма: c - количество кластеров; m - экспоненциальный вес; ε - параметр остановки алгоритма. *Шаг 2.* Случайным образом сгенерировать матрицу нечеткого разбиения.	
Шаг 3. Рассчитать центры кластеров: $c_j = \frac{\sum_{i=1}^{N} (u_{ij})^m \cdot x_i}{\sum_{i=1}^{N} (u_{ij})^m}$	*Шаг 3.* Рассчитать центры кластеров: $c_j = \frac{\sum_{i=1}^{N} (u_{ij})^m \cdot x_i}{\sum_{i=1}^{N} (u_{ij})^m}$

Базовый алгоритм нечетких k-средних	Алгоритм Густафсона-Кесселя
Шаг 4. Рассчитать расстояния между объектами из X и центрами кластеров: $$D_{ij} = \sqrt{\left\| x_i - c_j \right\|^2}$$	*Шаг 4.*Вычисляем матрицу ковариации для j-ого кластера: $$A_j = \frac{\sum_{i=1}^{N} (u_{ij})^m \cdot \left(x_i - c_j\right)^T \cdot \left(x_i - c_j\right)}{\sum_{i=1}^{N} (u_{ij})^m}$$ *Шаг 5.* Рассчитать расстояния между объектами из X и центрами кластеров: $$D_{A_j} = \left(x_i - c_j\right) \cdot \left[\left(\det\left(A_j\right)\right)^{\frac{1}{N}} \cdot A_j^{-1}\right] \cdot \left(x_i - c_j\right)^T$$

Базовый алгоритм нечетких k-средних	Алгоритм Густафсона-Кесселя
Шаг 5. Пересчитать элементы матрицы нечеткого разбиения: если $D_{ij}>0$ $$u_{ij}=\frac{1}{\left(D_{ij}^{2}\sum_{k=1}^{K}\frac{1}{D_{ik}^{2}}\right)^{\frac{1}{m-1}}}$$ если : $D_{ij}=0$ $$u_{ik}=\begin{cases}1, k=j\\0, k\neq j\end{cases}\quad k=\overline{1,K}$$	*Шаг 6.* Пересчитать элементы матрицы нечеткого разбиения: если $D_{A_j}>0$ $$u_{ij}=\frac{1}{\left(D_{ij}^{2}\sum_{k=1}^{K}\frac{1}{D_{ik}^{2}}\right)^{\frac{1}{m-1}}}$$ если : $D_{ij}=0$ $$u_{ik}=\begin{cases}1, k=j\\0, k\neq j\end{cases}\qquad k=\overline{1,K}$$

Базовый алгоритм нечетких k-средних	Алгоритм Густафсона-Кесселя
Шаг 6. Проверить условие $\left\|U - U^*\right\|^2 < \varepsilon$, где U^* - матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 7, иначе - к шагу 3. *Шаг 7*. Конец.	*Шаг 7.* Проверить условие $\left\|U - U^*\right\|^2 < \varepsilon$, где U^* - матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 8, иначе - к шагу 3. *Шаг 8.* Конец.

Алгоритм нечетких с-эллипсоидов



Шаг 1. Установить параметры алгоритма:

c - количество кластеров; m - экспоненциальный вес;

ε - параметр остановки алгоритма.

Шаг 2. Случайным образом сгенерировать матрицу нечеткого разбиения u_{ij} .

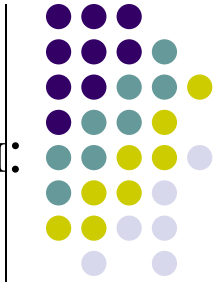
Шаг 3. Рассчитать центры кластеров: $c_j = \dfrac{\sum_{i=1}^{N} (u_{ij})^m \cdot x_i}{\sum_{i=1}^{N} (u_{ij})^m}$

Шаг 4. Рассчитать расстояния между объектами из X и центрами кластеров: $d^2\left(x_i, c_j\right) = \left\| x_i - c_j \right\|^2 - \alpha \cdot \sum_{s=1}^{r} \left[S_{js}^T \left(x_i - c_j \right) \right]^2$

где $\left\| x_i - c_j \right\|^2$ евклидово расстояние s=1,..,r, r-количество собственных векторов, S_{js} - s-ый собственный вектор ковариационной матрицы кластера j. A_j

Параметр $\alpha = 1 - \dfrac{\lambda_2}{\lambda_1}$, где λ_1, λ_2 max и min собственное значение матрицы A_j

Алгоритм нечетких с-эллипсоидов



Шаг 5. Пересчитать элементы матрицы нечеткого разбиения:

если $D_{ij}>0$ $$u_{ij}=\frac{1}{\left(D_{ij}^{2}\sum\limits_{k=1}^{K}\frac{1}{D_{ik}^{2}}\right)^{\frac{1}{m-1}}}$$

если : $D_{ij}=0$ $$u_{ik}=\begin{cases}1, k=j\\0, k\neq j\end{cases} \quad k=\overline{1,K}$$

Шаг 6. Проверить условие $\left\|U-U^{*}\right\|^{2}<\varepsilon$, где $U*$ матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 7, иначе - к шагу 3.

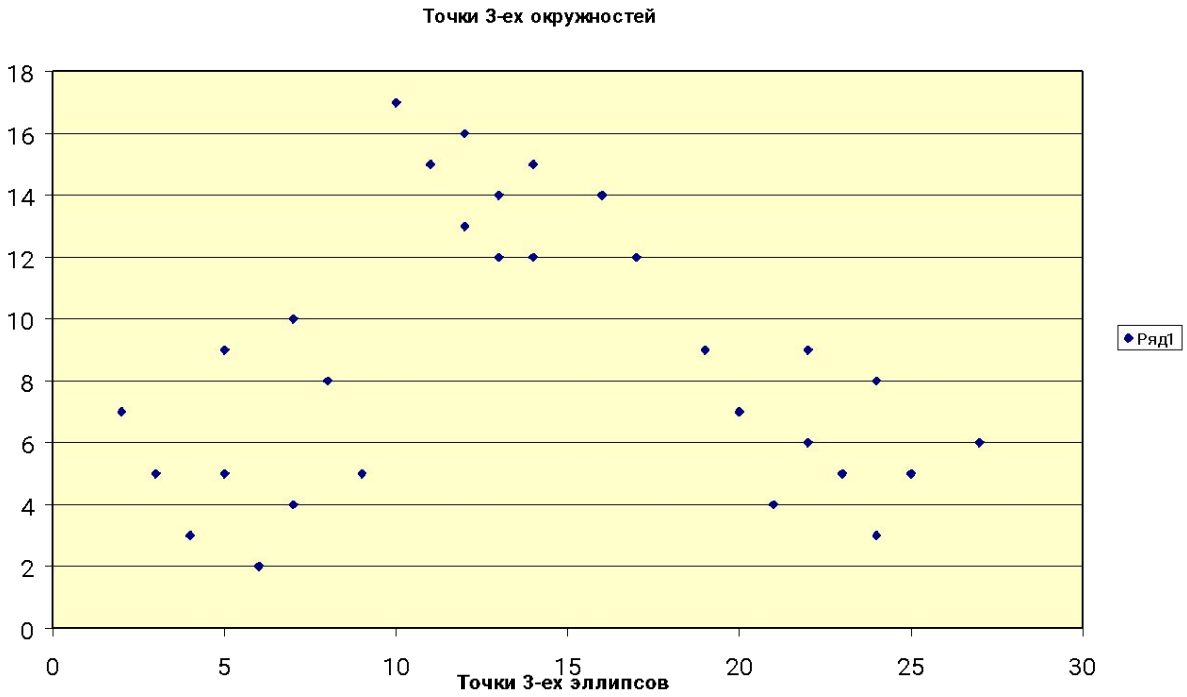
Шаг 7. Конец.

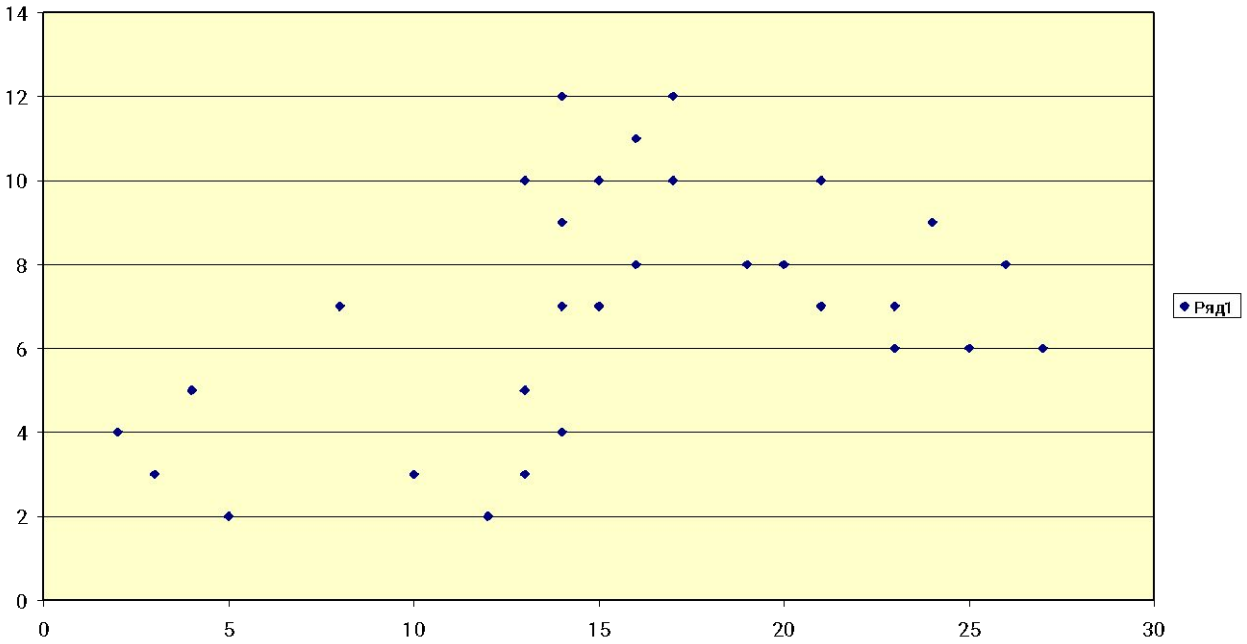
Shell-clustering algorithm

Основное новшество алгоритма состоит в том, что в данном алгоритме прототип кластера описывается помимо центра ещё и радиусом r_j .

Формула для вычисления расстояния имеет следующий вид:

$$d^2\left(x_i,\left(c_j,r_j\right)\right)=\left(\left\|x_i-c_j\right\|-r_j\right)^2$$


Точки 3-ех окружностей
18
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
Ряд1


Точки 3-ех эллипсов
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30
Ряд1

Результаты для окружностей

FCM algorithm

№	Clus1	Clus 2	Clus 3
1	0.072	0.877	0.049
2	0.057	0.908	0.034
3	0.077	0.879	0.043
4	0.024	0.961	0.014
5	0.049	0.919	0.031
6	0.092	0.839	0.069
7	0.094	0.848	0.057
8	0.066	0.906	0.027
9	0.057	0.917	0.026
10	0.135	0.816	0.049
11	0.881	0.065	0.055
12	0.957	0.019	0.024
13	0.794	0.126	0.080
14	0.929	0.023	0.048
15	0.858	0.075	0.067
16	0.861	0.063	0.075
17	0.729	0.153	0.117
18	0.604	0.254	0.142
19	0.689	0.197	0.113
20	0.919	0.018	0.064

GK-algorithm

№	Clus 1	Clus 2	Clus 3
1	0.066	0.906	0.027
2	0.078	0.878	0.043
3	0.058	0.916	0.025
4	0.025	0.965	0.008
5	0.069	0.899	0.031
6	0.094	0.848	0.057
7	0.135	0.816	0.048
8	0.451	0.473	0.07
9	0.196	0.689	0.114
10	0.512	0.423	0.064
11	0.907	0.066	0.026
12	0.908	0.062	0.030
13	0.919	0.052	0.028
14	0.603	0.254	0.142
15	0.975	0.017	0.007
16	0.940	0.045	0.013
17	0.932	0.050	0.016
18	0.907	0.064	0.028
19	0.948	0.037	0.014
20	0.638	0.312	0.049

C-elliptotypes

№	Clus 1	Clus 2	Clus 3
1	0.076	0.881	0.043
2	0.061	0.909	0.029
3	0.087	0.878	0.036
4	0.025	0.963	0.012
5	0.050	0.924	0.027
6	0.093	0.848	0.059
7	0.093	0.861	0.047
8	0.073	0.906	0.021
9	0.067	0.913	0.020
10	0.170	0.794	0.037
11	0.940	0.038	0.023
12	0.943	0.031	0.026
13	0.949	0.035	0.016
14	0.910	0.036	0.053
15	0.996	0.003	0.002
16	0.975	0.013	0.012
17	0.910	0.057	0.034
18	0.783	0.150	0.067
19	0.889	0.077	0.034
20	0.751	0.075	0.173

Shell-clustering

№	Clus 1	Clus 2	Clus 3
1	0.003	0.997	0.000
2	0.002	0.997	0.000
3	0.009	0.990	0.001
4	0.000	1.000	0.000
5	0.001	0.999	0.000
6	0.008	0.990	0.002
7	0.020	0.977	0.004
8	0.029	0.970	0.001
9	0.022	0.977	0.001
10	0.164	0.831	0.005
11	0.996	0.003	0.001
12	0.996	0.002	0.002
13	0.999	0.000	0.000
14	0.986	0.004	0.010
15	1.000	0.000	0.000
16	1.000	0.000	0.000
17	0.998	0.001	0.001
18	0.985	0.011	0.004
19	0.997	0.003	0.001
20	0.902	0.011	0.088

Результаты для окружностей

FCM algorithm

№	Clus 1	Clus 2	Clus 3
21	0.240	0.033	0.726
22	0.026	0.006	0.966
23	0.090	0.039	0.869
24	0.025	0.008	0.965
25	0.068	0.026	0.905
26	0.130	0.063	0.805
27	0.114	0.046	0.838
28	0.159	0.064	0.775
29	0.085	0.024	0.890
30	0.067	0.014	0.918

GK-algorithm

№	Clus 1	Clus 2	Clus 3
21	0.109	0.182	0.708
22	0.080	0.078	0.840
23	0.097	0.088	0.814
24	0.127	0.115	0.757
25	0.071	0.090	0.837
26	0.042	0.049	0.908
27	0.014	0.015	0.970
28	0.041	0.049	0.909
29	0.047	0.057	0.895
30	0.047	0.041	0.911

C-elliptotypes

№	Clus 1	Clus 2	Clus 3
21	0.260	0.079	0.661
22	0.080	0.036	0.884
23	0.056	0.037	0.907
24	0.005	0.003	0.992
25	0.013	0.007	0.980
26	0.055	0.037	0.907
27	0.033	0.019	0.949
28	0.067	0.037	0.896
29	0.025	0.011	0.964
30	0.066	0.024	0.910

Shell-clustering

№	Clus 1	Clus 2	Clus 3
21	0.128	0.007	0.865
22	0.012	0.001	0.987
23	0.002	0.000	0.997
24	0.000	0.000	1.000
25	0.000	0.000	1.000
26	0.003	0.001	0.996
27	0.001	0.000	0.999
28	0.008	0.001	0.991
29	0.001	0.000	0.999
30	0.009	0.001	0.990

Результаты для эллипсов

FCM algorithm

№	Clus 1	Clus 2	Clus 3
1	0.194	0.154	0.650
2	0.166	0.128	0.704
3	0.183	0.143	0.672
4	0.208	0.168	0.622
5	0.266	0.136	0.596
6	0.152	0.078	0.769
7	0.042	0.028	0.929
8	0.110	0.071	0.818
9	0.139	0.082	0.777
10	0.136	0.098	0.765
11	0.494	0.200	0.304
12	0.672	0.194	0.132
13	0.818	0.153	0.028
14	0.517	0.280	0.201
15	0.427	0.291	0.281
16	0.520	0.364	0.115
17	0.402	0.561	0.036
18	0.441	0.472	0.085
19	0.376	0.542	0.081
20	0.580	0.304	0.115

GK-algorithm

№	Clus 1	Clus 2	Clus 3
1	0.016	0.057	0.925
2	0.030	0.062	0.907
3	0.028	0.052	0.919
4	0.009	0.020	0.970
5	0.007	0.017	0.975
6	0.013	0.045	0.940
7	0.016	0.050	0.932
8	0.028	0.064	0.907
9	0.014	0.037	0.948
10	0.019	0.042	0.938
11	0.689	0.197	0.113
12	0.729	0.153	0.116
13	0.929	0.023	0.047
14	0.603	0.254	0.142
15	0.654	0.248	0.097
16	0.618	0.262	0.119
17	0.635	0.218	0.146
18	0.662	0.233	0.104
19	0.621	0.244	0.134
20	0.717	0.189	0.093

C-elliptotypes

№	Clus 1	Clus 2	Clus 3
1	0.022	0.003	0.975
2	0.005	0.001	0.995
3	0.013	0.001	0.985
4	0.036	0.005	0.959
5	0.606	0.070	0.325
6	0.658	0.023	0.320
7	0.010	0.001	0.989
8	0.096	0.018	0.886
9	0.301	0.045	0.654
10	0.047	0.002	0.951
11	0.979	0.004	0.017
12	0.984	0.006	0.010
13	0.998	0.001	0.001
14	1.000	0.000	0.000
15	0.993	0.001	0.006
16	1.000	0.000	0.000
17	0.988	0.010	0.002
18	0.996	0.003	0.001
19	0.962	0.033	0.005
20	0.986	0.007	0.008

Shell-clustering

№	Clus 1	Clus 2	Clus 3
1	0.257	0.114	0.628
2	0.218	0.096	0.687
3	0.260	0.101	0.639
4	0.278	0.124	0.599
5	0.227	0.076	0.697
6	0.157	0.034	0.810
7	0.058	0.020	0.923
8	0.106	0.045	0.848
9	0.122	0.046	0.832
10	0.243	0.058	0.698
11	0.732	0.050	0.218
12	0.771	0.071	0.158
13	0.839	0.083	0.078
14	0.995	0.001	0.003
15	0.889	0.032	0.079
16	0.955	0.021	0.024
17	0.723	0.195	0.082
18	0.808	0.116	0.076
19	0.643	0.249	0.108
20	0.796	0.093	0.111

Результаты для эллипсов

FCM algorithm

№	Clus 1	Clus 2	Clus 3
21	0.123	0.870	0.006
22	0.208	0.771	0.019
23	0.252	0.705	0.041
24	0.331	0.619	0.049
25	0.351	0.565	0.082
26	0.372	0.532	0.095
27	0.336	0.574	0.089
28	0.371	0.506	0.121
29	0.359	0.520	0.120
30	0.371	0.483	0.144

GK-algorithm

№	Clus 1	Clus 2	Clus 3
21	0.057	0.908	0.034
22	0.077	0.879	0.043
23	0.024	0.961	0.014
24	0.048	0.919	0.031
25	0.092	0.839	0.068
26	0.094	0.848	0.057
27	0.066	0.906	0.027
28	0.057	0.916	0.025
29	0.135	0.816	0.048
30	0.072	0.877	0.049

C-elliptotypes

№	Clus 1	Clus 2	Clus 3
21	0.041	0.908	0.052
22	0.125	0.872	0.003
23	0.042	0.957	0.002
24	0.013	0.986	0.001
25	0.000	1.000	0.000
26	0.001	0.999	0.000
27	0.001	0.999	0.000
28	0.002	0.998	0.000
29	0.005	0.995	0.001
30	0.011	0.988	0.002

Shell-clustering

№	Clus 1	Clus 2	Clus 3
21	0.141	0.808	0.052
22	0.084	0.886	0.030
23	0.094	0.874	0.032
24	0.028	0.959	0.013
25	0.028	0.957	0.015
26	0.057	0.910	0.033
27	0.060	0.911	0.029
28	0.099	0.840	0.061
29	0.110	0.827	0.063
30	0.143	0.764	0.094

Библиографический список

1. Осовский С. Нейронные сети. М.: Финансы и статистика, 2002.
2. Сокал Р.Р. Кластер-анализ и классификация: предпосылки и основные направления [Текст]/ Р.Р.Сокал. Под ред. Дж. Вэн Райзина. – М.:Мир, Классификация и кластер, 1980.
3. Bezdek J.C. Pattern recognition with fuzzy objective function algorithms. – Plenum Press, New York. – 1982.
4. Gustafson D.E., Kessel W.C. Fuzzy clustering with a fuzzy covariance matrix - http://www.egr.uh.edu/ece/faculty/karayiannis/Karayiannis_tnn_16(2)_05.pdf
5. Jain A. K. Data Clustering: A Review [Текст] / A. K. Jain, M. N. Murty, P. J. Flynn - http://www.csee.umbc.edu/nicholas/clustering/p264-jain.pdf, 2006.
6. Ahmed Ismail Shihab**.** Fuzzy clustering algorithmes and their application to medical image analysis, PhD dissertation, 2000.

Спасибо за внимание!

Волкова Елена Петровна
Блюмин Семен Львович
Шуйкова Инесса Анатольевна
volkova.lenochka@mail.ru
sabl@lipetsk.ru
shujkova_i_a@inbox.ru