VERİ KÜMELEME ÜZERİNE HİBRİT YAKLAŞIM

Mahmut HEKİM[1] **Umut ORHAN[2]** **Günsel DURUSOY[3]**

[1]Gaziosmanpaşa Üniversitesi, Tokat MYO, Elektronik Programı, Taşlıçiftlik, 60250 Tokat

[2]Gaziosmanpaşa Üniversitesi, Tokat MYO, Bilgisayar Programı, Taşlıçiftlik, 60250 Tokat

[3]Elektronik ve Haberleşme Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi, İTÜ 80626, Maslak, İstanbul

[1]e-posta: mhekim@gop.edu.tr [2] umutorhan@gop.edu.tr [3]gunsel@ehb.itu.edu.tr

Anahtar sözcükler: Keskin kümeleme, Bulanık kümeleme, Çıkarımlı kümeleme, Hibrit Kümeleme

ABSTRACT

Data clustering method is a process of putting similar data into groups. A clustering method partitions a data set into several groups such that the similarity within a group is larger than among groups. In this paper, the most representative clustering methods (hard clustering, fuzzy clustering and subtractive clustering) are investigated in detail, and a new hybrid approach is developed by combined with the others. These techniques have been implemented against a data set for lowering blood pressure during surgery. In addition, performance and accuracy of all methods are studied and compared.

1. GİRİŞ

Veri kümeleme, verideki benzerlikleri bulma ve veriyi gruplara ayırmak için bir yaklaşımdır. Bu, bir grup içindeki benzerliğin gruplar arasındaki benzerlikten daha büyük olması için, veri kümesinin birçok gruba ayrılması demektir [1]. Veri gruplama veya kümeleme düşüncesi insan düşünce yapısına yakındır. Büyük miktarda veri ele alındığında, onun analizini basitleştirmek için bu büyük sayıdaki veriyi daha küçük boyutlu gruplara ya da kategorilere ayırmaya çalışılır. Bu grupları bulmak ve kategorize etmek, veri boyutu küçük olmadığı (en fazla iki veya üç boyut) sürece basit değildir ve problemi çözmek için birçok metot önerilmiştir. Bu metotlar "veri kümeleme yöntemleri" olarak adlandırılır. Kümeleme algoritmaları, yalnızca veriyi kategorize etmek için değil, ayrıca veri sıkıştırma ve model oluşturma için de yararlıdır. Ayrıca, veri grupları saptanabilir ise, bu gruplara dayalı olarak problemi çözmek için bir model oluşturulabilir [2].

Bu çalışmada en çok üzerinde çalışılan kümeleme teknikleri ele alınmış ve bazı iyileştirmeler yapılarak yeni bir yaklaşım önerilmiştir. Genel kümeleme yöntemleri arasında bulunan Keskin Kümeleme, Bulanık Kümeleme ve Çıkarımlı Kümeleme yöntemleri aşağıda incelenmekte ve bu çalışmada önerilen Hibrit Kümeleme Yöntemi (HKY) sunulmaktadır.

2. GENEL KÜMELEME YÖNTEMLERİ

Keskin Kümeleme Yöntemi (KKY), veri (görüntü, ses vb) sıkıştırma gibi birçok alanda kullanılmaktadır. Bu yöntem, uzaklık ölçütlerinin *J* maliyet fonksiyonunu minimize etmeye çalışarak küme merkezlerini bulmaya dayanır. Genellikle, uzaklık ölçütü olarak "Öklid Uzaklığı" kullanılır. n adet x_j kümesi c adet gruba (G_i) bölümlenir ($j=1,...,n$ ve $i=1,...,c$). j grubundaki bir x_k vektörü ve uygun küme merkezi c_i arasındaki Öklid uzaklığına dayalı olarak maliyet fonksiyonu şu şekilde tanımlanabilir [3]:

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

Burada,

$J_i=\sum_{k,x_k\in G_i}\left\|x_k-c_i\right\|^2$ ve G_i'nin maliyet fonksiyonudur.

Elde edilen gruplar $c\times n$ ikili üyelik matrisi *U* ile tanımlanır. Burada *j*nci veri noktası x_j G_i'ye ait ise, u_{ij} elemanı 1'dir; diğer durumlarda 0'dır. Küme merkezleri c_i sabitlendiğinde, Eşitlik 1 için u_{ij} minimizasyonu aşağıdaki gibi elde edilir.

$$u_{ij}=\begin{cases}1, & eger\ \left\|x_j-c_i\right\|^2\le\left\|x_j-c_k\right\|^2 ise,\ \forall k\neq i\\ 0, & diger\end{cases} \quad (2)$$

Eşitlik 2, c_i tüm merkezler arasında en yakın merkez ise x_j'in grup *i*'ye ait olduğunu ifade eder. Öte yandan, üyelik matrisi sabitlenmiş ise, o zaman Eşitlik

1'i minimize eden optimum merkez c_i, G_i içerisindeki tüm vektörlerin ortalamasıdır:

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

Burada $|G_i|$, G_i'nin boyutudur veya başka bir deyişle G_i'ye ait olan vektör sayısıdır ($|G_i| = \sum_{j=1}^{n} u_{ij}$).

KKY'nin performansı küme merkezlerinin başlangıç değerlerine bağlıdır ve bu yüzden birçok defa farklı başlangıç değerleri kullanılması gerekir.

KKY üzerinde iyileştirme yapılarak elde edilen Bulanık Kümeleme Yöntemi'nde (BKY), her bir veri noktası bütün kümelere bir üyelik derecesi ile aittir. Bulanık kümeleme de keskin kümeleme gibi uzaklık ölçütüne ait maliyet fonksiyonunun minimize edilmesine dayanır [1],[2]. Ancak, KKY'de her bir veri noktası sadece bir kümeye aittir ve diğerlerine ait değildir. Oysa, BKY'de her bir veri noktası 0 ve 1 değerleri arasında üyelik derecesi ile birçok gruba bağlıdır. Bu yüzden, üyelik matrisi *U*, elemanların 0 ve 1 arasında değer almasına imkan tanır. Bir veri noktasının tüm kümelere aitlik derecelerinin toplamı her zaman 1'dir.

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

BKY için maliyet fonksiyonu Eşitlik 1'in genelleştirilmiş bir halidir.

$$J(u, c_1, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \quad (5)$$

Burada; u_{ij} 0 ve 1 arasındadır; c_i bulanık G_i'nin küme merkezidir; $d_{ij} = \|c_i - x_j\|$, *i*nci *i* küme merkezi ve *j*nci veri noktası arasındaki Öklid uzaklığıdır; $m \in [1, \infty]$ üstel ağırlıktır.

Eşitlik 5'te minimuma ulaşmak için gerekli koşullar Eşitlik 6 ve 7'dir.

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

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

BKY'de, veri kümesi bölümlenirken minimize edilen maliyet fonksiyonu kullanılır. KKY'de olduğu gibi BKY'nin performansı da başlangıç üyelik matrisine bağlıdır. Bu yüzden, farklı üyelik dereceleri ile algoritma birçok defa çalıştırılmalıdır[3].

Çıkarımlı Kümeleme Yöntemi (ÇKY), veri noktalarının yoğunluk ölçütüne dayanır. Bu yöntem, yüksek yoğunluklu veri noktalarının özellik uzayındaki bölgelerinin saptanmasına dayanır [3]. En yüksek sayıda komşusu olan nokta, bir küme için merkez olarak seçilir. ÇKY'de, her bir veri noktası x_i, küme merkezi için aday olduğundan dolayı birinci küme merkezi için yoğunluk ölçütü şu şekilde tanımlanmıştır [4]:

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

Burada, r_a komşuluk yarıçapını gösteren pozitif bir sabittir. Buna göre, bir veri noktası çok sayıda komşu veri noktalarına sahip ise, bu veri noktası yüksek bir yoğunluğa sahip olacaktır. En büyük yoğunluk değeri D_{c_1}'e sahip nokta, birinci küme merkezi x_{c_1} olarak seçilir. Her bir veri noktasının yoğunluğu aşağıdaki eşitlik kullanılarak elde edilir:

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

Buradaki r_b, yoğunluk değerindeki yeterli azaltmaya sahip komşuluğu tanımlayan pozitif bir sabittir. Bu yüzden, birinci küme merkezi x_{c_1}'e yakın veri noktaları önemli ölçüde azaltılmış yoğunluk değerine sahip olacaktır. Eşitlik 9 yinelenerek, en büyük yoğunluk değerine sahip nokta sonraki küme merkezi olarak seçilir. Bu süreç bildirilen sayıda küme elde edilene kadar devam eder. Ancak, ÇKY'de küme merkezleri olarak seçilen noktalar veri uzayı elemanlarından olacağından dolayı her uzay için uygun olmayacaktır. Aşağıda çalışılan yöntem de bu problemler ortadan kaldırılmıştır.

3. HİBRİT KÜMELEME YÖNTEMİ

Genel kümeleme tekniklerinde veri boyutu ve küme sayısına göre bazı uygulama alanlarında istenen doğrulukta sonuç elde edilemeyebilir. Keskin Kümeleme ve Bulanık Kümeleme Yöntemleri'nde küme sayısı, algoritmanın doğruluğu ile direkt ilgilidir [3],[4]. Küme sayısı eksik ya da fazla verildiği zaman algoritmanın saptadığı küme merkezleri, olması gerektiğinden farklı çıkabilir.

Kümeleme yöntemlerindeki problemleri ortadan kaldırmak için KKY, BKY ve ÇKY'nin avantajları birleştirilerek geliştirilen Hibrit Kümeleme Yöntemi, küme sayısını bulan ve hesaplama sayısının önemli ölçüde azaltılmasına imkan tanıyan yeni bir yaklaşımdır. Yoğunluk fonksiyonu D_i, Eşitlik 8 ve 9'da iyileştirmeler yapılarak elde edilmiştir. Burada *ks,* iterasyon sayısıdır ve sonlandırma koşulu $c_i = c_{i-1}$ sağlandığı zaman elde edilen son değer, optimum

küme sayısıdır. Bu koşul sağlandığında $ks = i - 1$ yapılır.

$$D_i = \begin{cases} \sum_{j=1}^{n} \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{(r_a / 2)^2} \right), & \text{eger ks} = 1 \text{ ise} \\ D_i - D_{c_{ks-1}} \exp\left(-\frac{\left\| x_i - x_{c_i} \right\|^2}{(r_b / 2)^2} \right), & \text{eger ks} \neq 1 \text{ ise} \end{cases} \quad (10)$$

Eşitlik 2 ile üyelik matrisi *U,* Eşitlik 1 ile de maliyet fonksiyonu hesaplanır. Maliyet fonksiyonu belirtilen tolerans değerinin altında ise iterasyona son verilir ve Eşitlik 3'e göre küme merkezleri güncellenir. Böylece, küme merkezleri kesin olarak saptanmış olur.

HKY'de her bir veri noktası bütün kümelere 0 ve 1 arasında bulunan bir üyelik derecesi ile aittir. Ancak, bu yöntemde BKY'de kullanılan *m* parametresinden bağımsız olarak üyelik derecelerini saptayan yeni bir eşitlik kullanılmıştır (11).

$$u_{ij} = \frac{S_{\min}}{S_j} \sum_{j=1}^{c} \sum_{i=1}^{ks} \left(\frac{\sum_{i=1}^{ks} \frac{1}{d_{ij}^2}}{d_{ij}^2} \right)$$

$$S_j = \sum_{i=1}^{ks} \left\| c_i - x_j \right\| \quad (11)$$

$$S_{\min} = \min S_j \qquad \text{ve} \qquad d_{ij} = \left\| c_i - x_j \right\|^2$$

Burada, $S_{\min}$ değeri, S_j değerleri arasındaki en küçük değerdir.

HKY, aşağıda gösterilen algoritma adımlarını kullanarak c_i küme merkezlerini ve U üyelik matrisini saptar:

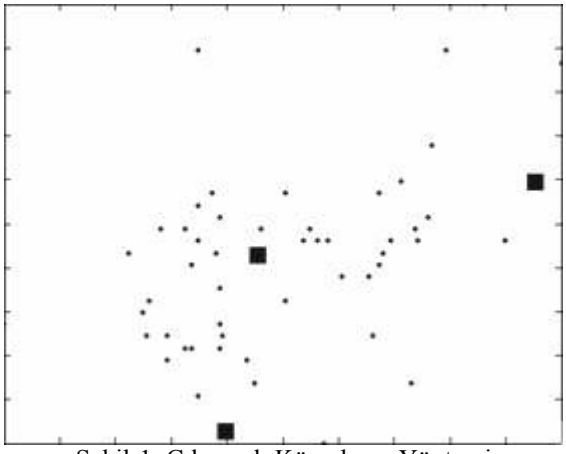
Adım 1. Veri normalize edilir.
Adım 2. Her bir veri noktasının yoğunluk değeri Eşitlik 10 ile hesaplanır.
Adım 3. En büyük yoğunluklu veri noktası *ks*'nci küme merkezi olarak seçilir. Bu adıma küme sayısı saptanana kadar devam edilir.
Adım 4. Adım 3'te elde edilen küme merkezleri, başlangıç değerleri olarak alınır.
Adım 5. Eşitlik 2 ile U üyelik matrisi saptanır.
Adım 6. Eşitlik 1'e göre maliyet fonksiyonu hesaplanır. Eşik değerinin altında ise Adım 8'e gidilir.
Adım 7. Eşitlik 3'e göre küme merkezleri güncellenir ve Adım 5'e dönülür.
Adım 8. Eşitlik 11'e göre U üyelik matrisi bulanıklaştırılır.

Genel kümeleme yöntemlerinde gözlemlenen uygun küme sayısı saptama problemi bu önerilen yöntemde giderilmiştir.

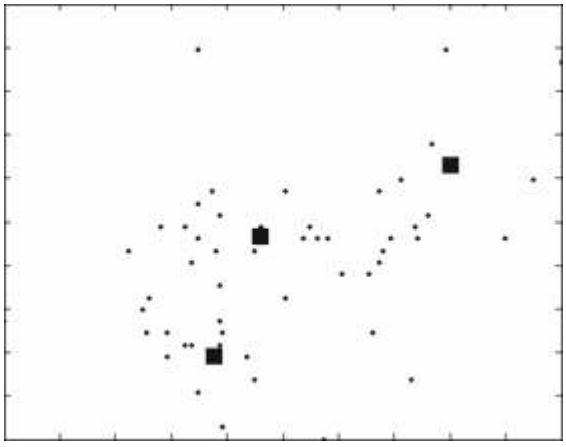
4. NÜMERİK ANALİZ

Bu makalede ele alınan ve geliştirilen yöntemlerin karşılaştırılmasında kullanılan veriler, ameliyat süresince düşen kan basıncının incelenmesi üzerine Robertson ve Armitage tarafından yapılan çalışmadan (sayfa 53-64) alınmıştır. Veri kümesinde kan basıncı ve buna göre verilen ilacın miktarı girdi değişkenleri olarak kullanılmıştır.

Nümerik analiz sonuçlarına göre; KKY ve BKY yöntemlerinin performansı ve doğruluğu, küme sayısının uygun değeri verildiğinde bile optimum sonuçtan uzaklaşabilmektedir ve yapılan her denemede farklı küme merkezleri bulmaktadır. Başka bir deyişle, birçok defa yapılan simülasyon denemelerinde bu kümeleme teknikleri farklı karakteristik sonuçlar göstermiştir.[5],[6] Bu yüzden, en iyi ve tek sonuç veren ÇKY ve HKY grafiksel olarak Şekil 1 ve Şekil 2'de gösterilmektedir.



Şekil-1. Çıkarımlı Kümeleme Yöntemi



Şekil-2. Hibrit Kümeleme Yöntemi

Aşağıdaki Tablo-1'de incelenen yöntemlerde elde edilen üyelik dereceleri ve Tablo-2'de küme merkezleri verilmiştir:

Tablo-1. KKY, BKY, ÇKY ve HKY yöntemlerinde elde edilen üyelik değerleri

	KKY			BKY			ÇKY			HKY		
	Küme1	Küme2	Küme3	Küme1	Küme2	Küme3	Küme1	Küme2	Küme3	Küme1	Küme2	Küme3
1	0	0	1	0.40	0.35	0.25	0.59	0.27	0.14	0.46	0.36	0.18
2	0	0	1	0.28	0.31	0.41	0	0	1	0.12	0.05	0.83
3	0	1	0	0.30	0.39	0.31	0.87	0.06	0.07	0.83	0.07	0.10
4	0	1	0	0.28	0.34	0.37	0.74	0.07	0.19	0.54	0.08	0.38
5	0	0	1	0.27	0.52	0.21	0.88	0.07	0.05	0.85	0.09	0.07
6	0	1	0	0.30	0.38	0.32	0.88	0.05	0.07	0.82	0.06	0.11
7	1	0	0	0.36	0.34	0.29	0.23	0.67	0.10	0.21	0.68	0.11
8	1	0	0	0.47	0.31	0.22	0.52	0.37	0.11	0.37	0.51	0.12
9	0	0	1	0.23	0.27	0.50	0.40	0.04	0.56	0.01	0.00	0.99
10	1	0	0	0.43	0.33	0.24	0.34	0.58	0.08	0.10	0.87	0.03
11	0	0	1	0.05	0.91	0.04	0.92	0.04	0.04	0.88	0.06	0.06
12	0	0	1	0.26	0.30	0.44	0.31	0.04	0.65	0.04	0.01	0.95
13	0	1	0	0.29	0.32	0.39	0.44	0.10	0.47	0.33	0.11	0.56
14	0	0	1	0.26	0.53	0.21	0.88	0.07	0.05	0.81	0.10	0.09
15	0	1	0	0.29	0.37	0.34	0.88	0.04	0.08	0.78	0.06	0.16
16	0	0	1	0.26	0.29	0.46	0.07	0.01	0.92	0.06	0.02	0.93
17	0	1	0	0.28	0.42	0.30	1	0	0	0.97	0	0.03
18	0	1	0	0.28	0.38	0.34	0.96	0.01	0.03	0.86	0.03	0.11
19	0	0	1	0.30	0.48	0.22	0.84	0.10	0.07	0.75	0.15	0.10
20	0	1	0	0.07	0.09	0.84	0.38	0.04	0.58	0.02	0	0.97
21	0	1	0	0.29	0.37	0.34	0.92	0.03	0.05	0.82	0.04	0.14
22	0	1	0	0.26	0.31	0.43	0.65	0.06	0.29	0.31	0.05	0.64
23	0	0	1	0.33	0.40	0.27	0.75	0.12	0.12	0.60	0.18	0.22
24	1	0	0	0.42	0.34	0.24	0.48	0.43	0.09	0.24	0.70	0.06
25	0	1	0	0.22	0.26	0.52	0.29	0.04	0.67	0.06	0.01	0.92
26	0	1	0	0.29	0.36	0.35	0.83	0.06	0.12	0.69	0.07	0.23
27	1	0	0	0.53	0.27	0.20	0.26	0.66	0.07	0.13	0.82	0.05
28	1	0	0	0.40	0.33	0.27	0	1	0	0.08	0.88	0.04
29	0	0	1	0.27	0.31	0.42	0.14	0.03	0.83	0.08	0.02	0.90
30	1	0	0	0.37	0.34	0.29	0.13	0.82	0.06	0.16	0.75	0.09
31	0	0	1	0.43	0.34	0.23	0.57	0.31	0.12	0.43	0.42	0.15
32	0	1	0	0.33	0.36	0.31	0.56	0.28	0.16	0.48	0.33	0.19
33	1	0	0	0.42	0.32	0.26	0.09	0.87	0.04	0.15	0.76	0.09
34	0	1	0	0.27	0.35	0.38	0.89	0.02	0.09	0.53	0.04	0.43
35	0	0	1	0.24	0.28	0.48	0.52	0.05	0.43	0.04	0.01	0.95
36	0	1	0	0.26	0.30	0.44	0.58	0.06	0.36	0.25	0.04	0.71
37	1	0	0	0.40	0.34	0.26	0.19	0.76	0.05	0.02	0.97	0.01
38	0	1	0	0.22	0.26	0.51	0.46	0.05	0.49	0.10	0.02	0.88
39	0	0	1	0.34	0.34	0.31	0.48	0.19	0.33	0.37	0.21	0.42
40	0	1	0	0.17	0.19	0.64	0.38	0.04	0.58	0.04	0.01	0.96
41	1	0	0	0.37	0.34	0.29	0.23	0.68	0.08	0.18	0.75	0.08
42	0	1	0	0.28	0.36	0.37	0.89	0.03	0.08	0.67	0.05	0.28
43	1	0	0	0.54	0.26	0.19	0.37	0.53	0.10	0.25	0.64	0.10
44	0	0	1	0.37	0.36	0.26	0.63	0.22	0.15	0.50	0.28	0.22
45	0	1	0	0.28	0.44	0.27	0.97	0.01	0.01	1	0	0
46	0	1	0	0.32	0.42	0.27	0.85	0.09	0.06	0.83	0.11	0.06
47	0	0	1	0.29	0.38	0.33	0.83	0.05	0.12	0.54	0.07	0.39
48	0	0	1	0.34	0.33	0.33	0.34	0.22	0.45	0.31	0.23	0.45
49	0	1	0	0.30	0.39	0.31	0.93	0.03	0.04	0.91	0.03	0.06
50	1	0	0	0.93	0.04	0.03	0.36	0.54	0.09	0.22	0.70	0.08
51	0	1	0	0.24	0.28	0.47	0.45	0.05	0.50	0.15	0.03	0.82
52	0	0	1	0.31	0.32	0.37	0.13	0.06	0.81	0.22	0.12	0.66
53	0	1	0	0.25	0.30	0.46	0.63	0.05	0.32	0.10	0.02	0.88

Tablo-2. KKY, BKY, ÇKY ve HKY yöntemlerinde elde edilen küme merkezleri

	KKY		BKY		ÇKY		HKY	
	İlaç Miktarı	Kan Basıncı	İlaç Miktarı	Kan Basıncı	İlaç Miktarı	Kan Basıncı	İlaç Miktarı	Kan Basıncı
Küme 1	2.46	74.33	2.36	69	1.9	67	1.91	68.43
Küme 2	1.71	66	2.04	68	2.7	73	2.46	74.33
Küme 3	2.05	61.44	1.72	59	1.81	52	1.78	58.33

5. SONUÇ

Bu makalede, en çok kullanılan kümeleme yöntemleri incelenmiş ve yeni bir yöntem olarak dinamik kümeleme tekniği sunulmuştur. Genel kümeleme yöntemlerinin ve önerilen dinamik kümeleme yönteminin performansı ile doğruluğu incelenmiş ve karşılaştırılmıştır. Özellikle, genel kümeleme teknikleri arasında en yüksek doğruluğa sahip olan Keskin Kümelemenin, simülasyon denemelerinde veri kümesi büyüdükçe performansının oldukça düştüğü

gözlemlenmiştir. Diğer BKY ve ÇKY tekniklerinin ise performansları yüksek olmakla birlikte, doğruluklarının KKY'den daha kötü olduğu tespit edilmiştir. Bu makalede önerilen yöntem, simülasyon denemelerinin tekrarını gerektirmeksizin uygun küme sayısını saptayan ve daha güvenilir sonuç veren yeni bir tekniktir.

Önerilen yöntem, pratikteki uygulamalar için, daha karmaşık sistemlerin bünyesine eklenerek mevcut yapılara uyarlanabilir.

KAYNAKLAR

[1] Jang, J. S.R., Sun, C.T., Mizutani, E., Neuro-Fuzzy and Soft Computing – a Computational Approach to Learning and Machine Intellence, *Prentice Hall* , 1997.

[2] Yu, J., General C-Means Clustering Model, TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol.27, No.8, pp.1197-1211, 2005.

[3] Ross, T.J., Fuzzy Logic with Engineering Applications, *McGraw-Hil.,* 1995.

[4] Baraldi, A., Blonda, P., Survey of Fuzzy Clustering Algorithm for Pattern Recognition-Part I and II, TRANSACTIONS SYSTEMS, MAN, and CYBERNETICS, Part B, Vol. 29, No.6, pp.778-801, 1999.

[5] Han, J., Kanber, B., Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers,* 2000.

[6] Krishnapuram, R., Keller, J.M., A Possibilistic Approach to Clustering, TRANSACTIONS ON FUZZY SYSTEMS, Vol.1, No.2, pp.98-110, 1993.