itüdergisi/***d***
mühendislik
Cilt:10, Sayı:1, 11-17
Şubat 2011

Bulanık c-means kümeleme yöntemine çıkarımlı yaklaşım

Mahmut HEKİM*, Umut ORHAN

Gaziosmanpaşa Üniversitesi, Elektronik Programı, 60250, Taşlıçiftlik, Tokat

Özet

Görüntü işleme, uzaktan algılama, veri madenciliği, örüntü tanıma ve benzeri konularda yaygın olarak kullanılan kümeleme yöntemleri, bir grup içindeki benzerliklerin gruplar arasındaki benzerliklerden daha büyük olmasını amaçlamaktadır. Farklı yoğunluklara sahip kümeler içeren veri uzayları için kümeleme işlemi zordur ve bu problemi çözmeye odaklanan birçok çalışma ileri sürülmüştür. K-means ve bulanık c-means kümeleme yöntemlerinin performansı küme merkezlerinin başlangıç değerlerine bağlıdır. Bu yüzden her iki algoritmanın da farklı küme merkezi başlangıç değeri için birçok defa çalıştırılması gerekir. Çıkarımlı kümeleme yöntemi ise veri noktalarının konumlarından veri uzayının yoğun bölgelerini tespit etmeye ve en çok komşuluğa sahip olan veri noktalarını küme merkezi olarak seçmeye dayanır. Bu özelliğiyle başlangıç koşulundan bağımsızdır ve algoritmanın bir kez çalıştırılması yeterlidir. Ancak, küme merkezleri veri noktalarından başka konumlardan saptanamadığı için bu yöntem her veri uzayına uygun olmayabilir. Bu makalede önerilen kümeleme yöntemi sayesinde genel kümeleme yöntemlerindeki başlangıç koşulu, ayrıca çıkarımlı kümeleme yönteminde küme merkezlerinin veri noktalarından seçilme zorunluluğu ortadan kaldırılmıştır. Dört yapay veri uzayı ile test edilen yeni yöntem, k-means, bulanık c-means ve çıkarımlı kümeleme yöntemleri ile karşılaştırılmıştır. Sonuç olarak bulanık c-means ve çıkarımlı kümeleme yöntemlerinin avantajlarını birleştiren yeni yöntem ile bulanık c-means yönteminin başlangıç koşuluna bağımlılığı ve küme merkezlerinin veri noktalarından seçilmesi zorunluluğu ortadan kaldırılmıştır.

Anahtar Kelimeler: *başlangıç koşulu, k-means, bulanık c-means, çıkarımlı kümeleme.*

*Yazışmaların yapılacağı yazar: Mahmut HEKİM. mhekim@gop.edu.tr; Tel: (356) 252 16 16 dahili: 2623.
Makale metni 09.04.2008 tarihinde dergiye ulaşmış, 23.10.2009 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 30.06.2011 tarihine kadar dergiye gönderilmelidir.
Bu makaleye "Hekim, M., Orhan, U., (2011) 'Bulanık c-means kümeleme yöntemine çıkarımlı yaklaşım', İTÜ Dergisi/D Mühendislik, 10: 1, 11-17" şeklinde atıf yapabilirsiniz.

Subtractive Approach to fuzzy c-means clustering method

Extended Abstract

Data clustering is an important part of cluster analysis. Based on various theories, numerous clustering algorithms have been developed, and new algorithms continue to appear in the literature. The aim of clustering is to obtain the most different groups in a dataset. A clustering method finds the similar data points and puts them into groups. If the groups in a dataset are found, then the dataset can be represented by fewer symbols. In the literature, researchers have proposed many solutions for this issue based on different theories. But there are still some problems such as the optimum cluster centers and the initial condition.

The best-known and earliest clustering method is K-means clustering algorithm. Its main advantageous is the capacity of fast converging. In spite of having many successful applications in several fields, it has many drawbacks. The membership values being only 0 or 1 may not always reflect the practical relationship between the data point and the cluster. In order to cope with this drawback, fuzzy c-means method employs fuzzy partitioning so that each data point can belong to several clusters with membership values between 0 and 1. Both clustering techniques try to group the data into given the number of clusters. Another method, subtractive clustering, finds the largest cluster by using the density function, then the second one, and so on. Subtractive clustering method uses the locations of the data points to calculate the density function.

K-means method tends to making homogenous distribution. Fuzzy c-means clustering method makes clusters with soft edges. Subtractive clustering usually tries to find the discreteness.

The locations of cluster centers in K-means and fuzzy c-means clustering may not be same for each time because of depending on initial condition. Therefore, they should be run several times for all datasets. Subtractive clustering method has only one solution independent of initial condition; consequently, it is enough to run once. But the main problem of subtractive clustering method is that the cluster centers are selected among data points. Because the cluster centers selected among data points may not represent the clusters of dataset.

In this paper, we offer a new approach which combines fuzzy c-means and subtractive clustering methods. The novel approach takes account of both discreteness and soft edges distribution; so the result has a similar appearance to average of other methods. The three main contributions of new approach can be summarized as: it becomes a more sophisticated technique by taking advantages of fuzzy c-means and subtractive clustering methods; it removes the initial condition. It has also only one solution independent of initial condition as in subtractive clustering method.

The novel algorithm consists of the following steps:

Step1. Normalize the data points.
Step2. Calculate the density value of each data point by Equation (10).
Step3. Select the point having the highest density value as cluster center.
Step4. Update the densities of each data point by Equation (11). If the number of detected cluster centers is less than the desired number, then go to Step3.
Step5. Compute the membership matrix by Equation (7).
Step6. Update the cluster centers by Equation (6).
Step7. Calculate the cost function by Equation (5). If it is bigger than the selected threshold value, go to Step5.

Clustering methods are usually evaluated and tested by using the artificial datasets. These methods must be able to analyze the datasets with different feature and sampling size. Artificial datasets used in the literature have some properties such as symmetric, discrete, and identical form. Therefore, we have used many special datasets in the numeric examples.

Finally, the novel approach is successful for both symmetric-identical and asymmetric-non-identical datasets. It also removes dependence on the initial condition in contrast to common KM and FCM clustering methods.

Keywords: *K-means; fuzzy c-means; subtractive clustering.*

Giriş

Görüntü işleme, uzaktan algılama, veri madenciliği ve örüntü tanıma gibi konularda yaygın olarak kullanılan kümeleme yöntemlerinin amacı bir veri uzayını genellikle benzerliğe dayalı olarak gruplara ayırmaktır. Bir kümeye ait olan veri noktalarının kendi içinde en çok benzeşen diğer kümelere ait veri noktaları ile en az benzeşen olması istenir. Bu yüzden kümeleme işlemi veri uzayını benzer özellikli homojen gruplara bölmeyi hedefler. Literatürde farklı teorilere dayalı birçok kümeleme yöntemi önerilmiş ve birçok araştırmada ise özel kümeleme tekniklerine odaklanılmıştır (Berkhin, 2006; Hekim ve Orhan, 2007).

En yaygın kullanılan bölümlemeli kümeleme algoritması olan K-means (KM) yönteminin avantajı yüksek kapsama hızı ve düşük miktarda saklama kapasitesiyle çalışabilmesidir. Bununla beraber, birçok dezavantajı vardır. Örneğin, üyelik fonksiyonu değerlerinin sadece 0 veya 1 olması, nokta ve küme arasındaki pratik ilişkiyi yeterli derecede yansıtamaz (Berkhin, 2006). Bu gibi dezavantajları ele alan bölümlemeli kümeleme algoritmaları arasında, yoğunluk tahminine dayalı ortalama kaydırma prosedürü, ayrık veri için yoğunluk temelli en yakın durağan noktayı kullanan ortalama kaydırma prosedürü ve değişken band genişlikli ortalama kaydırma prosedürü vardır (Cheng, 1995; Comaniciu ve Meer, 2002; Comaniciu, 2003).

Bulanık C-means (FCM) yöntemi, KM algoritması üzerinde iyileştirmeler yapılarak geliştirilmiştir. Bu teknikte her bir veri noktası bir üyelik derecesiyle birçok kümeye ait olabilir. KM yönteminde olduğu gibi FCM yöntemi de benzersizlik ölçütü olan maliyet fonksiyonunun indirgenmesine dayanır (Baraldi ve Blonda, 1999; Cheng, 1995).

KM ve FCM yöntemlerinin performansı küme merkezlerinin başlangıç değerlerine bağlıdır. Bu yüzden her iki algoritma da farklı başlangıç küme merkezi değerleri ile birçok defa çalıştırılmalıdır. Genellikle kümeleme işleminin yerine getirilebilmesi için uygun parametre seçimi de önemlidir. Örneğin küme sayısı ve bulanık kümeleme algoritmalarında kullanılan üssel ağırlık parametrelerinin uygun değerlerde olması zorunludur (Baraldi ve Blonda, 1999; Berkhin, 2006; Chopra vd., 2004).

Çıkarımlı kümeleme yöntemi ise veri noktalarının konumlarından veri uzayının yoğun bölgelerini tespit etmeye ve en çok komşuluğa sahip olan veri noktalarını küme merkezi olarak seçmeye dayanır. Bu sayede başlangıç koşulundan bağımsız olur ve algoritmanın bir kez çalıştırılması yeterlidir (Ross, 1995; Chopra vd., 2004). Ancak, küme merkezleri veri noktalarından başka konumlarda saptanamadığı için bu yöntem her veri uzayına uygun olmayabilir.

Bu makalede, başlangıç koşulunu ortadan kaldıran ve küme merkezlerini optimum konumlarda saptayan yeni bir bulanık kümeleme yöntemi önerilmektedir. Önerilen yöntem, çıkarımlı kümeleme yöntemi ile yoğun komşuluğa sahip veri noktalarından başlangıç küme merkezi konumlarının seçilmesine ve FCM yöntemi ile bu seçilen konumları kullanarak optimum küme merkezlerini saptamaya dayanmaktadır.

Veri kümelemeye genel bakış

K-means kümeleme

n adet x_j veri noktasına sahip bir veri uzayı c adet G_i grubuna bölünür (j=1,…,n ve i=1,…,c). Toplam maliyet fonksiyonu her bir G_i grubuna ait x_k noktası ve ilgili küme merkezi CC_i arasındaki Öklid uzaklığına dayalı maliyet fonksiyonlarının toplamıdır (Ross, 1995):

$$J = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \left\| x_k - CC_i \right\|^2 \right) \quad (1)$$

Kümelenen veri noktaları bir $c \times n$ ikili U üyelik matrisi ile tanımlanır ve bu matrisin u_{ij} üyelik değerleri aşağıdaki Eşitlik (2) ile bulunur.

$$u_{ij} = \begin{cases} 1 \text{ eger } \left\| x_j - CC_i \right\|^2 \le \left\| x_j - CC_k \right\|^2; \ \forall \text{k}, \ k \ne i \\ 0 \quad \text{diger} \end{cases} \quad (2)$$

Üyelik matrisi sabitlendiği zaman Eşitlik (1)'i indirgeyen küme merkezi, $i$nci küme içindeki tüm noktaların ortalaması olur:

$$CC_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (3)$$

burada $|G_i| = \sum_{j=1}^{n} u_{ij}$ 'dir.

KM algoritmasının performansı küme merkezi başlangıç değerlerine bağlıdır, bu yüzden farklı başlangıç küme merkezleri kullanılarak algoritmanın birçok kez tekrarlanması yararlı olacaktır.

Bulanık C-means kümeleme

FCM yönteminde, her bir veri noktası 0 ve 1 aralığında üyelik değerleriyle birkaç kümeye ait olabilmesi için bulanık kümeleme kullanılır. Bir veri noktasının tüm kümelere üyeliklerinin toplamı her zaman 1'dir (Ross, 1995):

$$\sum_{i=1}^{c} u_{ij} = 1, \ \forall j = 1, ..., n \quad (4)$$

Eşitlik (5) ile verilen maliyet fonksiyonu Eşitlik (1)'den geliştirilmiştir.

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| CC_i - x_j \right\|^2 \quad (5)$$

burada $u_{ij} \in [0,1]$; CC_i $i$nci küme merkezidir; $\left\| CC_i - x_j \right\|$ $i$nci küme merkezi ve $j$nci veri noktası arasındaki Öklid uzaklığıdır; ve $m \in [1, \infty]$ üssel ağırlıktır. Eşitlik (5)'in minimum değerine ulaşmak için gerekli koşullar Eşitlik (6) ve (7) ile verilmiştir:

$$CC_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \quad (6)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (7)$$

KM yönteminde olduğu gibi, FCM yönteminin performansı da başlangıç üyelik matrisine bağlıdır. Bu yüzden, küme merkezlerinin farklı başlangıç değerleri için algoritmanın birçok defa tekrarlanması gerekir.

Çıkarımlı kümeleme

Çıkarımlı kümeleme yönteminde, tüm veri noktaları küme merkezi adayıdır ve en çok sayıda yakın komşusu olan veri noktası küme merkezi olarak seçilir. Her bir veri noktası, birinci küme merkezine aday olduğu için birinci yoğunluk değeri şu şekilde tanımlanmıştır:

$$D_i = \sum_{j=1}^{n} \exp\left(- \frac{\left\| x_i - x_j \right\|^2}{(r_a / 2)^2} \right) \quad (8)$$

burada, r_a komşuluk yarıçapını gösteren pozitif bir sabittir. Bu yarıçapın dışında kalan veri noktaları komşuluk yoğunluğu üzerinde çok az etkiye sahiptir. Buna göre, bir veri noktası çok sayıda komşu veri noktasına yakın ise yüksek bir yoğunluklu olacaktır. En büyük yoğunluk değeri D_{CC_1}'e sahip veri noktası, birinci küme merkezi x_{CC_1} olarak seçilir. Sonra, her bir noktanın yoğunluğu D_i^* aşağıdaki eşitlik ile tekrar bulunur:

$$D_i^* = D_i - D_{CC_k} \exp\left(- \frac{\left\| x_i - x_{CC_k} \right\|^2}{(r_b / 2)^2} \right) \quad (9)$$

burada k ait olduğu döngüdeki küme sayısıdır. r_b, ikinci komşuluk yarıçapını gösteren pozitif bir sabittir. Sonraki yoğunlukları önemli derecede azaltmayı sağlar. Birinci küme merkezine yakın küme merkezlerinin saptanmasından sakınmak için r_a'dan çok az büyük olmalıdır ve genellikle r_a değerinin 1.25 katı alınır. Bu sayede, birinci küme merkezine yakın veri noktaları önemli derecede azalmış yoğunluğa sahip olacaktır. Bu yüzden bu veri noktalarının sonraki küme merkezi olarak seçilme olasılığı kalmayacaktır. Eşitlik (9) ile hesaplanan en büyük yoğunluk değerine sahip veri noktası, ikinci küme merkezi olarak seçilir ve bildirilen sayıda küme merkezi saptanana kadar tekrarlanır (Chopra, 2004).

Özellikle başlangıç koşulundan bağımsız olması nedeniyle iyi bir yaklaşım olan çıkarımlı kümeleme yönteminde, veri noktaları dışında küme merkezleri saptanmaz. Bu yüzden saptanan küme merkezleri her veri uzayına uygun olmayabilir. Ayrıca r_a ve r_b sabitlerinin uygun değerlerinin belirlenmesi veri uzayına bağlı olduğu için uzman görüşü gereklidir.

Önerilen yeni yöntemde, yukarıda anlatılan temel kümeleme yöntemlerinin avantajlarından yararlanılmıştır.

Bulanık c-means kümeleme yöntemine çıkarımlı yaklaşım

Bu çalışmada, çıkarımlı kümeleme yönteminde olduğu gibi komşuluk yoğunluklarına dayalı olarak başlangıç küme merkezi konumları bulunur. Bu konumlar FCM algoritması için küme merkezlerinin başlangıç değerleri olarak kullanılır. Yeni yaklaşım sayesinde KM ve FCM yöntemlerindeki başlangıç koşulu, ayrıca çıkarımlı kümeleme yönteminde küme merkezlerinin veri noktalarından seçilme zorunluluğu ortadan kaldırılır.

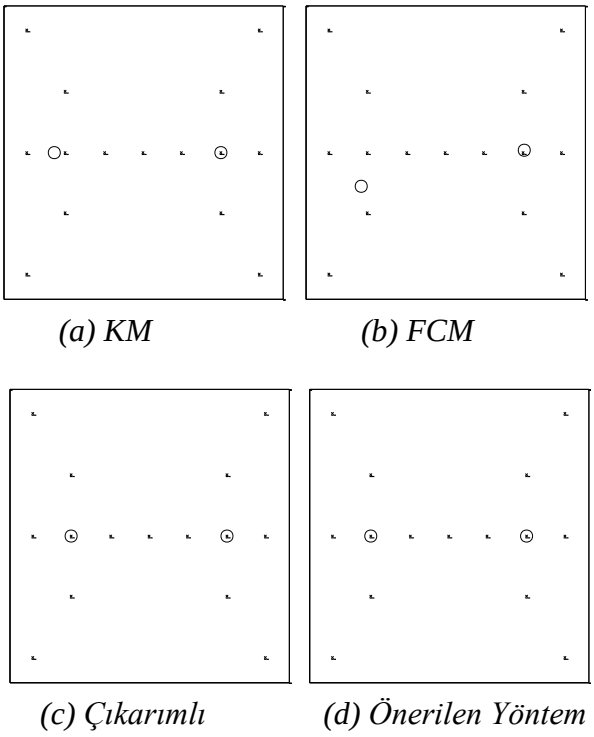
Önerilen yaklaşımda aşağıdaki adımlar kullanılarak küme merkezi konumları ve bulanık üyelik matrisi bulunur:

Adım1. Veri noktalarını [0 1] aralığına taşı.
Adım2. Eşitlik (8) ile her veri noktasının yoğunluklarını hesapla.
Adım3. En büyük yoğunluğa sahip veri noktasını küme merkezi olarak seç.
Adım4. Eşitlik (9) ile her veri noktasının yoğunluğunu güncelle. Eğer bulunan küme sayısı istenenden küçük ise Adım3'e git.
Adım5. Eşitlik (7) ile üyelik matrisini hesapla.
Adım6. Küme merkezlerini Eşitlik (6) ile güncelle.
Adım7. Eşitlik (5) ile maliyet fonksiyonunu hesapla. Seçilen eşik değerinden büyük ise Adım5'e git.

Nümerik örnekler

Bu bölümde, önerilen yöntemin anlaşılmasını sağlamak için üç yapay veri uzayı kullanılmıştır ve önerilen yöntem genel kümeleme yöntemleri ile karşılaştırılmıştır. Aşağıdaki nümerik örneklerde kullanılan veri uzayları [0 1] aralığına taşınmıştır.

Örnek 1. İki boyutlu 15 veri noktasına sahip kelebek veri uzayı iki kümeye sahiptir (Ross, 1995). Simetrik dağılıma sahip olan bu veri uzayı birçok çalışmada algoritma geçerliliğini test etmede kullanılmaktadır. Şekil 1'de veri uzayı için yöntemlerin kümeleme sonuçları gösterilmektedir.



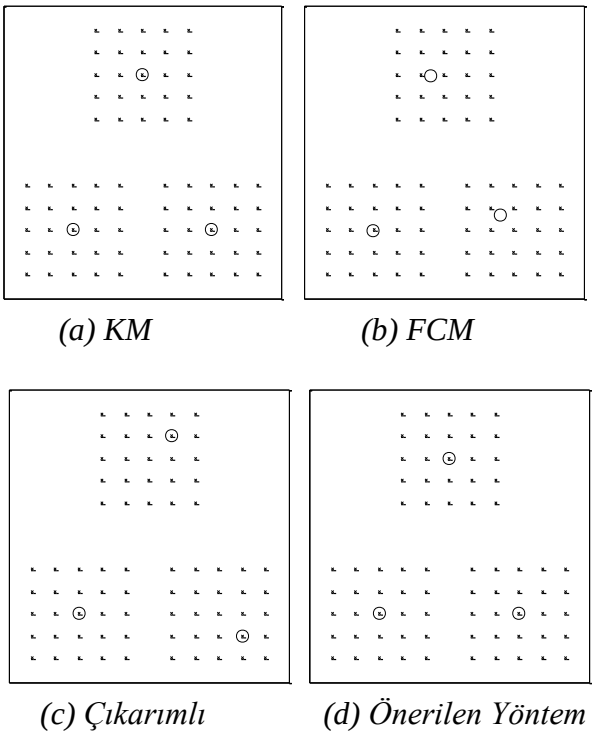
(a) KM *(b) FCM*

(c) Çıkarımlı *(d) Önerilen Yöntem*

Şekil 1. Kelebek veri uzayı için kümeleme yöntemleri

Örnek 2. İki boyutlu 75 veri noktasına sahip yapay veri uzayı üç kümeye sahiptir ve her biri 25 adet veri noktası içeren eş karesel yapıdadır (Rhee ve Oh, 1996). Gerçek küme merkezleri eş karesel üç yapının orta noktalarıdır. Şekil 2'de kümeleme sonuçları gösterilmektedir.

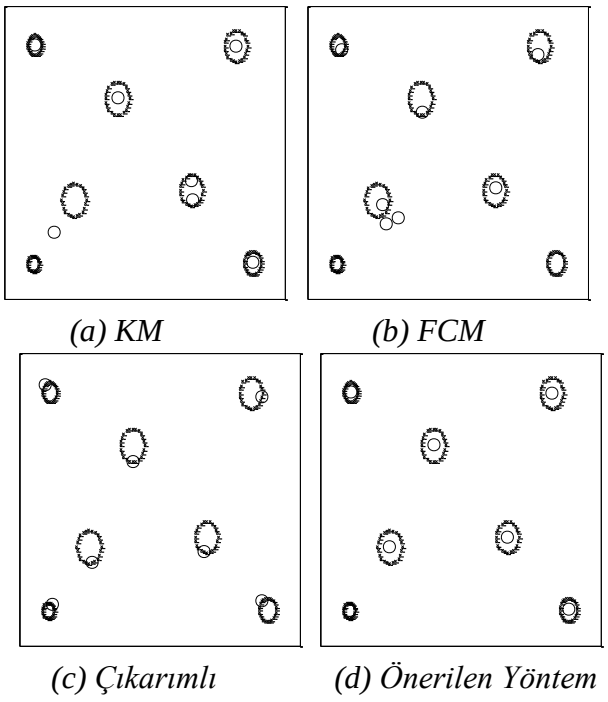
Örnek 3. İki boyutlu 168 noktaya sahip dairesel yedi kümeli veri uzayı hazırlanmıştır. Her bir kümede 24 veri noktası bulunmasına rağmen farklı çaplarda kümeler oluşturulmuştur. Bu veri uzayı için küme merkezleri dairesel yapıların veri noktası içermeyen orta noktalarıdır. Şekil 3'te kümeleme sonuçları gösterilmektedir.



(a) KM *(b) FCM*

(c) Çıkarımlı *(d) Önerilen Yöntem*

Şekil 2. Üç küme içeren simetrik veri uzayı için kümeleme yöntemleri

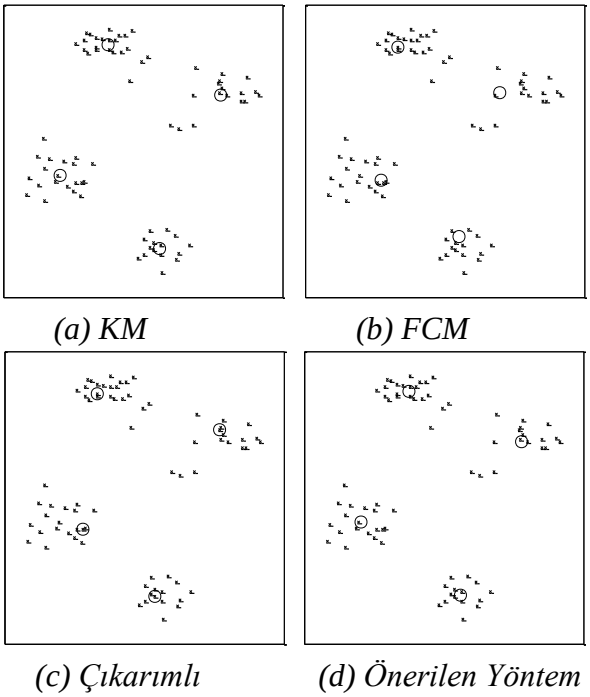


(a) KM *(b) FCM*

(c) Çıkarımlı *(d) Önerilen Yöntem*

Şekil 3. Yedi küme içeren veri uzayı için kümeleme yöntemleri

Örnek 4. Kümeleme tekniklerini göstermek için literatürde çok popüler olan Ruspini veri uzayı 75 noktadan oluşur ve dağınık yapıya sahip dört küme içerir (Ruspini, 1970). Şekil 4'te kümeleme sonuçları gösterilmektedir.



(a) KM *(b) FCM*

(c) Çıkarımlı *(d) Önerilen Yöntem*

Şekil 4. Ruspini veri uzayı için kümeleme yöntemleri

Sonuç

Bu makalede önerilen kümeleme yöntemi, optimum küme merkezi konumlarını bulan ve başlangıç koşulunu ortadan kaldıran yeni bir yaklaşımdır. FCM yöntemi ve çıkarımlı kümeleme yöntemini hibrit biçimde kullanan yeni yaklaşım sayesinde KM ve FCM yöntemlerindeki başlangıç koşulu, ayrıca çıkarımlı kümeleme yönteminde küme merkezlerinin veri noktalarından seçilme zorunluluğu ortadan kaldırılmıştır. Bazı yapay veri uzayları için test edilen yeni yöntemin sonuçlarına göre küme merkezi konumlarının iyileştiği görülmüştür. Önerilen yeni yöntemin özellikle bulanık kural tabanlı uygulamalar için yararlı olacağı düşünülmektedir.

Kaynaklar

Baraldi, A. ve Blonda, P., (1999). A survey of fuzzy clustering algorithms for pattern recognition-part I and II, *IEEE Trans. Systems, Man, and Cybernetics, Part B*, **29**, 6, 778-801.

Berkhin, P., (2006). Survey of clustering data mining techniques, grouping multi-dimensional data, *Recent Advances in Clustering*, 25-71.

Cheng, Y., (1995). Mean shift, mode seeking, and clustering, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **17**, 8, 790-799.

Chopra, S., Mitra, R. ve Kumar, V., (2006). Reduction of fuzzy rules and membership functions and its application to fuzzy PI and PD type controllers, *International Journal of Control, Automation, and Systems*, **4**, 4, 438-447.

Comaniciu, D. ve Meer, P., (2002). Mean shift: a robust approach toward feature space analysis, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **24**, 5, 603-619.

Comaniciu, D., (2003). An algorithm for data-driven bandwidth selection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **25**, 2, 281-288.

Hekim, M. ve Orhan, U., (2007). A validity measure for a new hybrid data clustering, *International Symposium on Innovations in Intelligent Systems and Applications*, 70-74.

Orhan, U. ve Hekim, M., (2007). Mass action based data clustering method and its weighted fuzzification, *5th International Conference on Electrical and Electronics Engineering*, 386-390.

Rhee, H.S. ve Oh, K.W., (1996). A validity measure for fuzzy clustering and its use in selecting optimal number of clusters, *IEEE International Conference Fuzzy Systems*, **2**, 1020-1025.

Ross, T.J., (1995). *Fuzzy logic with engineering applications*, McGraw-Hill.

Ruspini, E. H., (1970). Numerical methods for fuzzy clustering, *Information Sciences*, **2**, 319-350.