BULANIK KÜMELEMEDE EN UYGUN KÜME SAYISININ YAPAY SİNİR AĞLARI VE DİSKRİMİNANT ANALİZİ İLE BELİRLENMESİ

Faruk ALPASLAN[1]
Necati Alp ERİLLİ[2]
Ufuk YOLCU[3]
Erol EĞRİOĞLU[4]
Ç.Hakan ALADAĞ[5]

Özet: Bir kümeleme probleminde, kümeler birbirinden belirgin bir şekilde ayrılmıyorsa ya da bazı birimler küme üyeliğinde kararsızsa, klasik kümeleme yöntemleri yerine bulanık kümeleme yöntemleri tercih edilmesi yararlı olacaktır. Kümeleme analizinde, anlamlı ve sağlıklı sonuçlara ulaşabilmek için en uygun küme sayısının belirlenmesi önemlidir. Ancak, birçok kümeleme algoritması küme sayısının önceden bilinmesini gerektirir. En uygun küme sayısının belirlenme işlemlerine genel olarak küme geçerliliği adı verilmektedir. Bulanık kümeleme ile ilgili literatürde en uygun küme sayısı, küme geçerlilik indeksleri ile belirlenmektedir. Bazı karmaşık yapılar içeren verilerde, küme üyeliklerindeki karasızlıklar nedeniyle, küme geçerlilik indeksleri en uygun küme sayısını belirlemede birbirleri ile çelişen sonuçlar verebilmektedir. Ayrıca hangi indeksin en uygun küme sayısını belirlediğini ortaya koyan bir ölçüt de bulunmamaktadır. Bu çalışmada, en uygun küme sayısını belirlemede, ileri beslemeli yapay sinir ağları ve diskriminant analizi kullanılmış; sonuçlar PC, CE gibi küme geçerlilik indekslerinden elde edilen sonuçlar ile karşılaştırılarak en uygun küme sayısı hakkında karar verilmeye çalışılmıştır.

Anahtar Kelimeler: Bulanık Kümeleme, Küme geçerlilik indeksi, Yapay sinir ağları, Diskriminant Analizi.

Abstract: In a clustering problem, it would be better to use fuzzy clustering if there was an uncertainty in determining clusters or memberships of some units. Determining the number of cluster has an important role on obtaining sensible and sound results in clustering analysis. In many clustering algorithm, it is firstly need to know number of cluster. However, there is no pre information about the number of cluster in general. The process of determining the most proper number of cluster is called as cluster validation. In the available fuzzy clustering literature, the most proper number of cluster is determined by utilizing cluster validation indices. When the data contain complexity are being analyzed, cluster validation indices can produce conflictive results. Also, there is no criterion point out the best index. In this study, artificial neural networks and discriminant analysis are employed to determine the number of cluster and the proposed method are applied some data and obtained results are compared to those obtained from validation indices like PC and CE.

Keywords: Fuzzy clustering, Cluster validation index, Artificial neural network, Discriminant analysis.

[1] Prof. Dr., Ondokuz Mayıs Üniversitesi Fen Edebiyat Fak.İstatistik Bölümü
[2] Arş. Gör., Ondokuz Mayıs Üniversitesi Fen Edebiyat Fak.İstatistik Bölümü
[3] Arş. Gör. Ondokuz Mayıs Üniversitesi Fen Edebiyat Fak. İstatistik Bölümü
[4] Yrd. Doç. Dr., Ondokuz Üniversitesi Fen Edebiyat Fak. İstatistik Bölümü
[5] Öğr.Gör. Dr., Hacettepe Üniversitesi Fen Fak. İstatistik Bölümü

I. Giriş

Kümeleme analizi, son yıllarda iş ve bilim alanında sıkça kullanılmaya başlanan çok değişkenli veri analiz yöntemlerinden biridir. Kümeleme Analizi, bireylerin ya da nesnelerin sınıflandırılmasını ayrıntılı bir şekilde açıklamak amacıyla geliştirilmiştir (Erilli, 2009). Kümeleme Analizi, bir araştırmada incelenen birimleri aralarındaki benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlamalar yapmayı sağlayan bir yöntemdir. Burada amaç; gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya uygun, işe yarar özetleyici bilgiler elde etmede yardımcı olmaktır (Tatlıdil, 2002). Başka bir ifade ile veriler arasındaki benzerlikler dikkate alınarak benzer verilerin aynı grupta veya kümede toplanmasını sağlamaktır. Kümeleme Analizi, önceden belirlenen seçme kriterine göre birbirine çok benzeyen birey ya da nesneleri aynı küme içinde sınıflandırır. Analizin sonucunda oluşan kümelerin kendi içindeki homojenliği yüksek ve kümeler arası heterojenliği düşük olacaktır (Kalaycı, 2005).

Bulanık kümelerin kümelemede kullanımı ilk kez Bellman ve ark.(1966) tarafından önerilmiştir. Bulanık Kümeleme; verileri kümelemek için bulanık teknikler kullanır ve bu tekniklerde bir nesne birden fazla kümeye sınıflandırılabilir. Bu tip algoritmalar gerçek sayıların belirsizliğini ele aldığından, günlük yaşamın tecrübelerine uygun kümeleme şekillerinin ortaya çıkmasına yardımcı olur (Erilli, 2009). Bulanık kümeleme analizinde en iyi küme sayısının belirlenmesi önemli bir problemdir. Literatürde en iyi küme sayısının belirlenmesi için çeşitli küme geçerlilik indeksleri önerilmiştir. Bu çalışmada bulanık kümelemede en iyi küme sayısının belirlenmesi için yapay sinir ağlarına dayalı bir ölçüt önerilmiştir.

Çalışmanın ikinci bölümünde bulanık kümeleme hakkında bilgi verilmiştir. Üçüncü bölümde literatürde çok iyi bilinen Bulanık C-Ortalamalar yöntemi tanıtılmıştır. Dördüncü bölümde bulanık küme geçerlilik indeksleri, beşinci bölümde yapay sinir ağları ve altıncı bölümde de diskriminant analizi hakkında özetleyici bilgi verilmiştir. Yedinci bölümde önerilen ölçüt üç benzetim ve birde gerçek veriye uygulanarak açıklanmıştır. Son bölümde ise elde edilen sonuçlar tartışılmıştır.

II. Bulanık Kümeleme

Bu yaklaşımda, kümeler birbirinden belirgin bir şekilde ayrılmıyorsa ya da üyeliklerinde bazı birimler küme üyeliğinde kararsızsa uygun bir yöntem olarak ortaya çıkmaktadır. Bulanık Kümeler kümedeki birimin üyeliği olarak tanımlanan 0 ile 1 arasındaki her birimi belirleyen fonksiyonlardır. Birbirine çok benzeyen birimler aynı kümede yüksek üyelik derecesine göre yer alırlar (Erilli, 2009).

Diğer kümeleme yöntemlerine benzer olarak Bulanık Kümeleme de uzaklık ölçümlerine dayanır. Bu uzaklık ölçütlerinden hangisinin seçileceği küme yapısına ve kullanılan algoritmaya bağlıdır. Bulanık Kümelemenin kullanışlı bazı özelliklerini şu şekilde sıralayabiliriz:

i. Yorum açısından kullanışlı olan üyelik değerleri sağlar.

ii. Uzaklık kullanımı konusunda esnektir.

iii. Üyelik değerlerinin bazıları bilindiğinde sayısal optimizasyonla birleştirilebilir (Naes ve Mevik, 1999).

Bulanık Kümelemenin klasik kümeleme yöntemlerine göre avantajı, veri hakkında daha detaylı bilgi vermesidir. Diğer taraftan dezavantajları da vardır. Çok sayıdaki birey ve küme durumunda çok fazla çıktı olacağından, özetlemek ve bilgiyi tasnif etmek zordur. Ayrıca bulanık kümeleme algoritmaları genellikle karmaşık yapıdadırlar ve daha çok belirsizlik söz konusu olduğunda kullanılır (Şahinli, 1999).

III. Bulanık C-Ortalamalar (BCO) Algoritması

Bulanık C-Ortalamalar algoritması, amaç fonksiyonuna dayanan bütün kümeleme tekniklerinin temelini oluşturmaktadır. Bezdek (1974) tarafından geliştirilmiştir. BCO algoritması sonuçlandığında, p boyutlu uzaydaki noktalar küresel bir şekil halini alır. Bu kümelerin yaklaşık olarak aynı boyutta olduğu varsayılır. Her bir kümeyi, küme merkezleri temsil eder ve bunlara prototip denir. Uzaklık ölçüsü olarak veriler ile küme merkezi arasındaki Öklid uzaklığını kullanır.

$$d_{ik} = d(x_i, v_k) = \left[\sum_{r=1}^{p} (x_{ri} - v_{rk})^2\right]^{1/2}$$

Burada x_i gözlem değerinin koordinat sistemindeki konumunu, v_k ise küme merkezini simgelemektedir.

Bu tekniğin uygulanabilmesi için küme sayısının ve bireylerin kümeye üyelik derecelerinin önceden bilinmesi gerekmektedir. Bu tür parametrelerin önceden bilinmesi zor olduğundan, bu değerler deneme yanılma yoluyla ya da geliştirilen bazı tekniklerle bulunabilir.

Bu kümeleme yöntemi için kullanılan amaç fonksiyonu şu şekildedir:

$$J(u, v) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}{}^{m} \left\| x_{ri} - v_{rk} \right\|^2$$

Bu fonksiyon ağırlıklandırılmış en küçük kareler fonksiyonudur. n parametresi gözlem sayısını, c ise küme sayısını gösterir. u_{jk}^{m} ise k. kümedeki x_j ' nin üyeliği, $f(u,v)$ değeri ise tüm ağırlıklandırılmış hata karelerinin toplamının bir ölçüsüdür (Şahinli, 1999).

Eğer $f(u,v)$ fonksiyonu c'nin her değeri için minimize edilecek olursa, diğer bir deyişle v_i'lere göre 1. dereceden türevi alınıp 0'a eşitlenirse BCO Algoritmasının prototipi şu şekilde olacaktır;

$$v_{jk} = \frac{\sum_{j=1}^{n} u_{jk}^{m} . x_{ik}}{\sum_{j=1}^{n} u_{jk}^{m}}$$

BCO Algoritması için gerekli adımlar ise şu şekildedir:

Adım 1: Başlangıç değerlerini belirle: Küme sayısı c, bulanıklık indeksi m, işlem bitirme kriteri ε ve üyelik dereceleri matrisi U veya V küme prototiplerini rasgele üretilir.

Adım 2: U küme prototiplerinin rasgele üretildiği varsayılırsa bu değerleri kullanarak üyelik dereceleri matrisini hesaplanır.

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ji}}{d_{jk}} \right)^{2/(m-1)} \right]^{-1}$$

Adım 3: Adım 2 eşitliğine göre U küme prototiplerini güncellenir.

Adım 4: $\left\| U^{(z)} - U^{(z-1)} \right\| \leq \varepsilon$ ise iterasyon durdurulur, aksi takdirde Adım 2'ye geri dönülür.

BCO Algoritması uygulandıktan sonra hangi bireyin hangi kümeye gireceğine karar vermek için üyelik dereceleri kullanılır. Her bir bireyin hangi kümeye olan üyeliğinin en büyük olduğuna bakılır ve bu bireyler o kümeye dâhil edilir. Ancak her bir birey diğer kümelere de belli bir üyelik dereceleri ile girebilir.

BCO Algoritmasının sonucu başlangıçta rasgele üretilen değerlere oldukça bağlıdır. Bu yüzden rasgelelikten kaynaklanan problemleri ortadan

kaldırmak için çeşitli algoritmalar geliştirilmiştir ve geliştirilmeye devam edilmektedir.

BCO, küme merkezlerini ve her veri noktası için üyelik derecelerini iterasyon yöntemi ile günceller ve küme merkezlerini veri seti içinde olması gereken yere taşır.

Küme merkezlerinin ilk yerleri, başlangıçta değeri rasgele atanan U matrisi kullanılarak oluşturulduğu için, BCO optimal sonuca yaklaşmayı garanti etmeyecektir (Sintas vd., 1999).

Performans; merkezlerin başlangıç yerlerine bağlıdır. Daha güçlü bir yaklaşım için aşağıda tanımlanan iki yol vardır.

i. Tüm merkezleri tanımlamak için bir algoritma kullanmak.

ii. BCO'yı farklı başlangıç merkezleri ile tekrarlı olarak çalıştırmak (Yeniden Başlama Stratejisi).

IV. Bulanık Kümeleme Geçerlilik İndeksleri

Kümeleme Analizi, benzer nesneleri aynı gruplara yerleştirmeyi amaçlamaktadır. Böylece büyük veriler içeren örneklemlerde, örnek dağılımları ve değişkenler arası korelasyonlar hakkında fikir edinmeyi amaçlamaktadır. Bununla birlikte birçok kümeleme algoritması küme sayısının önceden bilinmesini gerektirir. Birçok çalışmada, araştırmacının küme sayısı hakkında ön bilgisinin olmaması, bulunan küme sayısının gerçek küme sayısından az ya da çok olup olmadığının bilinmemesine yol açmaktadır. Eğer bulunan küme sayısı gerçek küme sayısından az çıkarsa, mevcut kümelerden bir veya birkaçı birleşmek durumunda olacaktır, çok çıkarsa mevcut kümelerden bir veya birkaçı bölünmelere uğrayacaktır. Optimal küme sayısının belirlenme işlemlerine genel olarak Küme Geçerliliği (Cluster Validity) adı verilmektedir. Böylece kümeleme işlemleri yapıldıktan sonra bulunan küme sayısının doğruluğunu tespit edebiliriz.

Veriler iki boyutlu uzayda olduğunda küme sonuçlarını görsel olarak yorumlayarak küme sayısına karar verilebilmektedir. Ama uzaydaki boyut sayısı arttıkça görsellik zorlaşmakta ve geçerlilik indekslerine ihtiyaç duyulmaktadır.

Sonuç olarak, kümeleme değeri ve en uygun kümeleme planlaması için iki kriterden bahsedilebilir.

1.Yoğunluk: Küme elemanlarının birbirlerine yakınlıklarını ölçer. Buna en iyi örnek olarak varyansı verebiliriz.

2.Ayrılma: İki kümenin birbirlerinden ne kadar ayrıldıklarını gösterir. İki farklı küme arasındaki mesafeyi ölçer.

A. Bölünme Katsayısı (Partititon Coefficient) (PC)

Bezdek (1974) tarafından önerilen bu yöntem; 1/c ile 1 arasında değer alır. Burada c küme sayısıdır. Bulanık bölünme sonucunda bütün üyelik

değerleri eşit çıkarsa, $u_{ij} = 1/c$ olacaktır. Bu aynı zamanda PC'nin en küçük değeridir. PC değerinin uygun kümeleme işlemindeki değerinin 1'e yakın olması istenen durumlardandır. PC değeri 1/c değerine yaklaştıkça kümeleme bulanıklaşacaktır. Ayrıca, 1/c değerine yakın bir değer kümeleme algoritmasının başarısız olduğunu gösterir.

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u^2{}_{ij}$$

B. Sınıflama Bölümlemesi (Classification Entropy) (CE)

Bu yöntem de Bezdek (1974) tarafından önerilmiştir.

$$V_{CE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_a u_{ij}$$

Burada a logaritma e tabanıdır. CE değerinin 0'a yakın olması istenmektedir. En iyi küme sayısı $2 \leq c \leq n-1$ aralığında olacaktır.

C. Xie-Beni İndeksi (XB)

Xie ve Beni tarafından (1991) geliştirilen bu indeks, yoğunluk ve ayrılma geçerlilik fonksiyonu olarak da bilinir ve şu şekildedir.

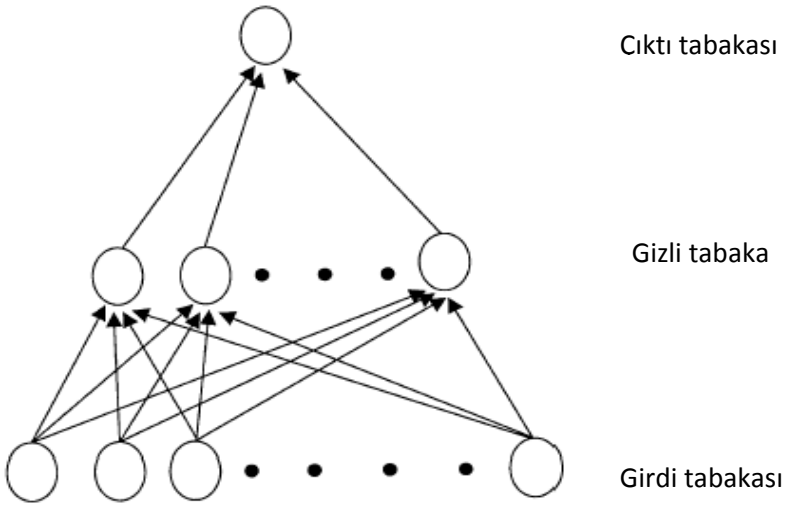
$$V_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\|x_k - v_i\|^2}{n \min_{i,j}\|v_i - v_j\|^2}$$

V. Yapay Sinir Ağları Yöntemi

Yapay sinir ağları biyolojik sinir ağlarını taklit ederek ortaya atılmış etkili bir analiz aracıdır. Yapay sinir ağları basitliği ve etkinliği nedeniyle birçok bilim dalında uygulama alanı bulmuştur. Amaca yönelik olarak uygulamada farklı yapay sinir ağı bileşenleri kullanılmıştır. Öngörü problemlerinde, genellikle ileri beslemeli yapay sinir ağı mimarileri tercih edilmiş ve başarılı sonuçlar alınmıştır. İleri beslemeli yapay sinir ağlarının bileşenleri genel olarak aşağıda verilmiştir.

Mimari Yapı**:** En basit hali ile çok tabakalı ileri beslemeli bir yapay sinir ağı mimari yapısı Şekil 1 de verilmiştir. Şekilde de görüldüğü gibi çok tabakalı ileri beslemeli bir yapay sinir ağı mimarisi üç kısımdan oluşur. Bunlar girdi tabakası, gizli tabaka (ya da tabakalar) ve çıktı tabakasıdır. Tabakalar, nöron (düğüm) adı verilen birimlerden oluşmaktadır. Yapay sinir ağlarında

nöronlar birbirlerine ağırlıklar ile bağlıdırlar. İleri beslemeli ağlarda bu bağlantılar tek yönlü ve ileri doğrudur. Aynı tabakanın birimleri arasında bağlantı yoktur.



Şekil 1. *Çok Tabakalı İleri Beslemeli Yapay Sinir Ağı*

Öğrenme Algoritması: Yapay sinir ağlarında ağırlıkların belirlenmesinde kullanılan bir çok öğrenme algoritması vardır. En yaygın kullanılan öğrenme algoritmalarından biri Geri Yayılım algoritmasıdır. Geri yayılım algoritması eldeki veri ile ağın çıktısı arasındaki farka dayalı olarak ağırlıkların güncellenmesini gerçekleştirir. Geri yayılım algoritmasında kullanılan öğrenme parametresi optimal sonuca yeterli derecede yaklaşılmasında önemli rol oynar. Öğrenme parametresi sabit olarak alınabileceği gibi, algoritma içinde dinamik olarak da güncellenebilir.

Aktivasyon Fonksiyonu: Aktivasyon fonksiyonu girdi ve çıktı birimleri arasındaki eğrisel eşleşmeyi sağlar. Aktivasyon fonksiyonunun doğru seçilmesi, ağın performansını önemli derecede etkiler. Aktivasyon fonksiyonu genelde tek kutuplu, çift kutuplu ya da doğrusal olarak seçilebilir. Seçilen aktivasyon fonksiyonu doğrusal olmadığında, eğim parametresinin belirlenmesi gerekmektedir. Eğim parametresi de optimal sonuca yeterli derecede yaklaşılmasında önemli rol oynayan bir faktördür.

VI. Diskriminant Analizi

Diskriminant Analizi, birimleri (bireyleri) en az hata ile ait oldukları kitlelere ayırmak için yapılan işlemler topluluğu olarak tanımlanabilir (Tatlıdil, 2002). Kümeleme analizi ile benzerlikler gösterdiğinden diskriminant analizi ile

kümeleme analizinin aynı amaca yönelik olduğu yanılgısına literatürde sıklıkla karşılaşılmaktadır. Gerçekte bu iki yöntem arası benzerlikler bulunsa da, küme sayısının önceden tam olarak bilinmemesi ve gelecekte kullanılabilirlik özelliği olmaması gibi nedenlerden dolayı kümeleme analiz diskriminant analizinden farklılıklar göstermektedir. Bu nedenle diskriminant analizinin temeli, incelenen bireyin kitlesinin belirlenmesini sağlayacak bir fonksiyonunun bulunmasıdır. Bu fonksiyonun bulunmasında, belirlenecek grupların ortalamaları arasındaki farklılığın maksimum olması amaçlanmaktadır (Tatlıdil, 2002).

Diskriminant Analizinin kullanım amaçlarını şu şekilde sıralayabiliriz:

i. Grup üyeliğini tahmin etmek, başka bir deyişle, bir verinin hangi değişken grubuna gireceğine kara vermek için kullanılabilir.

ii. Ayırma fonksiyon eşitliğini kullanarak, verilerin gruplara ayrılmasına yardımcı olur.

iii. Bağımsız değişkenlerin aritmetik ortalamalarının gruplar arasında nasıl değiştiğini tespit etmek için kullanılabilir.

iv. Grupları ayırmada etkili olan ve olamayan değişkenleri belirlemek için kullanılabilir.

v. Bağımlı değişkenin varyansının ne kadarının bağımsız değişkenler tarafından açıklanabildiğini belirlemek için kullanılabilir (Kalaycı, 2005).

Bütün bunların yanında diskriminant analizinde yanlış sınıflandırma ihtimalini ortadan kaldırmak için değişkenlerin çoklu normal dağılıma sahip olmaları, bütün gruplar için kovaryans matrislerinin eşit olması ve bağımsız değişkenler arasında çoklu bağlantı probleminin olmaması varsayımlarının gerçekleşmesi beklenmektedir.

VII. Yapay Sinir Ağları veya Diskriminant Analizine Dayalı Yeni Bir En İyi Küme Belirleme Yaklaşımı ve Uygulaması

Bulanık Kümelemede en iyi küme sayısının belirlenmesi, özellikle kümeler belirgin bir şekilde birbirlerinden ayrılmıyorsa daha da önem kazanmaktadır. Kararsızlık durumlarında, kümeleme indeksleri kesin kararlar vermede araştırmacıya kolaylık sağlamaktadır. Literatürdeki birçok küme geçerlilik indeksleri, karmaşık yapılar içeren verilerde küme sayıları hakkında çelişkili sonuçlar vermektedir. Bulanık kümeleme yöntemi uygulandıktan sonra her veri en yüksek üyelik derecesine sahip olduğu kümeye atanır. Bu sonuçlara göre yapılacak bir sınıflama sonucunda herhangi bir sınıflama tekniğinden yüksek doğru sınıflama yüzdesi beklenir. Sınıflama yöntemi olarak yapay sinir ağları veya diskriminant analizi kullanılırsa, sinir ağının girdisi veya diskriminant analizindeki faktörler; veri matrisimiz ve yapay sinir ağının hedef değeri veya dikriminant analizinin grup değişkeni; bulanık kümeleme sonucunda her bir verinin atandığı küme numarası olacaktır.

Herhangi bir sınıflama tekniği kullanıldığında, yüksek doğru sınıflama yüzdesi beklentisi fikrinden hareketle yapay sinir ağlarına dayalı aşağıda verilen algoritma ile bulanık kümelemede en uygun küme sayısı belirlenebilir.

Algoritma; Adım 1: Veriye uygun olabilecek en düşük ve en yüksek küme sayısına karar verilir. Belirleyeceğimiz en uygun küme sayısı bu aralıkta olacaktır. En uygun küme sayısı c_{opt}, en düşük küme sayısı c_{min} ve en yüksek küme sayısı c_{maks} ise; $c_{min} \leq c_{opt} \leq c_{maks}$ olacaktır.

Adım 2: Belirlenen aralıktaki küme sayıları için BCO yöntemi uygulanır. Sonuç olarak $c_{maks}+ c_{min}-1$ kez BCO yöntemi uygulanmaktadır.

Adım 3: Girdisi veri matrisimiz ve hedef değeri bulanık kümeleme sonucunda her bir verinin atandığı küme numarası olacak şekilde ileri beslemeli yapay sinir ağları (veya diskriminant analizi) mümkün küme sayılarının her biri için (sinir ağında çeşitli gizli tabaka birim sayılarına göre) uygulanır.

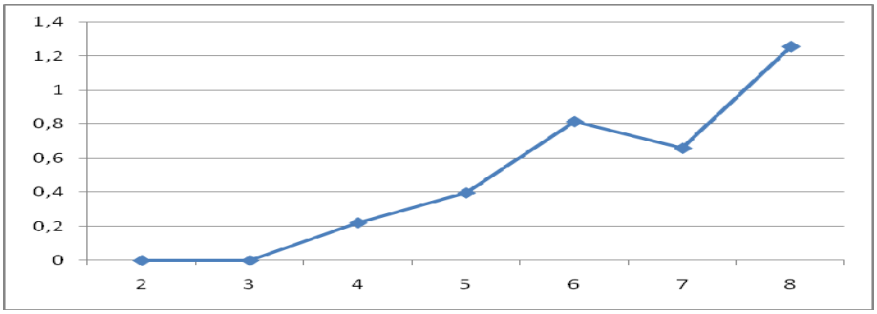
Adım 4: Her bir küme sayısı için, çeşitli gizli tabaka birim sayılarına göre yapay sinir ağlarından elde edilen RMSE (Hata kareler ortalaması karekök değeri) değerlerinin medyanı hesaplanır. Eğer diskriminant analizi yapıldıysa sınıflama hatası değerleri hesaplanır.

Adım 5: Her bir küme sayısı için elde edilen medyan değerlerinin veya sınıflama hatası grafiği çizilerek, ilk sıçramanın olduğu (RMSE medyan değerinin ilk aşırı büyüdüğü) küme sayısından bir önceki değer en uygun küme sayısı olarak belirlenir.

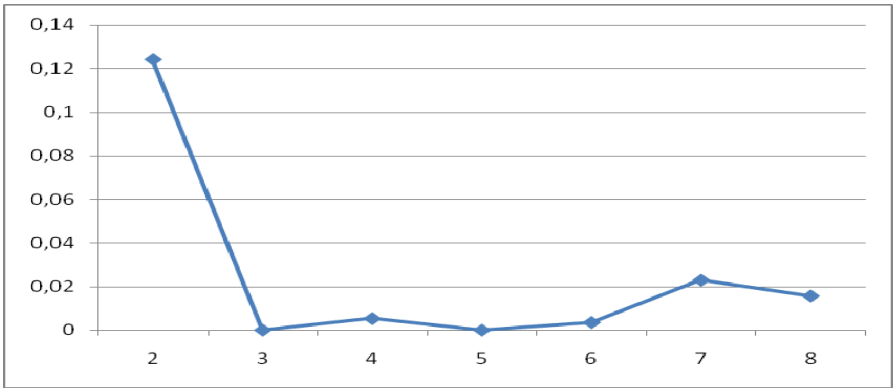
Önerilen algoritma 2 adet benzetim ve bir gerçek hayat verisine uygulanmıştır. İlk olarak gerçekte 3 kümeli olan veriye önerilen yöntem uygulanmıştır. Çalışmada kullanılan küme geçerlilik indeksleri değerleri elde edilerek sonuçlar Tablo 1 de verilmiştir. Tablo 1 incelenirse PC, CE ve XB ölçütleri için en uygun küme sayısı 3 dür. Bununla birlikte diskriminant analizinden elde edilen sınıflama hatası değerleri için de en uygun küme sayısının da 3 olduğunu söyleyebiliriz. Sınıflama hatası değerlerine baktığımızda ilk sıçrama 3 den 4 e gerçekleştiğinden küme sayısı 3 almaktayız. YSA sütununda önerdiğimiz yöntemden elde edilen sonuçlardan da ilk sıçramanın 4'de olduğu ve uygun küme sayısının 3 olduğu görülmektedir. Önerilen yaklaşımın çeşitli küme sayıları için grafiği, YSA için Şekil 2 de ve Diskriminant analizi sınıflama hatası değerleri için de şekil 3 de verilmiştir.

Tablo 1. *3 Kümeli Benzetim Verisi İçin Sonuçlar*

	Küme Geçerlilik İndeksleri				
Küme Sayısı	PC	CE	XB	YSA	Disk.
2	0,8167	0,2765	1,4454	0,000093	0,123988
3	0,9992	0,0034	49,2682	0,000198	0
4	0,8996	0,1562	43,2621	0,221739	0,005319
5	0,8808	0,1954	24,8372	0,396981	0
6	0,7817	0,3464	19,2458	0,814937	0,003546
7	0,8215	0,3078	43,596	0,658898	0,02294
8	0,743	0,4512	13,6646	1,255811	0,015676



Şekil 2. 3 *Kümeli Veri İçin YSA Yaklaşımından Elde Edilen Değerlerin Çeşitli Küme Sayılarına Göre Grafiği*

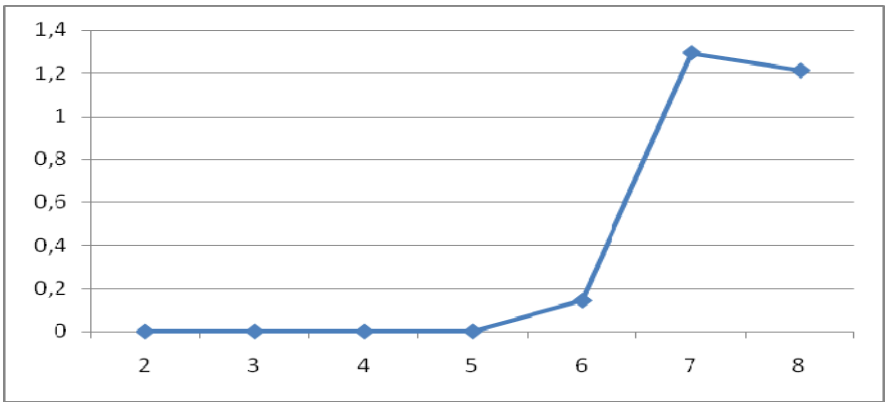


Şekil 3. *3 Kümeli Veri İçin Diskriminant Analizi Sınıflama Hatası Değerlerinin Çeşitli Küme Sayılarına Göre Grafiği*

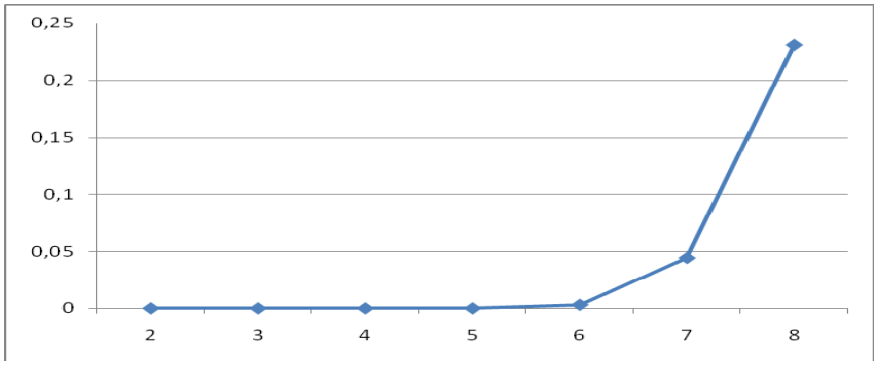
İkinci olarak gerçekte 5 kümeli olan veriye önerilen yöntem uygulanmıştır. Çalışmada kullanılan küme geçerlilik indeksleri değerleri elde edilerek sonuçlar Tablo 2 de verilmiştir. Tablo 2 incelenirse PC, CE ve XB ölçütleri için en uygun küme sayısı 5 tir. Bununla birlikte diskriminant analizinden elde edilen sınıflama hatası değerleri için de en uygun küme sayısının da 5 olduğunu söyleyebiliriz. Sınıflama hatası değerlerine baktığımızda ilk sıçrama 5 den 6 ya gerçekleştiğinden küme sayısı 5 almaktayız. YSA sütununda önerdiğimiz yöntemden elde edilen sonuçlardan da ilk sıçramanın 6'de olduğu ve uygun küme sayısının 5 olduğu görülmektedir. Önerilen yaklaşımın çeşitli küme sayıları için grafiği, YSA için Şekil 3 de ve Diskriminant analizi sınıflama hatası değerleri için de şekil 4 de verilmiştir.

Tablo 2. *5 Kümeli Benzetim Verisi İçin Sonuçlar*

	Küme Geçerlilik İndeksleri				
Küme Sayısı	PC	CE	XB	YSA	Disk.
2	0,6949	0,471	0,99	0,000094	0
3	0,9	0,2286	1,1935	0,0002	0
4	0,9265	0,1702	5,9476	0,000325	0
5	0,9999	0,00049	26,0438	0,000483	0
6	0,966	0,0555	14,7612	0,144221	0,003086
7	0,9262	0,131	25,4019	1,293433	0,044204
8	0,8677	0,2123	10,6709	1,210771	0,230596



Şekil 4. *5 Kümeli Veri İçin YSA Yaklaşımından Elde Edilen Değerlerin Çeşitli Küme Sayılarına Göre Grafiği*

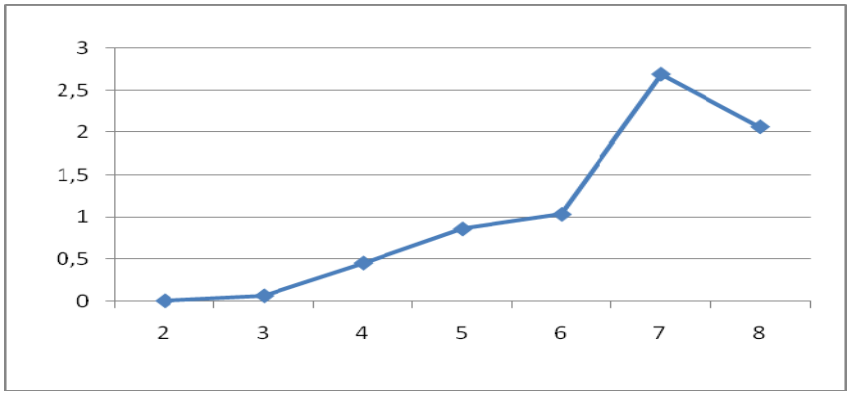


Şekil 5. 5 *Kümeli Veri İçin Diskriminant Analizi Sınıflama Hatası Değerlerinin Çeşitli Küme Sayılarına Göre Grafiği*

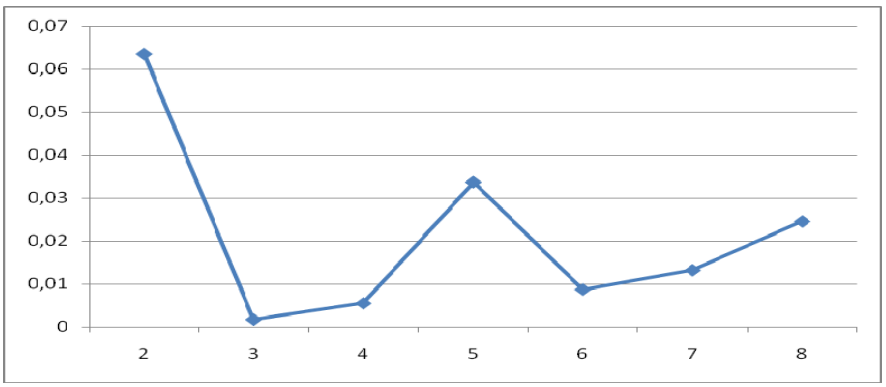
Son olarak gerçekte 3 kümeli olan Sentetik verisine önerilen yöntem uygulanmıştır. Çalışmada kullanılan küme geçerlilik indeksleri değerleri elde edilerek sonuçlar Tablo 3 de verilmiştir. Tablo 3 incelenirse PC ölçütü için en uygun küme sayısı 3 tür. CE ve XB küme geçerlilik indeksleri için küme sayısı 2 bulunmuştur. Bununla birlikte diskriminant analizinden elde edilen sınıflama hatası değerleri için de en uygun küme sayısının da 3 olduğunu söyleyebiliriz. Sınıflama hatası değerlerine baktığımızda ilk sıçrama 3 den 4 e gerçekleştiğinden küme sayısı 3 almaktayız. YSA sütununda önerdiğimiz yöntemden elde edilen sonuçlardan da ilk sıçramanın 4'de olduğu ve uygun küme sayısının 3 olduğu görülmektedir. Önerilen yaklaşımın çeşitli küme sayıları için grafiği, YSA için Şekil 5 de ve Diskriminant analizi sınıflama hatası değerleri için de şekil 6 de verilmiştir.

Tablo 3. *Sentetik Verisi İçin Sonuçlar*

	Küme Geçerlilik İndeksleri				
Küme Sayısı	PC	CE	XB	YSA	Disk.
2	0,8262	0,2864	46,3832	0,000162	0,063433
3	0,8537	0.3014	26,5862	0,059574	0,001642
4	0,7368	0,5036	20,6003	0,4495	0,005414
5	0,6781	0,6201	18,0638	0,855033	0,033604
6	0,6141	0,7652	11,821	1,028642	0,008617
7	0,5756	0,8481	11,7565	2,685111	0,013115
8	0,5539	0,9137	22,7856	2,059358	0,024461



Şekil 6. *Sentetik Verisi İçin YSA Yaklaşımından Elde Edilen Değerlerin Çeşitli Küme Sayılarına Göre Grafiği*



Şekil 7. *Sentetik Verisi İçin Diskriminant Analizi Sınıflama Hatası Değerlerinin Çeşitli Küme Sayılarına Göre Grafiği*

VIII.Sonuç ve Tartışma

Kümeleme analizinde, anlamlı ve sağlıklı sonuçlara ulaşabilmek için en uygun küme sayısının belirlenmesi önemli bir problemdir. Bazı karmaşık yapılar içeren verilerde, küme üyeliklerindeki karasızlıklar nedeniyle, küme geçerlilik indeksleri en uygun küme sayısını belirlemede birbirleri ile çelişen sonuçlar verebilmektedir. Bu çalışmada, en uygun küme sayısını belirlemede, ileri beslemeli yapay sinir ağları kullanılmış ve PC, CE gibi küme geçerlilik indekslerinden elde edilen sonuçlar ile karşılaştırılmıştır. Önerilen yöntem 2 adet benzetim ve bir gerçek hayat verisine uygulanmıştır. Benzetim verisi için elde edilen sonuçlarda hem PC, CE ve XB kriterleri hem de önerilen yöntemler en uygun küme sayısını doğru olarak tespit etmiştir. Sentetik isimli gerçek hayat verisi için ise sadece PC ölçütü ve önerdiğimiz yöntemler en uygun küme

sayısını doğru olarak belirlemiştir. Uygulamalar sonucunda, önerilen YSA'na dayalı yaklaşım ve Diskriminant analizi sınıflama hatası değerleri ile bulanık kümelemede en uygun küme sayısının belirlenebileceği görülmektedir.

Kaynaklar

Bellman R. E., Kalaba R., Zadeh L.A.,1966. Abstraction and pattern classification. J. Math. Anal. Appl., 1-7.

Bezdek J. C., 1974. Cluster validity with fuzzy sets. J. Cybern. 3, 58-73.

Erilli N.A., 2009. Kümeleme Analizine Bulanık Yaklaşım Algoritmaları ve Uygulamaları, 19 Mayıs Üniv., Yayınlanmamış Yüksek Lisans Tezi, Samsun.

Kalaycı Ş., 2005. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayıncılık, Ankara.

Naes T., Mevik T.H., 1999. The Flexibility of Fuzzy Clustering Illustred By Examples, Journal Of Chemo Metrics.

Sintas A.F., Cadenas J.M., Martin F., 1999. Membership functions in the Fuzzy c-Means Algorithm, Fuzzy Sets and Systems 101.

Şahinli F., 1999. Kümeleme Analizine Fuzzy Set Teorisi Yaklaşımı, Gazi Üniversitesi, Yayınlanmamış Yüksek Lisans Tezi, Ankara.

Tatlıdil H., 2002. Uygulamalı Çok Değişkenli İstatistiksel Analiz, Akademi Matbaası, Ankara.

Xie L., Beni G., 1991. A Validity Measure For Fuzzy Clustering, IEEE Trans. On Pattern Analysis And Machine Int. 13(4),pp 841-846.