Object Fuzzy Clustering

Eva Ocelíková, Jana Výrostková

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Letná 9, 041 20 Košice, Slovak Republic
Eva.Ocelikova@tuke.sk, Jana.Vyrostkova@tuke.sk

Abstract: Data clustering is possible to describe in a simple way as the classification of the data into the groups and so the similar objects belong to the same group while dissimilar objects belong to various groups. This document deals with data clustering based on membership function (fuzzy clustering). Two main clustering algorithms are described as: fuzzy c-means and Gustafson-Kessel algorithm.

Keywords: clustering, fuzzy clustering, fuzzy c-means algorithm, Gustafson&Kessel algorithm

1 Introduction

Currently exist a lot of accesses to data classification to the groups, beginning with classic statistical methods and ending with accesses utilizing artificial intelligence means (neuron nets, expert systems, machine learning or fuzzy sets).

2 Cluster Analysis

Cluster analysis defined Tryon [5] in the 1939 by following way:

Cluster analysis is general logical process formulated as procedure, by means of which objective subjects are formed into groups based on their similarity and dissimilarity.

Data representation with smaller number of clusters loses precision of details, but simplification is reached this way. Clustering allows represents large set of object by several clusters and so it models the data by their clusters. From historic point of view modelling the data range the clustering into mathematics, statistics and numeric analysis.

In the classic clustering analysis each object belong to just one cluster, in case of fuzzy clustering belong the object to more cluster at the same time. Membership of the *k* object into *i* cluster is representing by his membership and can be gain of values from interval <0, 1>.

Generally **cluster** is characterized as set containing examined objects, that are each other similar and at the same time they are not similar to objects from others clusters.

Objects can create clusters of various geometrical shapes, sizes and densities. Success of individual clustering algorithms is dependent not only on geometric shape cluster and density of individual clusters, but also on their space relations and distances between them. Clusters can be well separable, each other interconnected together; eventually they can be recovered.

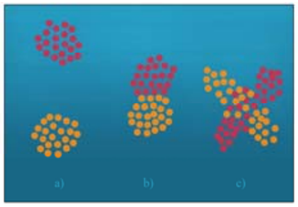


Figure 2.1
Cluster arrangement types: a) well separable, b) each other interconnected together, c) recovered

Parameter of the indefinites *m* is a weight exponent, which determine indefinites of final partitions at fuzziness clustering. Whereby is the parameter of values *m* bigger, thereby is indefinites of clustering bigger. Most frequently choosing value is *m = 2*.

2.1 Fuzzy C-means

Fuzzy C-Means algorithm tries to find the division of the data set, by minimizing the primary c-means functional.

$$J(\mathrm{X};\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^{m} d\left(x_k, \mathbf{v}_i\right)^2 \quad , \qquad (1)$$

where

u_{ik} membership degree of element x_i into *j* cluster,

v_k centre of *k* cluster,

m indefinites parameter (fuzziness), which is higher than 1,

$d\left(x_k, v_i\right)$ expresses the distance between the element $\mathbf{x}_i$ and the centre of the cluster *j*

(Euclidean metrics $d\ (\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$)

Entry data:

X	set of objects,
U	set of membership,
c	number of objects**,**
m	fuzziness (m > 1),
ε	tolerance

Fuzzy c – means algorithm consisting of the following steps:

1 Incidental initialization of set of membership

2 Compute centre of clusters by relation

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m . x_k}{\sum_{k=1}^{n}(u_{ik})^m} \tag{2}$$

3 Refresh of matrix with appropriateness U according to relation

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d(x_k, v_i)^2}{d(x_k, v_j)^2}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{3}$$

4 Algorithm finishes, if is difference between actual and previous step is smaller than choices tolerance ε. In the case, that deviation is bigger, algorithm repeat from 2 step, while condition (4) is accomplished

$$\left\| U^{(t)} - U^{(t-1)} \right\| < \varepsilon \tag{4}$$

2.2 Gustafson&Kessel Algorithm

Method Gustafson & Kessel is extension fuzzy c-means algorithm about adaptive norm, which allows get clusters of various shapes in one set of data. Each cluster Z_i is characterized by its standardization matrix A_i. Matrixes A_i are applied as optimization variables in c-means functional. Each cluster is able to adapt one’s own norm by topology data from the specific region. Objective function is definite as:

$$J\left(X;U,V,A_i\right)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^{m}d\left(x_k,v_i\right)^{2}_{ikA_i} \quad , \qquad (5)$$

where

u_{ik} membership degree of element x_i into j cluster,

m fuzziness (m > 1),

$d\left(x_k,v_i\right)$ distance between element $\boldsymbol{x_i}$ and centre of cluster i

Entry data***:***

X set of objects,

V set of clusters,

U set of membership,

c number of objects**,**

m fuzziness (m > 1),

ε tolerance.

Gustafson&Kessel algorithm consists of the following steps:

1 Incidental initialization of set of membership

2 Compute matrix of covariance F by relation

$$Fi=\frac{\sum_{k=1}^{n}\left(u_{ik}\right)^{m}.\left(x_k-v_i\right)\left(x_k-v_i\right)^{T}}{\sum_{k=1}^{n}\left(u_{ik}\right)^{m}} \qquad (6)$$

3 Updating of set of membership U by relation

$$u_{ik}=\frac{1}{\sum_{j=1}^{c}\left(D_{ikAi}/D_{jkAi}\right)^{\frac{2}{m-1}}} \quad , \qquad (7)$$

where

$$d(x_k, v_i)^2 = (x_k - v_i)^T . A_i . (x_k - v_i) = (x_k - v_i)^T \left[det(F_i)\frac{1}{m} F_i^{-1} \right] (x_k - v_i) \tag{8}$$

and standardization matrix A_i by relation

$$A_i = \left[det(F_i)\frac{1}{m} F_i^{-1} \right] \tag{9}$$

4 Iteration repeats of step 2. Clustering finish when is executed finishing condition (4).

3 Experimental Part

On the artificially created data we are apply algorithms FCMeans and G&K. The data are generated that, in order to created clumps will be uniform. At both of kind of algorithms FCMeans and G&K, where we start random generated functions of object of membership, was for mutual comparison applied equivalent values.

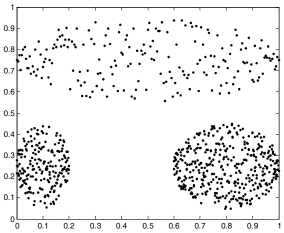


Figure 3.1
The set of artificially generation date

3.1 Fuzzy C – means

This algorithm has been realized for different values of the fuzziness *m* and by the various finishing conditions ε. The best result was acquire by $m = 2 \; a \; \varepsilon = 0.1$.

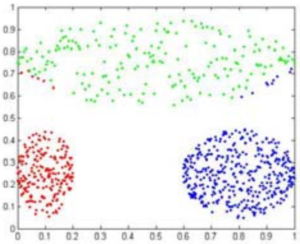
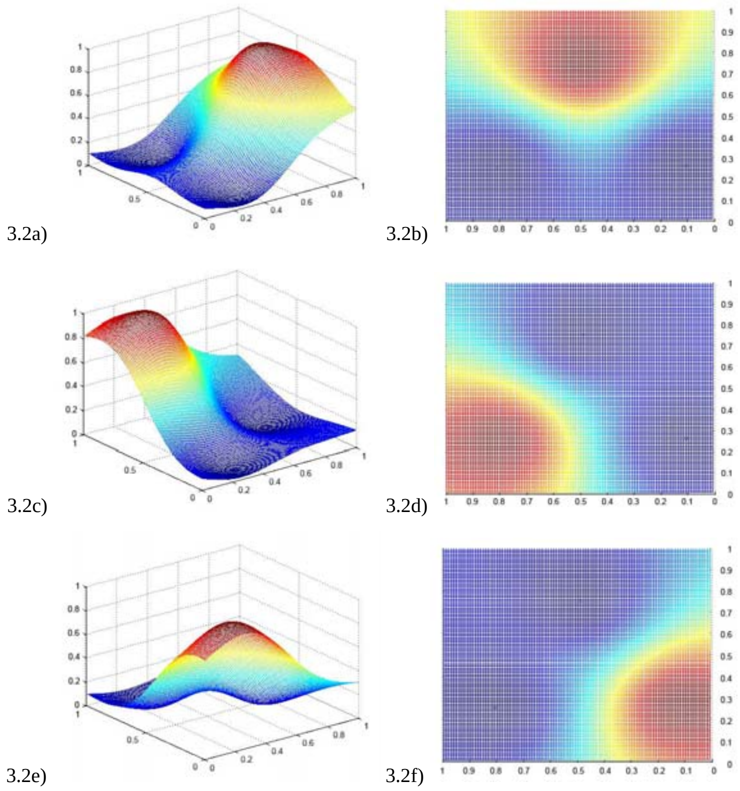


Figure 3.2
The clusters acquired by fuzzy c - means algorithm for $m = 2 \; a \; \varepsilon = 0.1$


3.2a) 3.2b)

3.2c) 3.2d)

3.2e) 3.2f)

Figure 3.2 a) c) e)

The shape of the function of reference related to referred clumps

Figure 3.2 b) d) f)

The depiction concrete function of membership into plaints xy

3.2 G&K Algorithm

This method has been realized also for different value of the fuzziness *m* and at various finishing conditions ε. Analogous to method fuzzy c – means by method G&K also was acruire best result by the option $m = 2\ a\ \varepsilon = 0.1$

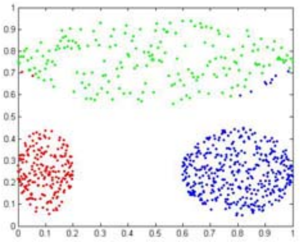
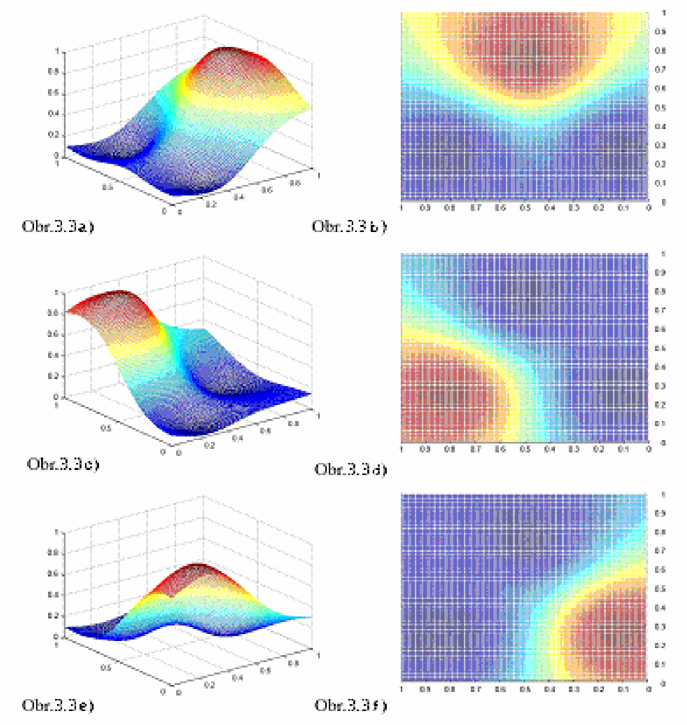


Figure 3.3

The clusters acquired by algorithm Gustafson&Kessel for $m = 2\ a\ \varepsilon = 0.1$



Obr.3.3a)

Obr.3.3 b)

Obr.3.3c)

Obr.3.3 d)

Obr.3.3 e)

Obr.3.3 f)

Figure 3.3 a) c) e)

The shape of the function of reference related to referred clumps

Figure 3.3 b) d) f)

The depiction concrete function of membership into plaints xy

The both methods fuzzy c - means and Gustafson &Kessel devide the set of artificial generation data always into three clusters and also by different values of parameter of the indefinites and tolerance. The best results was achieve for

$m = 2 \; a \; \varepsilon = 0.1$ by both methods on Figure 3.2 and Figure 3.3. By choosing higher values of parameter *m* is absolute success of clustering relatively low and the classification of object into given cluster is vague.

Conclusions

Elaboration of information is an activity, which is in the spotlight. Currently appear more and more applications, which on the strength of hidden coherence in observated data knows extraordinarily help to decisioner do right decision. This document have showen application of universal algorithm, which have got high level of classification successfulness. The results of experiment have showen, that the listed algorithms are excellent doing in same data type, however exist cases, in which the acquired results aren’t so good. To view from this special cases, these algorithms have a big applications potencial on many fields of a human activities.

References

[1] Aha, D. W. - Kibler D.- M. K. Albert M. K.: Instance-Based Learning Algorithms. Machine Learning, vol. 6, pp. 37-66, 1991

[2] Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function., Plenum Press, New York, 1981

[3] Baraldi, A., Blonda. P.: A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. ICSI, TR-98-038, 1998

[4] Hual, L., Hiroshi, M.: Extraction, Construction and Selection. Kluwer Academic Publishers, Boston, 1998

[5] Lukasová, A., Šarmanová, J.: Metody Shlukové Analýzy, SNTL, Praha, 1985

[6] Ocelíková, E.: Multikriteriálne Rozhodovanie, ELFA, Košice, 2002

[7] Tryon, R.C.: Cluster Analysis. Ann Arbor, Edwards Bros, 1939