

ON CLUSTER ANALYSIS BASED ON FUZZY RELATIONS BETWEEN SPATIAL DATA

Efendi NASIBOV

Dept. of Statistics, Dokuz Eylul University,
Buca 35160 Izmir, TURKEY
efendi.nasibov@deu.edu.tr

Gozde ULUTAGAY

Dept. of Statistics, Dokuz Eylul University,
Buca 35160 Izmir, TURKEY
gozde.ulutagay@ogr.deu.edu.tr

Abstract

Methods like DBSCAN are widely used in the analysis of spatial data. These methods are based on the neighborhood relations which use distance between points. However, these neighborhood relations consider to have at least a certain number of neighbors within a definite boundary. In this proposed work such a neighborhood analysis is done by using the benefits of fuzzy sets theory. Usage of fuzzy logic gives more sensitive and realistic results. In this paper, Fuzzy Joint Points (FJP) based on this theory is handled and some theoretical properties used in neighborhood analysis are investigated.

Keywords: Cluster analysis, fuzzy neighborhood relation, Fuzzy Joint Points (FJP).

1 Introduction

Cluster analysis has an important role in statistical data analysis. The main objective of clustering is to facilitate the analysis process by constructing similar objects in a cluster. Clustering methods can be divided into two groups such as hierarchical and prototype-based [3]. In hierarchical clustering, the remoteness of elements is the cornerstone. First of all, closer elements are put into same cluster, and in the next step, elements a little bit far away from the previous ones are put in another cluster, etc. In prototype-based methods, however, prototypes which have common features of some certain classes are formed and then the elements are taken into these classes with respect to the proximity degrees to the prototypes. Namely, in such a situation, not the remoteness of above-mentioned elements from each other, but their remoteness from the prototypes is considered. Some examples of these methods are k-means, k-medoids and FCM [1, 3, 4]. Besides, single-linkage (SLINK), complete-

linkage (CLINK), DBSCAN, OPTICS and FJP are some examples of hierarchical methods [2, 3, 5, 6, 7, 8]. In DBSCAN-like methods, in order to determine the core points of clusters or noise points, classical neighborhood density analysis is done. Thus, a point is conceived as a core point, if the number of points in a certain radius is larger than a specified threshold. However, in FJP-based methods, fuzzy neighborhood analysis that provides more sensitive results is used [7].

In this study, some researches are executed on fuzzy neighborhood analysis based on distance between points and it is showed that such a method can provide more sensitive and realistic results in comparison with non-fuzzy methods.

2 Fuzzy Neighborhood Based on Distance Between Points

Let us consider the data set $X = \{x_1, x_2, \dots, x_n\}$. The purpose of clustering is to determine sets X^1, X^2, \dots, X^k that constitute X which provides $\forall i, j : i \neq j \Rightarrow X^i \cap X^j = \emptyset$ and $\bigcup_{i=1}^k X^i = X$.

Let $F(E^p)$ indicate the set of all p -dimensional fuzzy sets in the space E^p . Let $\mu_A : E^p \rightarrow [0, 1]$ represent the membership function of the fuzzy set $A \in F(E^p)$.

Definition 1. A conical fuzzy point $A = (a, R) \in F(E^p)$ of the space E^p is a fuzzy set with membership function

$$\mu_A(x) = \begin{cases} 1 - \frac{d(x,a)}{R} & \text{if } d(x,a) \leq R \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $a \in E^p$ is the center of fuzzy point A , and $R \in E^1$ is the radius of its support, $\text{supp } A$.

α -level set of conical fuzzy point $A = (a, R)$ is calculated as follows:

$$\begin{aligned} A_\alpha &= \{x \in E^p \mid \mu_A(x) \geq \alpha\} \\ &= \{x \in E^p \mid d(x,a) \leq R \cdot (1 - \alpha)\}. \end{aligned}$$

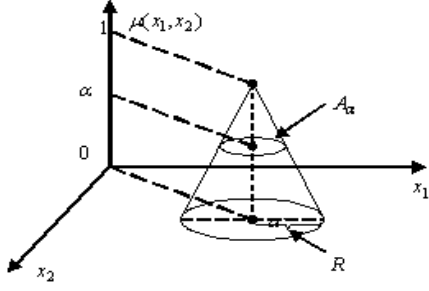


Figure 1: $A = (a, R) \in F(E^2)$ conical fuzzy point.

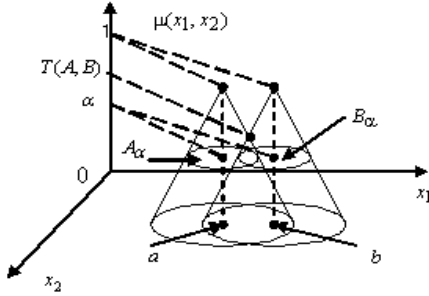


Figure 2: α -neighbor points $A = (a, R)$ and $B = (b, R)$ in fuzzy space E^2 .

The equivalent of the fuzzy conical point $A = (a, R) \in F(E^1)$ in the space E^1 is triangular and symmetric fuzzy number $A = (a, R, R)$.

Let $A = (a, R)$ and $B = (b, R)$, be fuzzy sets from the set $X \subset F(E^p)$. Let's denote a fuzzy similarity relation $T : X \times X \rightarrow [0, 1]$ on the set X as follows:

$$T(A, B) = 1 - \frac{d(a, b)}{2R}, \quad (2)$$

where $a \in E^p$ and $b \in E^p$ are the centers of the fuzzy points A and B respectively.

(2) can be written as follows:

$$d(a, b) = 2R \cdot (1 - T(A, B)). \quad (3)$$

It is obvious that the relation T is reflexive, thus $\forall A \in X, T(A, A) = 1$ is satisfied.

Lemma 2. *The fuzzy points $A = (a, R)$ and $B = (b, R)$ are called fuzzy α -neighbor points if and only if*

$$d(a, b) \leq 2R \cdot (1 - \alpha) \quad (4)$$

is satisfied.

Definition 3. Let A and B be fuzzy points on the set $X \subset F(E^p)$. If

$$T(A, B) \geq \alpha, \quad (5)$$

is satisfied for fixed $\alpha \in (0, 1]$, then the points A and B are called the α -neighbor fuzzy points and it is denoted by $A \sim_\alpha B$ (Fig.2).

Definition 4. If there is a sequence of α -neighbor fuzzy points $C^1, \dots, C^k, k \geq 0$, for fixed $\alpha \in (0, 1]$, between the points A and B , i.e.

$$A \sim_\alpha C^1, C^1 \sim_\alpha C^2, \dots, C^{k-1} \sim_\alpha C^k \text{ and } C^k \sim_\alpha B,$$

then the fuzzy points A and B are called α -joint fuzzy points.

Definition 5. Let $X \subset F(E^p)$ be a set of fuzzy points. If the fuzzy points A and B are α -joint for $\alpha \in (0, 1]$ and $\forall A, B \in X$, then the set X is called fuzzy α -joint set.

Suppose that the distance between the level sets A_α and B_α is defined as follows:

$$d(A_\alpha, B_\alpha) = \min\{d(x, y) \mid x \in A_\alpha, y \in B_\alpha\}.$$

Let the relation $\hat{T} : X \times X \rightarrow [0, 1]$ be the transitive closure of relation $T : X \times X \rightarrow [0, 1]$ by using max-min composition.

Lemma 6. *The fuzzy points A and B are called α -neighbor points if and only if*

$$A_\alpha \cap B_\alpha \neq \emptyset \quad (6)$$

is satisfied.

3 Fuzzy Neighborhood Membership Functions

Let us deal with two sets given in Fig. 3 within the same radius ε_1 , with equal number of elements, but with different densities. A method like DBSCAN that uses classical neighborhood relation, will consider the points x_1 and x_2 as core points and the cardinality of these sets will be equal within radius ε_1 .

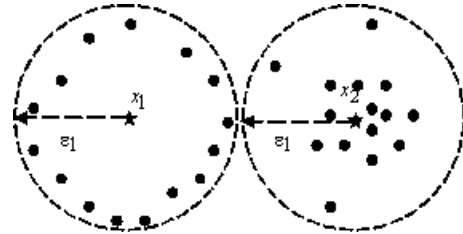


Figure 3: Points x_1 and x_2 are similar according to crisp neighborhood cardinality, but dissimilar according to fuzzy neighborhood cardinality.

Nevertheless, if these points are evaluated according to fuzzy neighborhood cardinality as done in FJP

method, since the proximity of the points around them will be taken into account, it is obvious that they will be different from each other. It can be said that analysis realized with such a point of view will produce more sensitive and realistic results.

As seen in Fig.4, since the classical neighborhood relation is used in DBSCAN method, although the point y_1 is closer to the point x , the neighborhood degree of the points y_1 and y_2 within radius ε_1 are equal with respect to the point x .

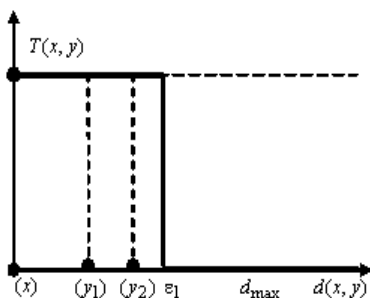


Figure 4: Neighborhood relation used in DBSCAN method.

However in FJP method, since the fuzzy neighborhood relation is used by taking into account the proximity of the points y_1 and y_2 to the point x is considered within radius ε_1 , a distinct difference is seen between the neighborhood degrees α_1 and α_2 of the points y_1 and y_2 respectively to the point x (Fig.5).

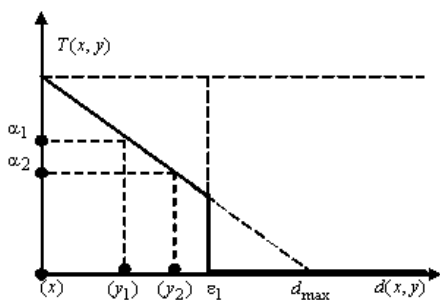


Figure 5: Neighborhood relation used in FJP method.

When dealing with fuzzy neighborhood analysis, different membership functions can be used in order to state neighborhood degrees. Also, the sensitivity of analysis can be increased by using different functions in calculation of the relation T (Fig. 5-6).

As seen from the figures, different sensitivity can be reached by using different membership functions in neighborhood analysis. For instance, the membership function in Fig. 4 handles the points within radius ε_1 as identical whereas there is an obvious distinction in Fig. 5. The difference becomes more evident in Fig.

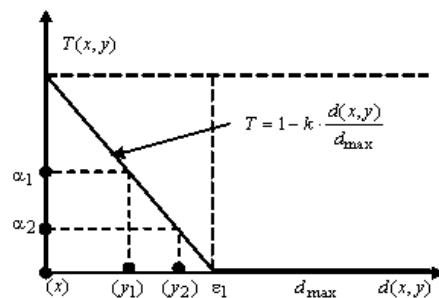


Figure 6: Linear neighborhood relation within an ε_1 radius.

6. However, the sensitivity is the same on near and far distance of the reference point. But in Fig. 7, the sensitivity differs exponentially.

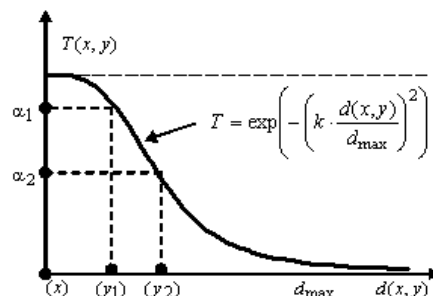


Figure 7: The effect of using different membership functions for a fuzzy point (exponential).

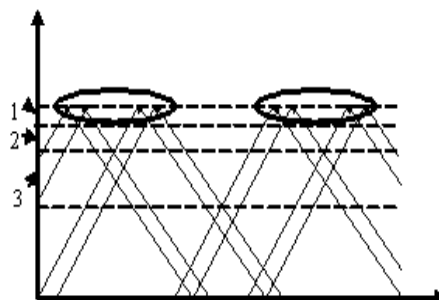


Figure 8: The effect of using different membership functions for a fuzzy point (linear).

Note that, in definition of the fuzzy point, if different nonlinear functions are used, different results can be found by the FJP algorithm. For example, consider a data set with 8 elements. If the membership function is linear as in Fig. 8, the widest change interval in which the α -parameter does not affect the number of clusters is found as the interval number 3 and according to the working principle of the algorithm, such a situation is appropriate for two clusters.

However, when the membership function of the fuzzy points looks like bell-shaped as in Fig. 9, the widest change interval found as number 2 and such a situation is suitable for four clusters. Thus, FJP algorithm finds two clusters if a membership function of a fuzzy point as in Fig. 8 is used whereas it detects four clusters if a membership function of a fuzzy point as in Fig. 9 is used.

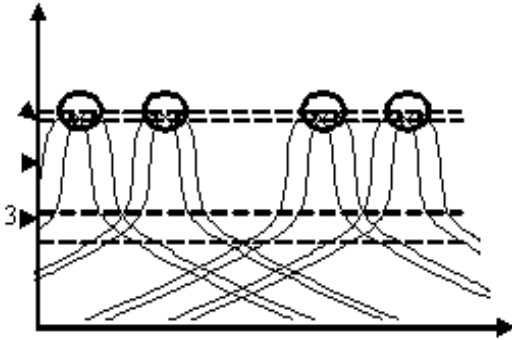


Figure 9: The effect of using different membership functions for a fuzzy point (bell-shaped).

Note that in FJP algorithm, the lowest layer that is suitable for one cluster is not taken into consideration.

4 Conclusion

In this study, the Fuzzy Joint Points (FJP) method is handled and fuzzy neighbor relations between points are investigated. FJP method may be more effectively used in the analysis of statistical databases with spatial data. As a result of our researches, since FJP method uses fuzzy logic, it produces more sensitive realistic results in comparison of methods based on classical neighborhood relation.

Also in our study, some theoretical results are given for the analysis of the cluster structure. Moreover, it is mentioned that using different membership functions for the neighborhood relation results in different partitions.

We think that FJP method may be more effective in areas such as mapping of national or regional geographical information, medical tomography, ultrasound or X-ray analysis, etc.

Acknowledgement

This study is supported as a research project No.106T312 by Scientific and Technological Research Institute of Turkey (TUBITAK).

References

- [1] J. C. Bezdek, *Fuzzy Mathematics in Pattern Classification*, PhD Thesis, Cornell Univ., New York, 1973.
- [2] M. Ester, H. Kriegel, J. Sander, X. Xu, A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (1996) pp.226–231.
- [3] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
- [4] J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, In Proc. 5th Berkeley Symp. Math. Statist. Prob. 1 (1967) pp.281–297.
- [5] E. N. Nasibov, An Alternative Fuzzy-Hierarchical Approach to Cluster Analysis, In Proc. 7th Int. Conf. on Application of Fuzzy Systems and Soft Computing, Germany, (2006) pp.113–123.
- [6] E. N. Nasibov, G. Ulutagay, A New Approach to Clustering Problem Using the Fuzzy Joint Points Method, *Automatic Control and Computer Sciences*,39 (2005) No.6, pp.8-17.
- [7] E. N. Nasibov, G. Ulutagay, On the Fuzzy Joint Points Method for Fuzzy Clustering Problem, *Automatic Control and Computer Sciences*, 40 (2006) No.5, pp.33-44.
- [8] E. N. Nasibov, G. Ulutagay, FJP: A New Hierarchical Method for Fuzzy Clustering, In Proc. 3rd Int. Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (2005) pp.212-220.
- [9] W. Pedrycz, F. Gomide, *An Introduction to Fuzzy Sets*, Massachusetts Institute, 1998.
- [10] J. Sander, M. Ester, H. P. Kriegel, X. Xu, Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications, *Data Mining and Knowledge Discovery 2* (1998), pp.169-194.