

Fuzzy Clustering of the Earth Screen Data

Jana Výrostková, Eva Ocelíková, *Dana Klimešová

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

*Czech University of Agriculture, Prague
Faculty of Economics and Management, Dept. of Information Engineering
Kamýcká 129, 165 21 Praha 6 – Suchbátka, Czech Republic
klimesova@pef.czu.cz

Abstract: Nowadays there are a lot of approaches used for referring the classification of the data into groups, starting with classical statistics methods and closing with the approaches using means of an artificial intelligence. The report is dealing with a fuzzy clustering of the Earth screen data following the function of its reference (fuzzy clustering). There were described and on a real data experimentally proved two of basic clustering algorithms there: fuzzy c-means and the Gustafson-Kessel algorithm.

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

1 Introduction

Clustering is to be one of the solutions of the case of a learning beyond control, where no class labeling information on the data is available. Clustering is a method of dividing the data into groups (the clusters), ‘seemingly’ making a sense. Clustering algorithms are usually a high-speed and quite simple. It does not need any in advance knowledge on the data and the form having been used, and makes a solution by comparing the given samples to each other and to the *clustering criterion*, as well.

The simplicity of the algorithms can also be a disadvantage: the results may vary greatly using a different kind of clustering criteria and therefore unfortunately, also nonsense solutions are possible. Also use of some algorithms the order, the original samples are introduced at, can cause a great difference in the results.

Despite the disadvantages, clustering is used in many fields of the science, including machine vision, life and medical sciences and information sciences, as well. One of the reasons for that is to be the fact that intelligent beings, the human

beings included, are known for using the idea of clustering in areas of many brain functions.

2 Definition of a Cluster

Define some basic concepts of clusters in a mathematical way. Let X be a set of data, that is $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where every $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ is described by k features. So called m -clustering of X is its partition into m parts (clusters) Z_1, \dots, Z_m , so that

- 1 None of the clusters is empty; $Z_i \neq \emptyset$
- 2 Every sample belongs to a cluster
- 3 Every sample belongs to a single cluster (crisp clustering); $Z_i \cap Z_j = \emptyset, i \neq j$.

Of course, it is assumed that vectors in cluster Z_i are in some way “more similar” to each other than to the vectors in other clusters. Figure 1 illustrates couples of different types of clusters; a compact, a linear and a circular one.



Figure 1

Couple of different kind of clusters: a) compact, b) linear, c) circular

3 Fuzzy Clustering Algorithms

While there at the classical fuzzy analysis each of the data must be allocated to a just one cluster, the Fuzzy Cluster Analysis moderates the postulate enabling a gradual membership, i.e. it offers an opportunity to deal with data belonging to more than one cluster at the same time. Affiliation of a “ k ” subject to an “ i ” cluster u_{ik} can reach values of $\langle 0,1 \rangle$ interval. There are describe two of basic clustering algorithms: fuzzy c-means and the Gustafson-Kessel algorithm.

3.1 Fuzzy C-Means

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

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

where

- u_{ik} degree of relationship of x_i factor to a j -th cluster,
- v_k center of the k -th cluster,
- m indeterminacy parameter, where $m > 1$,
- $d(x_k, v_i)$ expressing the distance between an x_i factor and the center of the j -th cluster,
- X set of subjects,
- V set of clusters,
- U function reference matrix of the x subjects to partial “Z” clusters is to be generated accidentally,

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} by

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

and the cluster centers c_j by

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m} \quad (3)$$

$$\text{This iteration will stop when } \left\| U^{(t)} - U^{(t-1)} \right\| < \varepsilon, \quad (4)$$

where ε is a termination criterion between 0 and 1, whereas i are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m .

The algorithm consists of the following steps:

- 1 Accidental initialization of the reference matrix:
- 2 Calculation of the v_i cluster fusion centres according to formula of (3)
- 3 Aktualization of the “U” reference matrix according to formula of (2)
- 4 Algorithm finishes at the moment the difference of the reference between the actual step and the previous one is less than the ε tolerance having been chosen. In a case the offset is more, the algorithm is repeated since the step 2 till the condition (4) is fulfilled.

3.2 Gustafson&Kessel Algorithm

Gustafson & Kessel is an extension fuzzy c-means algorithm on an adaptive norm, enabling to provide clusters of various shapes in one set of data. Every cluster is characterized by its normalization matrix A_i . The matrix A_i are applied as optimization of variables in c-means functional. Each of the clusters is able to adapt one's own norm, in accordance with a topology data of a specific region. The objective function is defined as:

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

where

u_{ik}	membership level (degree) of x_i element in j cluster,
m	fuzziness ($m > 1$),
$d(x_k, v_i)$	distance between element x_i and centre of cluster i
A_i	standardization matrix of each cluster,
X	set of objects,
V	set of clusters,
U	matica funkcí příslušnosti objektov x k jednotlivým zhlukom Z , je vygenerovaná náhodne,
c	number of objects,
ε	tolerance.

Gustafson&Kessel algorithm consisted of the following steps:

- 1 Incidental initialization of set of membership

2 Compute matrix of covariance F by relation

$$F_i = \frac{\sum_{k=1}^n (u_{ik})^m \cdot (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^n (u_{ik})^m} \quad (6)$$

3 Updating of matrix of the U- membership by relation

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d(x_k, v_i)}{d(x_k, v_j)} \right)^{\frac{2}{m-1}}} \quad (7)$$

where $d(x_k, v_i)$ distance between element x_k and centre of cluster i

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

and standardization matrix A_i by relation (9)

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

4 Iteration repeating by step 2. Clustering finishes, when there reached finish conditions there.

4 The Obtained Results on Clustering the Earth Screen Data

The data set consists of 368 152 specimen of the Earth surface, where one of them represents area of 30 x 30 metres, representing total of 332 sq kms of a land. Specimen of the Earth surface is characterized by a 7- dimensional vector at which the partial factors are describing the brightness of the seven spectral bands. The data were obtained from transmission research of the Earth by LANSAT satellite. Fusion the data was realized via *fuzzy c-means* and *Gustafson&Kessel algorithms* for varied values of “m” parameter of indeterminacy from interval of <1,5; 5>. Figure 2 shows a real picture of the Earth area having been researched.

The obtained data were classified to 7 classes (clusters): old city, new city, the land not used, agricultural area, mine area, forest, water.



Figure 2

Photography of the transmission researching the Earth

Figures 3 and 4 show the best results of clustering, using the described fuzzy methods of c-means and the Gustafson & Kessel Algorithm, having been reached at the parameter of indeterminacy $m=2$.

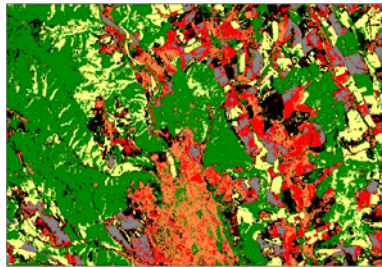


Figure 3

Result of fusion the FCMeans referring to parameter of indeterminacy $m=2$

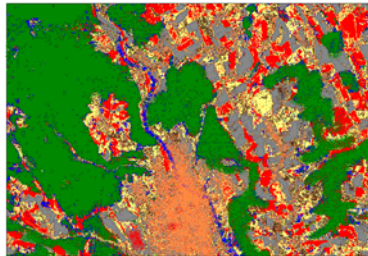









Figure 4

Result of fusion the G&K referring to parameter of indeterminacy $m=2$

Legend:

- | | | | |
|---|-------------------|---|-----------|
|  | old city |  | mine area |
|  | new city |  | forest |
|  | the land not used |  | water |
|  | agricultural area | | |

Conclusions

In the article there were described and via real Earth surface data experimentally attested fusion algorithms of c-means and of Gustafson&Kessel Algorithm. Comparing the results of classification the data via both of the methods, Fig. 2 and 4, it is possible to see a considerable deuce with a real picture of the Earth surface having been researched. For both of the fusion methods there were areas of mines and water areas problematic there. The reason for that is to be the fact that a specimen number was small and that these zones covered just a small part of the area having been researched.

The stated algorithms reach the high success in classification, they are also highly effective at the classification of a certain types of data.

Classification of a larger number of data of the same class is exact enough, anyway, as stated above, the problematic ones are the classes including a small number of specimen. Excluding these special cases, the algoritms described here are of a great application potential in many areas of human activities.

Acknowledgement

This work is supported by the VEGA project No 1/2185/05 and project MSM 6046070904.

References

- [1] Lukasová, A.-Šarmanová, J.: *Metody shlukové analýzy*, SNTL, Praha, 1985
- [2] Ocelíková, E.: *Multikriteriálne rozhodovanie*, ELFA, Košice, 2002
- [3] Tryon, R. C.: *Cluster Analysis*. Ann Arbor, Edwards Bros, 1939
- [4] Bin, W., Zhongzhi, S.: A clustering algorithm based on swarm intelligence. In: *Proceedings of Info-tech and Info-net, ICII 2001*, Vol. 3, pp. 58-66, 2001
- [5] Chatterjee, M., Das, S. K. and Turgut, D.: A weighted clustering algorithm for mobile ad hoc networks. *Journal of Cluster Computing (Special Issue on Mobile Ad hoc Networks)*, Vol. 5, No. 2, April 2002, pp. 193-204
- [6] Ben-Hur, Horn, D., Siegelmann, H. T and Vapnik, V.: A support vector clustering method. In: *International Conference on Pattern Recognition*, 2000
- [7] Hubert, L. J., Arabie, P. and De Soete, G.: *Clustering and Classification*. World Scientific, 1996
- [8] Klimešová D., Ocelíková E.: Spatial Data and Context Information. In: *Space without Frontiers, Cartography Brasil, Porto Alegre 2001*, pp. 22-25
- [9] Horváth, J., Zolotová, I.: Comparison of Two Clustering Methods in Image Segmentation. *Transactions of the VŠB -Technical University of Ostrava*, 2005, No. 2, Vol. LI, paper No. 1468, pp. 47-52, ISSN 1210-0471