Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, Cilt 22, Sayı1, 2013, Sayfa 189-212

ROBUST KÜMELEME YÖNTEMİ İLE GRUP SAPAN DEĞERLERİN BELİRLENMESİ

Gülsen KIRAL*
Nedret BİLLOR**
Asuman TÜRKMEN***

Özet

Gözlemlerin çoğunluğu tarafından önerilen modele uygun olmayan gözlem(ler) sapan değer olarak tanımlanır. Sapan değerler genellikle gözlemlerin çoğunluğu tarafından desteklenen bilginin yok olmasına neden olurlar. Sapan değer(ler)in belirlenmesi ile ilgili çok sayıda yaklaşım bulunmaktadır. Uygun testin seçimi verinin geldiği dağılıma, dağılım parametrelerine, beklenen sapan değerin tipine ve sayısına bağımlıdır. Bu çalışmada düşük ve yüksek boyutlu verilerde kümelemeye dayalı sapan değer belirleme yöntemi önerildi. Yöntemin etkinliği gerçek ve simüle edilmiş veri kümeleri üzerinde gösterildi.

Anahtar Kelimeler: Kümeleme, Sapan Değer belirleme, Yüksek Boyutlu Veri

Abstract

Outliers are minority observations, do not conform to the model suggested by the homogeneous majority of the observations.Outliers, causes lost of the information supported by the majority of the observations. Many approach are exist for determination of outliers. The choice of appropriate test is depend on the distribution of the data, the distribution of parameters and the type or/and the number of the outliers that expected. In this study, we proposed a new method for determining of outliers in low and high dimensional data depend on clustering.The effectiveness of the method is shown on real and simulated data sets.

Key Words: Clustering , Detecting of Outlier, High Dimensional Data

1. Giriş

Sapan değerlerin belirlenmesi çok değişkenli analizin pek çok uygulamasında önemlidir. Çünkü ayrıştırma analizi, faktör analizi, temel bileşenler analizi vb. pek çok istatistiksel yöntem sapan değer(ler)den etkilenir. Veri kümesi içerisinde yapısı önceden bilinmeyen iki ya da daha çok grubun olması durumunda sapan değerlerin belirlenmesi konusu üzerine çalışmalar yenidir.

Tek değişkenli sapan değerlerin belirlenmesi kutu ya da serpilme grafiklerinin kullanımı ile ortaya çıkarılabilirken çok değişkenli sapan değerlerin saptanması işlemi

*Yrd.Doç.Dr., Çukurova Üniversitesi, İİBF, Ekonometri Bölümü, İstatistik Anabilim Dalı, gkiral@cu.edu.tr
**Prof.Dr.; Auburn University, billone@auburn.edu
***Doç.Dr. Middle East Technical University, Department Of Geological Engıneering

çoğu zaman her bir noktanın merkeze olan uzaklığının hesabını verecek olan bir uzaklık formülünden yararlanılarak belirlenmektedir. Karesel form ile tanımlı uzaklık ölçüsü "Mahalanobis Uzaklığı" olarak bilinir. Önceden belirlenmiş olan sınır değerini geçen Mahalanobis Uzaklıklarına karşılık gelen gözlemler sapan değer olarak tanımlanırlar.

Araştırmacı veri kümesi iki ya da daha fazla grup içerisinde ki çoklu sapan değerleri bulmayı amaçlıyor ise incelemede yukarıda belirtilen uzaklık ölçüsü yanında (grupları belirlemeye yarayacak) kümeleme yapacak algoritmaların kullanımına da ihtiyaç duyacaktır. Kümeleme, veri kümeleri içerisinde birbirine benzer veri noktalarını gruplamamıza yardımcı olan bir istatistiksel tekniktir (Jain ve Dubes, 1988) ve sapan değer analizinde önemli bir araçtır.

Bu çalışmada kümelemeye dayalı yaklaşım ile tanımladığımız yöntemin etkinliği gerçek veri kümeleri üzerinde gösterilecektir. İncelememizde kümelemenin eliptik ve sapan değer belirleme yönteminin çok değişkenli normal dağılıma uygun olduğu kabul edilmektedir.

Şimdiye kadar yapılmış olan çalışmalarda çoklu sapan değer belirlemesi ile ilgili olarak çok sayıda yöntem bilinirken çok az sayıdakinin grup içindeki sapan değerlerin belirlenmesi ile ilgili olduğunu görmekteyiz.

Kümeleme ile ilgili yapılan çalışmalar: **İstatistik alanında:** Hartigan (1975), Silverman (1986), Kaufman ve Rousseeuw (1990), Scott (1992), Banfield ve Raftery (1993), Garcia-Escudero ve Gardoliza(99), Gallegos (2002), Gallegos ve Ritter (2005), Garcia-Escudero ve ark.(2008), Garcia-Escudero ve ark. (2009), **Veri Madenciliği** alanında: Ng ve Han (1994), Zhang ve ark. (1997), Bradley ve ark. (1998) ve Murtagh (2002), **Yüz tanımlama (Pattern Recognition)** alanında; Fukunaga (1990) ve Duda ve ark.(2000) olarak sayılabilir.

Bu çalışmalarda verinin kümeleme yapısı incelenmiş ancak sınıflandırılmış veri içerisindeki sapan değerlerin belirlenmesi ile ilgili bir çalışma yapılmamıştır. Pek çoğu da uygulamada pratik değillerdir. Glascock (1992) önerdiği yöntemde, kümeleme analizinde bazı optimizasyon yöntemlerini kullanmak yerine iteratif yeniden yapılandırma yöntemi uygulayarak grupları eşanlı olarak belirlemektedir. Beier ve Mommsen'nin (1994) yöntemi grup üyeliğinin belirlenmesinde Mahalanobis uzaklığını ve ki-kare ölçüsünü kullanır. Ancak gruplar eşanlı olarak değil ardışık olarak belirlenmektedir. Hardin ve Rocke (2002) tek sınıflı küme içerisinde sapan değerlerin belirlenmesi ile ilgili olarak F dağılımını kullanarak yeni bir yöntem geliştirdiler. Hardin ve Rocke (2004) daha sonra bu yöntemi bir sapan değer tanılama yöntemi ile birlikte robust kümeleme metodu olarak çoklu sınıf durumları için genişletmiştir. Yöntem Maronna (1976), Campbell (1980), Croux ve Haesbroeck (2000), Rousseeuw (1984), Davies (1987), Rousseeuw ve Leroy (1987) çalışmalarında da olduğu gibi, klasik kovaryans matrisinin robust kovaryans tahmin edicisi ile yer değiştirmesi ile tanımlanmıştır. Yöntemlerin çoğu belli sapan değere karşı duyarlı ve hesaplama problemleri içermektedir. Caroni ve Billor (2007) robust kümeleme yöntemi ile sapan değer belirleme çalışmalarında güzel sonuçlar elde ettiler. Ancak bu yöntem de diğerleri gibi sadece düşük boyutlu veri kümelerinde etkin olarak uygulanabilmekte yüksek

boyutlu verilerde çalışmamaktadır. Moh'd Belal Al-Zoubi (2009) PAM kümeleme yöntemini kullanmış ve veri kümesinden uzakta bulunan küçük grupları sapan değerler olarak tanımlamıştır.

Bu çalışmada veri kümesi içerisinde ki grup yapılarını ortaya çıkardıktan sonra her bir grupta ki gözlemleri ayrı ayrı değerlendirilip gruplardan uzakta bulunan gözlemleri sapan değer olarak tanımlayan bir yöntem önerilmiştir.

Veri kümesi içerisinde sapan değerlerin olması durumunda klasik kestiriciler kullanımı ile hesaplanan Mahalanobis Uzaklığının kullanımı uygun değildir. Doğru istatistiksel sonuçlar elde edebilmek için robust kestiricilerin kullanımı tercih edilmektedir (Rousseeuw ve Leroy, 1985; Rousseeuw ve von Zomeren, 1990; Rocke ve Woodruf, 1996; Becker ve Gather, 1999).

En popüler olarak kullanılan robust tahmin ediciler; M-tahmin edicisi (Maronna (1976), Campbell (1980)), MCD yöntemi (Rousseeuw(1984), Rousseuw ve Van Driessen (1999), Croux ve Haesbroeck (2000)), S tahmin edicisi (Davies (1987), Rousseeuw ve Leroy (1987)) olarak bilinir.

2. Yüksek Boyutlu Veri

Gözlem sayısı, parametre sayısından küçük olan veri kümeleri yüksek boyutlu veri kümeleri olarak bilinir. Son yıllarda teknolojik gelişmeler nedeni ile kemometri, bilgisayar, genetik, mühendislik vb. pek çok alanda yüksek boyutlu veri kümeleri ile karşı karşıya gelmekteyiz. Pek çok yöntem bu tip veri kümelerinde hesaplama problemi içerir ve/ya çalışmaz. Ayrıca pek çok yöntemin sapan değerlere karşı yüksek derecede hassasiyeti bulunmaktadır.

Yüksek boyutlu veri kümeleri ile işlem yapılırken genellikle iki adım takip edilir.

- Boyut indirgeme tekniklerini kullanarak yüksek boyutlu veri kümesini indirgeme,
- İndirgenmiş uzaya standart kümeleme tekniklerini uygulama (Dennis ve Lee 1999; Culhane ve ark. 2002; Ghosh 2002; Nguyen ve Rocke 2002 a,b; Hennig 2004).

2.1. Boyut indirgeme işlemi için kullanılabilecek yöntemler

2.1.1.Temel Bileşenler Analizi Yöntemi (PCA) (Klasik ya da robust kestiricilerin kullanımı ile uygulanabilir) Boyut indirgeme yöntemidir. Az sayıda bileşen yardımı ile kovaryans yapısının açıklanması için kullanılır.

2.1.2.Projection Pursuit (PP) Yöntemi: Veri kümesinin yansıtılacağı ardışık doğrultuları elde etmek için yayılım ile ilgili robust ölçüyü maksimize etmeye çalışır (Li ve Chen (1985), Croux, Ruiz Gazen, (2000), Hubert ve ark. (2002)). Yüksek hesaplama problemi içerirler ve gözlem sayısının parametre sayısından küçük olduğu yüksek boyutlu veri kümelerinde sonuç veremezler.

2.1.3. Projection Pursuit ve Robust Kovaryans tahmin edicilerin ortak kullanımı ile tanımlı yöntemler: ROBPCA; FAST–MCD (Hubert ve ark. 2005) yöntemine dayalı algoritma. PP yöntemi başlangıç boyut indirgemesi için kullanılır. Daha sonra MCD tahmin edicisine dayalı yöntem elde edilen düşük boyutlu veri uzayına uygulanır.

RBPCA; BACON (Billor ve ark. 2000) algoritmasına dayalı olarak tanımlı yöntem. Bu yöntem diğer yöntemlerle karşılaştırıldığında hem daha hızlı hem de yüksek kırılma noktasına (%25, %40) sahip olup yüksek boyutlu veri kümelerinde etkindir.

2.1.4.Küresel PCA: İlk olarak veri robust ortalama ile küre üzerine yansıtılır. Daha sonra yansıtılmış veriye PCA uygulanır.

Çoklu sapan değer belirleme yöntemleri **kullanılan tekniğe bağlı** olarak ise dört ana kategori içerisinde sınıflandırılır (Zhang ve Wang, 2006): Dağılıma dayalı, Uzaklığa dayalı, Yoğunluğa dayalı, Kümelemeye dayalı yaklaşımlar.

3. Kümeleme Analizi

3.1. En yaygın olarak kullanılan kümeleme algoritmaları:

3.1.1.k-Ortalama Tekniği ile Kümeleme: (McQuenn,1967 Hartigan ve Wong, 1979)

Gözlemlerin grup ortalamasına olan Öklidyen Uzaklıkları toplamını minimum yapmayı amaçlar. Robust bir yaklaşım değildir. Grup sayısının fazla olması durumunda etkin değildir.

3.1.2.k-Medoid Kümeleme (PAM) (Kaufman ve Rousseeuw (1990)) Grup ortalaması yerine gözlemlere ait uzaklıklar toplamını minimize edecek şekilde grupların medoidlerini bulmaya çalışır. Bu yöntem yüksek boyutlu veride etkin olarak kullanılabilmektedir.

3.1.3.Kırpılmış k-Ortalamalar Yöntemi (Later Cuesto Abertos ve ark. (1997)) k-ortalamalar yönteminin genişletilmiş bir formu olup bu yönteme göre daha robusttır. Sapan değer belirleme algoritması iki aşamada gerçekleştirilmektedir.

- Birinci aşamada k-medoid sürecini gerçekleştirme
- İkinci aşama sınıf medoidlerinden uzakta bulunan gözlemlerin ardışık olarak atılması.

Kümeleme durumlarında başlangıç veri kümesinin büyüklüğü bilinir, fakat bireysel sınıf büyüklüğü bilinmez. Bu yüzden kümeleme algoritması sapan değerler belirlenmeden önce uygulanmalıdır.

Kümeleme analizi yöntemi veri kümesi içindeki homojen grupların belirlenmesini amaçlar ancak robust olmayan kümeleme yöntemlerinin eksiklikleri nedeni ile çoğu zaman başarısız sonuçlar vermektedir. Örneğin; birbirinden farklı hareket eden farklı gruplar tek bir sınıfmış gibi görünüp bizi yanıltabilir ya da verinin yoğunlaştığı yerden

uzakta bulunan sapan değer olan gözlemler grubu sınıf olarak karşımıza çıkıp bizi yanıltabilir.

Bu nedenle kümeleme analizinde robustlığın kullanımı önemlidir. Robust kestiricilerin kullanımı sayesinde çok zararlı olabilecek gözlemler kolaylıkla ortaya çıkarılabilmekte ve yukarıda bahsedilen iki problem de kolay bir şekilde çözümlenebilmektedir.

4. Algoritmalar

Çalışmada yüksek boyutlu veride robust kümeleme yolu ile sapan değer belirlenmesi için bir yöntem önerildi.

Yöntem Caroni ve Billor (2007)'un önermiş olduğu algoritmanın yüksek boyutlu veri kümelerinde de kullanılacak şekilde yeniden uyarlanması ile tanımlanmıştır. İşlemler eşit ya da eşit olmayan varyans varsayımı altında uygulanabilen iki farklı kullanım şekli ile tanımlanmıştır.

Veri kümesine robust kümeleme algoritması uygulandıktan sonra her bir sınıf ayrı bir popülâsyondan gelmiş gibi düşünülür ve sapan değer belirleme yöntemleri bu sınıflar üzerinde bireysel olarak uygulanır.

Yöntem özellikle büyük veri kümelerinde başarılı olmamızı sağlayacak önemli bir tekniktir. İncelemede hiyerarşik olmayan kümeleme yöntemlerinden yararlanılarak k sınıfın belirlenme işlemi gerçekleştirilmektedir.

Amacımız grup yapısı ile ilgili detaylar önceden bilinmezken, grup yapısı olduğu düşünülen veri kümelerinde çoklu sapan değer belirlememize yardımcı olacak bir yöntem geliştirmektir.

Bu içerikte sapan değer olan gözlem belirlenen gruplardan uzakta bulunan gözlemler olarak tanımlıdır.

Yöntem iki aşamada gerçekleştirilir. Birinci aşamada k-medoid yöntemi veriye uygulanarak veri kümesi sınıflara ayrılır. İkinci aşamada ise sınıfların merkezinden uzakta bulunan gözlemlerin iteratif olarak ortadan kaldırılmasına ilişkin işlemleri içermektedir. Ortadan kaldırma işlemi uygun bir kritik değerinin seçimi ile yapılmaktadır.

4.1. BACON Algoritması

İki farklı şekilde kullanılabilmektedir. Birincisi hemen hemen affine equvariant, %40 a kadar varan yüksek kırılma noktasına sahip iken ikinci yaklaşım; affine equivariant, %20 lik kırılma noktasına sahiptir. Her iki yaklaşımda büyük veri kümelerinde daha az hesaplama problemi içermektedir.

4.1.1. Tam ranklı veri için BACON Algoritması (RD1-BACON)

Adım1. Klasik Mahalanobis (V1 için) ya da Öklidyen Uzaklıkların minimum değerine karşılık gelen r gözlemi başlangıç temel alt küme (x_b) içine seç

Adım 2. Temel alt kümenin ortalama ($\bar{x}_b$) ve kovaryans matrisini (S_b) kullanarak uzaklıkları hesapla.

$$d_i(\bar{x}_b, S_b) = \sqrt{(x_i - \bar{x}_b)^T S_b^{-1} (x_i - \bar{x}_b)} \quad i=1,2,..,n$$

Adım 3. $d_i < c_\alpha$ eşitsizliğini sağlayan gözlemler ile yeni temel alt kümeyi belirle. (c_α ; ki-kare dağılımından belirlenen kritik değer)
Adım4. Temel alt küme sayısı n ye eşit olana ya da temel alt küme içinde değişiklik olmayana kadar 2 ve 3 nolu adımları tekrar et. Temel olmayan alt küme içindeki gözlemleri (şayet varsa) sapan değer olarak tanımla.

4.1.2. Tam ranklı olmayan veri kümeleri için BACON Algoritması (RD2-BACON)

Tam ranklı olmayan durumlarda verinin kovaryans matrisinin tersi alınamadığından Mahalanobis Uzaklığı hesaplanamaz. Bu durumda BACON algoritmasının genişletilmiş formu olan RD2-BACON yönteminin kullanımı uygundur.

RD2-BACON yöntemi kovaryans matrisinin, ters alınamama problemini çözümlemek için kovaryans matrisinin yeniden düzenlenmesi ile tanımlıdır. Tanımlamada kovaryans matrisinin köşegen değerlerine pozitif bir sabit eklenerek shrinkage ya da ridge-tipi düzenlemesi ile tekillik problemi çözümlenir ve böylece robust uzaklık hesabında ki problem ortadan kalkar.

RD2-BACON Algoritması

Adım 1: Gözlemleri uygun bir şekilde seçilen başlangıç uzaklığına (di) bağlı olarak iki alt kümeye ayır.

$$di = \|xi - m\|$$

Başlangıç Alt küme (x_b): Minimum uzaklığa sahip c.p boyutundaki gözlem başlangıç alt kümeyi oluşturur. Bu gözlemlerin sapan değerlerden arındırılmış olduğu varsayılır. (**c**:4 ya da 5, **p**:parametre sayısı, **m**: medyan $\mathbf{x_i}$: X matrisinin i. satırı)

Temel olmayan alt küme : Geri kalan veriler

Adım 2 : Temel alt kümenin ortalama ve kovaryans matrisini hesapla

Kovaryans matrisinin hesabında aşağıdaki adımları takip et.

$$S_b = \frac{1}{n_b} \cdot \sum_{i=1}^{n_b} S(x_{ib} - \bar{x}_b) S(x_{ib} - \bar{x}_b)^T$$

x_{ib}: x_b nin i. satırı , n_b:temel alt küme gözlem sayısı
Öyleki

$$S(z) = \begin{cases} \frac{z}{\|z\|} & z \neq 0 \\ 0 & z = 0 \end{cases}$$

Adım 3: Elde edilen kovaryans matrisinin tekil değer ayrışımını (SVD) hesapla. Tekillik probleminden kurtulabilmek için bu kovaryans matrisini yeniden tanımla.

$$\Lambda^* = \Lambda \quad + \delta I_p \text{ iken} \quad \delta = \text{k. özdeğer}$$

Adım 4:

$$di = sqrt\left[(x_i - \bar{x}_b)\left(V_k(\Lambda_k{}^*)^{-1}V_k{}^T\right)(x_i - \bar{x}_b)\right] \; i = 1,2,\dots,n$$

V_k ve $\Lambda_k{}^*$ sırasıyla varyans-kovaryans matrisininin birinci özdeğer ve özvektörlerine karşılık gelen matrisler olmak üzere d_i uzaklıklarını hesapla.

Adım 5: Aşağıdaki eşitsizliği sağlayan yeni temel alt küme elemanlarını belirle

$$d_i \le med(d_i) + t_\alpha . IQR(d_i)$$

(**med** :medyan **IQR**: çeyreklik değeri , t_α : sabit $= 1.5, 2$ yada 2.5)

Adım 6 : 4. ve 5.adımlar temel alt küme büyüklüğünde değişiklik olmayana kadar tekrarla. Temel alt küme dışındaki gözlemleri (eğer varsa) sapan değer olarak belirle.

4.2. ÖNERİLEN ALGORİTMA(CBRD2)

Adım 1. k değerini uygun şekilde belirledikten sonra sınıf elemanlarını belirlemek için k-medoids yöntemini kullan.

Adım 2. Tüm gözlemlerin sınıf merkezlerine olan Öklidyen Uzaklıklarını hesapla. Gözlemleri minimum öklidyen uzaklığa göre sınıflar içine ata.

Adım 3.

Eşit varyans varsayımı altında, en küçük d_i uzaklığına sahip m^*gözlemli veri kümesini seç. m: seçilen m^* gözlemin L1 medyanı ya da spatial medyanı olmak üzere (x-m) nin kovaryans matrisini hesapla. Kovaryans matrisinin tekil değer ayrışımı (SVD) ve ardına skor matrisi hesapla.

Eşit olmayan varyans varsayımı altında, her grup içinden merkeze göre en düşük uzaklığa sahip olan m^* veri noktası her bir grup içine yerleştir. $\mathbf{m_j}$: j. sınıf içinden seçilen m^* veri noktasına ait L1 medyan ya da spatial medyan olmak üzere (x_i-m_j) nin kovaryans matrisi hesapla. Kovaryans matrisinin SVD ve skor matrislerini hesapla.

Adım 4: Skor matrisine BACON RD2 algoritması uygula ve sıralı k özdeğer ve özvektörler kullanarak Mahalanobis Uzaklıklarını hesapla. Minimum Mahalanobis uzaklıklarının tanımlanması ile her veri noktası en yakın merkez içine atanarak grupları yeniden tahsis edin.

Adım 5. Temel alt küme(ler) $d_i<C_\alpha$ i=1,2,…,n koşulunu sağlayan gözlemleri küme içine dahil ederek genişlet (C_α : Kritik değer).

Adım 6. Yukarıdaki iki adımı temel alt küme büyüklüğü değişmeyene kadar tekrarla. Temel alt küme dışında kalan gözlemleri (eğer varsa) sapan değer olarak tanımla.

Sapan değerler kritik değeri geçen robust uzaklık değerlerinin belirlenmesi ile tanımlanır. Kritik değer parametrik olmayan ölçülerden hesaplanır.

5. Grafiksel Gösterim (Uzaklık-Uzaklık (D-D) grafiği)

Robust uzaklık (RD) karşın indeks grafiği bize sapan değer olan gözlemleri göstermekte ancak bu sapan değerlerin birbirleri ile grup oluşturup oluşturmadıkları ve grup oluşması durumunda grup yapısının ne olduğu hakkında bilgi vermemektedir.

Veri içerisinde grup sapan değerlerin olduğu düşünülüyorsa sapan değerlerin yapısını araştırmak için Uzaklık–Uzaklık (D-D) grafiklerinden yararlanırız. Bu grafik sapan değerlerin çoğunluğunun ayrı bir sınıftan mı geldiği yoksa rastgele olarak mı dağıldığı hakkında bilgi vermektedir.

Uzaklık uzaklık grafiğinin çizimi için:

- Robust Mahalanobis uzaklıkları hesaplanır (RD1)
- Sapan değerleri içeren alt kümenin lokasyon ve kovaryans tahmin edicileri kullanılarak sapan değer gözlem grubu için Robust Mahalanobis Uzaklığı (RD2) hesaplanır.
- İki boyutlu uzayda RD1 karşın RD2 grafiği çizilir.

6. Uygulama

6.1. Gerçek Veri Kümeleri Uygulaması

Araştırmamızda daha önceden incelenmiş Swiss (Flury ve Riedwly 1988) ve Leukemedia veri (Golub ve ark. (1999)) kümeleri ele alındı. İncelemede R version 2.15.2 for Windows programından yararlanıldı.

Tablo1: İncelemede Kullanılacak Veri Kümelerinin Özellikleri

Veri Kümesi	#Örneklem	#Boyut	#Sınıf	Tipi
Swiss	200	6	2	Düşük Boyutlu
Leukemedia	72	500	2	Yüksek Boyutlu

6.1.1. Swiss Veri Kümesi

Swiss veri kümesinde ilk 100 gözlem Genuie Bankasından, geri kalan 100 gözlem ise Counterfeit bankasından alınan banknotlara ait bilgileri içermektedir.

x1: Banknotun uzunluğu

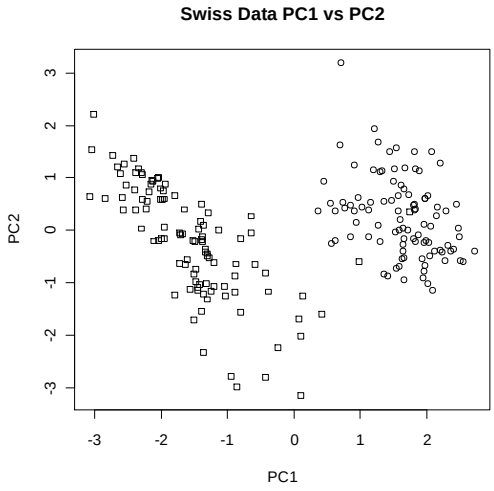
x2: Banknotun yüksekliği, soldan ölçüm,

x3: Banknotun yüksekliği, sağdan ölçüm

x4: Banknotun alt sınıra kadar ki iç çerçeve uzaklığı

x5: Banknotun üst sınıra kadar ki iç çerçeve uzaklığı

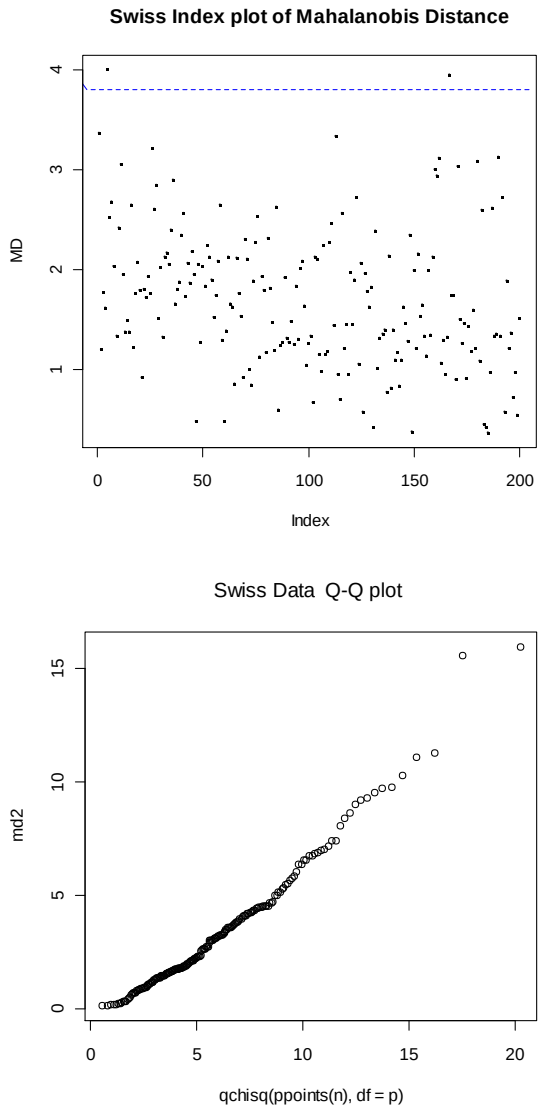
x6: Banknotun köşegen uzunluğu



Şekil 1: Swiss data 1. temel bileşen (PC1) karşın 2.temel bileşen (PC2) değerlerine ait serpilme grafiği

Swiss veri kümesinin ilk iki temel bileşenine ait serpilme grafiğine bakıldığında veri kümesi içinde 2 farklı grubun bulunduğu hemen söylenebilir. Bu nedenle de incelemede grup sayısı k=2 alınarak işlemler yapılmıştır.

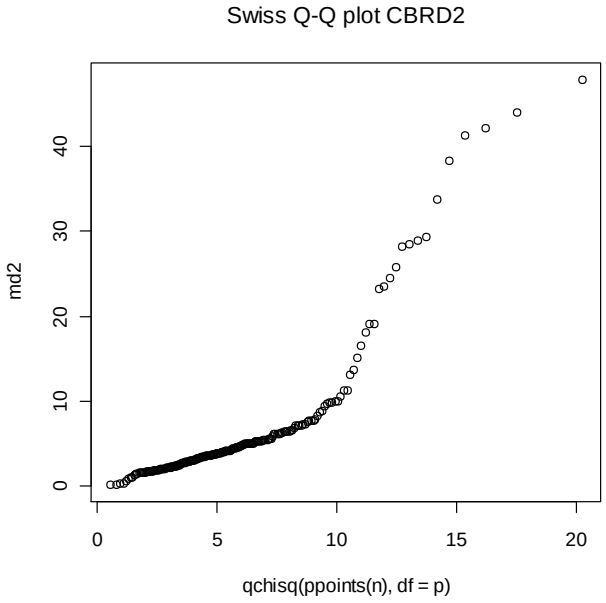
Swiss veri kümesine klasik kestiricilerin kullanımı ile tanımlı sapan değer belirleme algoritmalarını uyguladığımızda karşımıza sadece 5 ve 167 nolu gözlemler sapan değer olarak çıkmaktadır.



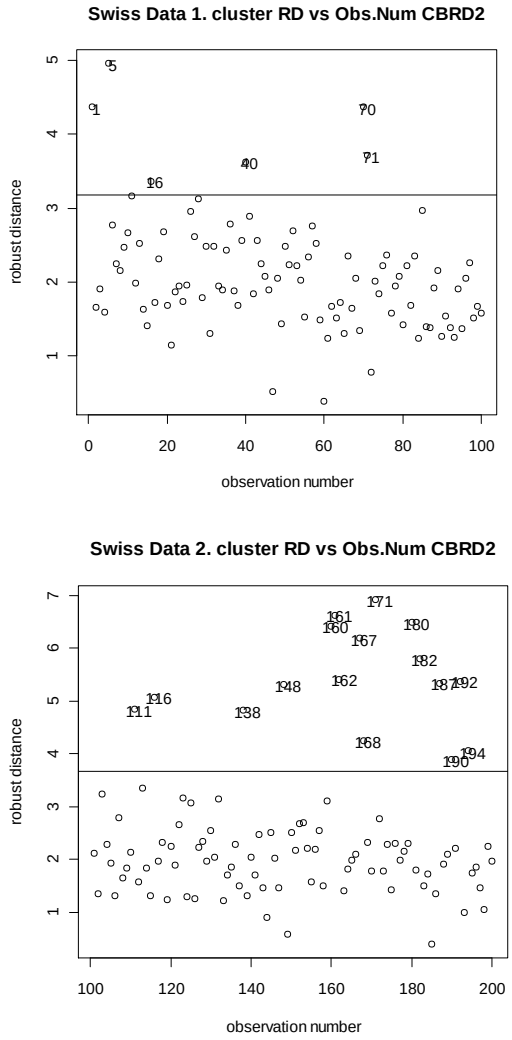
Şekil2: Klasik yöntem kullanılarak hesaplanan Swiss Data Mahalanobis Uzaklığı (MD) karşın İndeks ve Q-Q grafikleri

Önerilen yöntem için elde edilen Q-Q grafiği ve Robust uzaklık karşın gözlem numarasına ait serpilme grafikleri klasik yöntemin vermiş olduğu sonuçlardan çok farklıdır. Klasik yöntemin bulamadığı pek çok sapan değer önerilen yöntemin

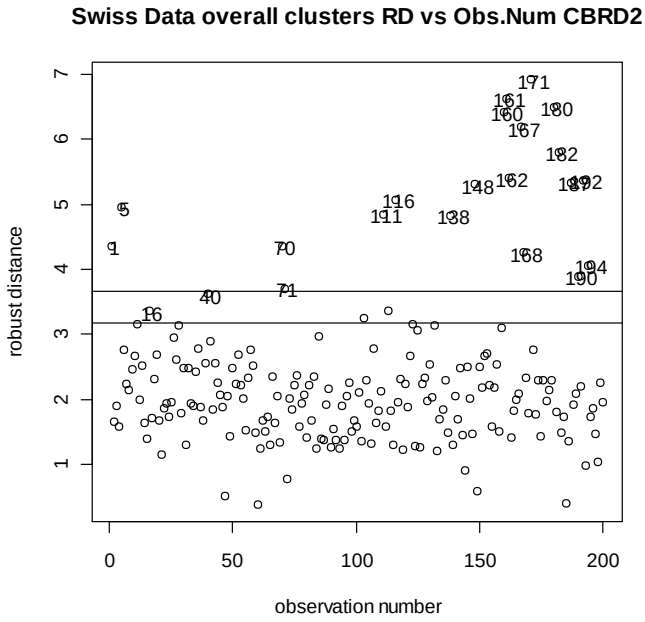
uygulanması ile kolayca belirlenebilmektedir. Elde edilen bu sonuçlar veriyi önceden inceleyen araştırmacıların sonuçları (Flury ve Riedwly, 1988) ile örtüşmektedir.



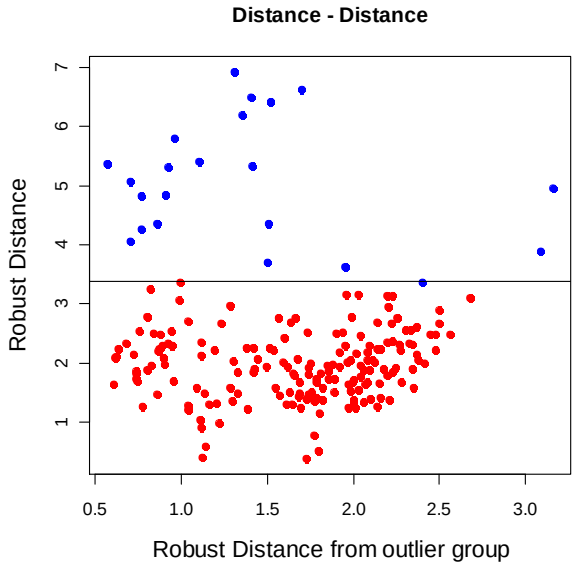
Şekil 3: Swiss veri kümesine CBRD2 yöntemi uygulandıktan sonra elde edilen Q-Q grafiği



Şekil4: Swiss veri kümesinin 1. ve 2. sınıfları için elde edilen RD karşın gözlem numarası grafikleri



Şekil5: Swiss veri kümesi RD karşın gözlem numaraları (1. ve 2 sınıf bilgisi üst üste)



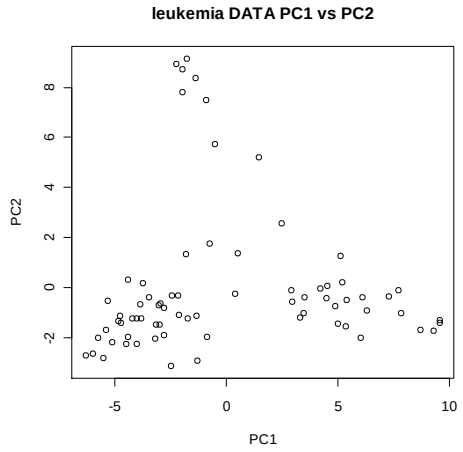
Şekil6: Swiss veri kümesine ait D-D grafiği

Tablo2: Swiss veri kümesi sonuçları

Swiss Veri Kümesi
Klasik kümeleme yöntemleri ile belirlenen sapan değerler 5 167
CBRD2 Yöntemi ile belirlenen sapan değerler 1 5 16 40 70 71 111 116 138 148 160 161 162 167 168 171 180 182 187 190 192 194

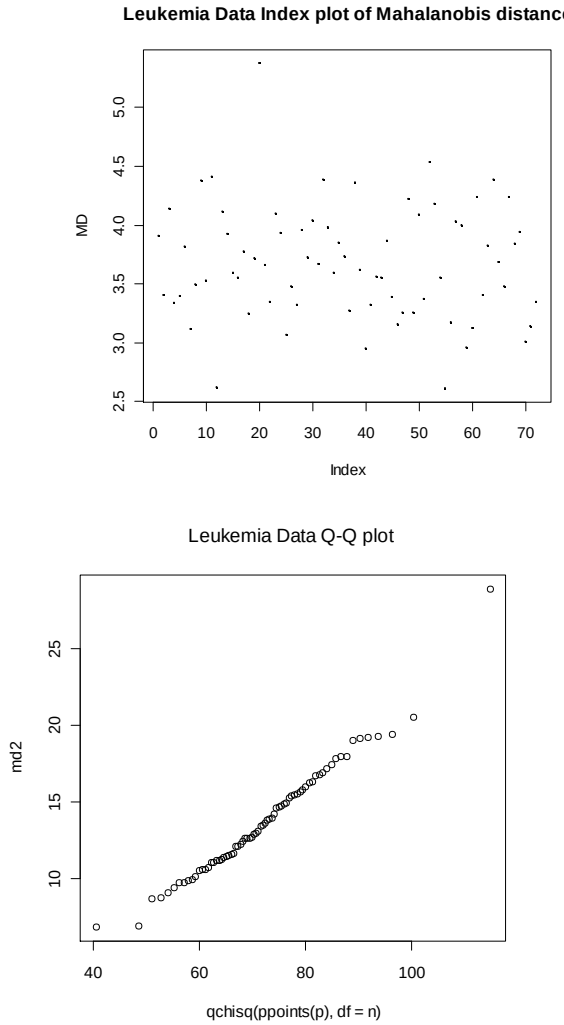
6.1.2.Leukemia Veri Kümesi

Leukemia veri kümesi; Akut Leukemia hastası olan 72 kişiden alınan 7129 genin incelenmesi ile elde edilmiş bir verdir. İncelemeye alınan 72 hastanın 34'ü Akut Leukemia (ALL), 38 i ise Akut Myelogenous Leukemia (AML) hastasıdır. İnceleme içerisinden sadece 500 gen seviyesi kullanılmıştır.



Şekil 7: Leukemia verisinin ilk iki temel bileşenine (PC1,PC2) ait serpilme grafiği

Leukemia veri kümesinenin ilk iki temel bileşeninin grafiğine (Şekil7) bakıldığında veri kümesi içinde 2 farklı grubun bulunduğu hemen söylenebilir. Bu nedenle de incelemede grup sayısı k=2 alınarak işlemler yapılmıştır.

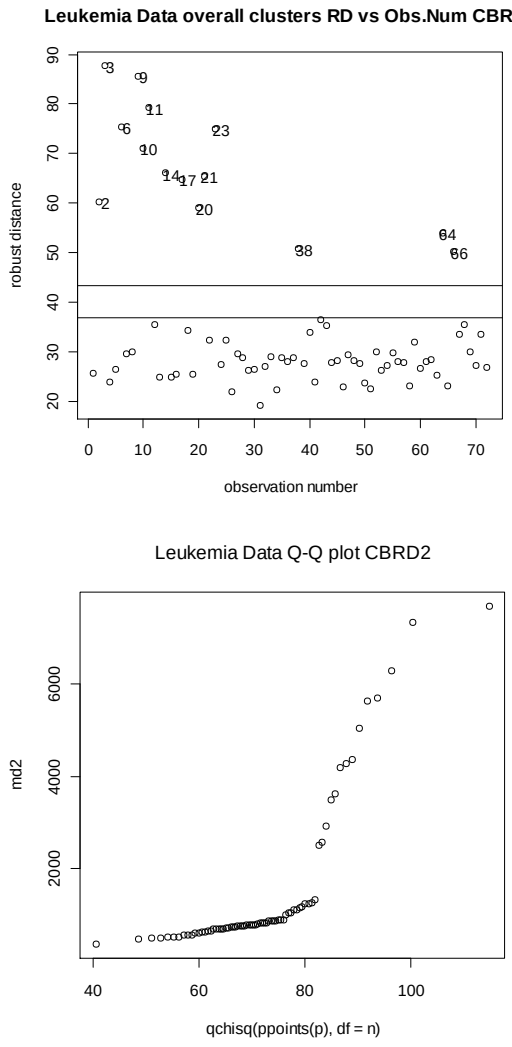


Şekil8: Klasik yöntem kullanılarak hesaplanan Leukemia Data MD karşın İndeks ve Q-Q grafikleri

Leukemia veri kümesine klasik kestiricilerin kullanımı ile tanımlı sapan değer belirleme algoritmalarını uyguladığımızda karşımıza sadece bir gözlem sapan değer olarak çıkmaktadır.

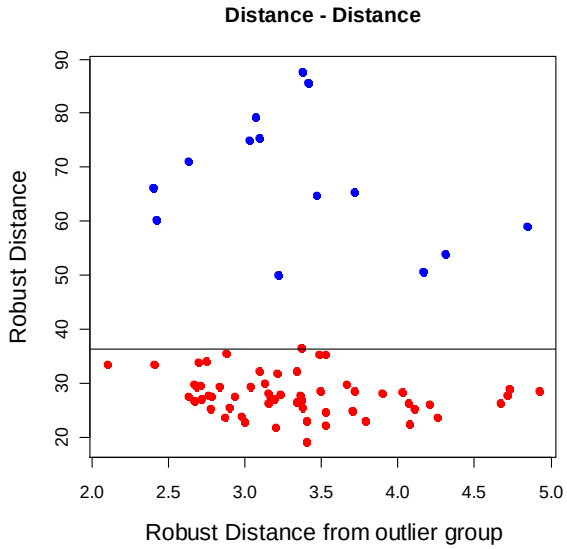
Tablo3: Leukemia veri kümesi sonuçları

Leukemia Veri Kümesi
Klasik yöntem ile bulunan sapan değerler 20
CBRD2 yöntemi ile bulunan sapan değerler 2 3 6 9 10 11 14 16 20 21 23 38 64 66



Şekil9: Leukemia Veri kümesi için CBRD2 yöntem kul.hesap.MD karşın İndeks ve Q-Q grafikleri

Önerilen yöntem için elde edilen grafiklere baktığımızda ise durumunun çok daha farklı olduğunu gerçekte birden çok sapan değer olduğunu görmekteyiz. Önerilen yönteme ait Q-Q grafiği ve Robust uzaklık karşın gözlem numarasına ait serpilme grafikleri klasik yöntemin vermiş olduğu sonuçlardan çok farklıdır. Klasik yöntemin bulamadığı pek çok sapan değer olan gözlem burada belirgin bir şekilde kendini göstermektedir.



Şekil10: Leukemia Veri Kümesine ait uzaklık uzaklık (D-D) grafiği

Grafikten sapan değer grubunun verilerin yoğunlaştığı kısımdan uzakta küçük bir sınıf oluşturduğunu görmekteyiz. İzole edilen sapan değerler grafikten rahatlıkla görülebilmektedir.

6.2.Simülasyon Çalışması

Klasik ve önerilen CBRD2 sınıflandırma yöntemini farklı konfigürasyonlar altında karşılaştırmak amacı ile yapılan simülasyon çalışmasında kullanılan parametreler ve aldığı değerler Tablo 4 de görülmektedir.

Tablo4: Simülasyonda kullanılan parametre değerleri

	Boyut	n	p	m*	c	Kont.Oranı	k
Durum 1	**n>p**	**250**	**10**	**2 ve 8**	**1ve 5**	**0, 10 ve 20**	**2 ve 3**
Durum 2	**n<p**	**50**	**250**	**2 ve 8**	**1 ve 5**	**0, 10ve 20**	**2 ve 3**

Burada **n**; gözlem sayısı, **p;** parametre sayısı, **m***;ortalama için kontaminasyon parametresi,**c**; kovaryans matrisi için kontaminasyon parametresi, **kont**; kontaminasyon oranı, **k;** küme sayısını göstermektedir. Simülasyon çalışması düşük (**n>p**) ve yüksek boyutlu (**n<p**) olmak üzere iki farklı veri tipi için yapıldı.

Veri kümeleri içerisindeki grup elemanları çok değişkenli normal dağılımdan

j temiz gözlem grubu;

$$N_{\mathbf{p}}(\mu_j , \sum_{\mathbf{j}}) \mathbf{I}_p + N_p(\mathbf{0}, 0.1*\mathbf{I})$$

j. sapan değer grubu ise

$$N_{\mathbf{p}}(\boldsymbol{\mu}_o , c\sum_{\mathbf{j}}) \mathbf{I}_p + N_{\mathbf{p}}(\mathbf{0}, 0.1*\mathbf{I}),$$

dağılımından gelecek şekilde üretildi.

$\boldsymbol{\mu}_1=(0,m^*,0,\ldots0)^T$,$\boldsymbol{\mu}_2=(m^*,0,0,\ldots,0)^T$ $\boldsymbol{\mu}_3=(-m^*,-m^*,0,\ldots,0)^T$ $\boldsymbol{\mu}_o=(m^*,m^*,\ldots,m^*)^T$ olacak şekilde tanımlandı. Ortalama için kontaminasyon parametresi m*=2 veya 8, kovaryans matrisi için kontaminasyon parametresi c=1 ya da 5 alınarak inceleme yapıldı. Kontaminasyon miktarı ise 0.10 ve 0.20 olarak seçildi.

Durum 1: Tüm kovaryans matrisleri eşit: $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma$=diagonal(1,1,…,1)

Durum 2: Tüm kovaryans matrisleri farklı

∑1=diag(c(1, m*,1,…,1))

∑2=diag(c(m*, m*,1,…,1))

∑3=diag(c(m*,m*,1,…1))+rbind(c(0,m*,0,…,0),c(m*,0,…,0)),O**mat**)

Omat ; (p-2xp) boyutunda sıfırların matrisidir. **rbind**; satır birleştirerek matris oluşturma işlemini temsil eden komut, **diag**; ise diagonal elemanlarını ilgili vektör değerleri olan diagonal bir matris tanımlama komutudur.

Yöntemleri karşılaştırma için 3 farklı ölçüden yararlanıldı.

(i)pp: doğru sapan değer belirleme yüzdesi

(pp= doğru sapan değer belirleme sayısı /toplam sapan değer sayısı)

(ii)po: yanlışlıkla sapan değer olarak belirlenen gözlem oranı

(po= gerçekte sapan değer olmamasına rağmen sapan değer olarak bulunan gözlem sayısı/temiz gözlem sayısı)

iii) missclass%: Gözlemlerin grup içlerine atanmalarında yanlış sınıflandırma oranı.

Elde edilen simülasyon sonuçları sırası ile Tablo 5 ve Tablo 6 da verildi.

Tablo5: Eşit kovaryans varsayım altında elde edilen simülasyon sonuçları

Durum 1:Eşit Kovaryans									
				Klasik Yöntem			**CBRD2**		
n,p	**k**	**m***	**Kont.oranı**	**pp**	**po**	**missclass%**	**pp**	**po**	**missclass%**
250,10	**2**	**2**	**%0**	1	0.38	0.32	0.99	0.004	0.06
			%10	0.99	0.40	0.49	0.97	0.015	0.370
			%20	0.99	0.48	0.50	0.88	0.020	0.184
		8	**%0**	1	0.54	0.42	1	0.02	0.05
			%10	1	0.71	0.50	0.99	0.014	0.053
			%20	1	0.79	0.50	0.98	0.003	0.096
	3	**2**	**%0**	0.99	0.33	0.56	0.98	0.018	0.123
			%10	0.99	0.39	0.64	0.92	0.021	0.365
			%20	0.99	0.46	0.68	0.86	0.010	0.550
		8	**%0**	1	0.5	0.45	1	0.023	0.042
			%10	0.99	0.65	0.66	1	0.019	0.059
			%20	1	0.78	0.67	0.95	0.013	0.087
50,250	**2**	**2**	**%0**	0.98	0.01	0.34	0.94	0.019	0.099
			%10	1	0.003	0.48	0.88	0.012	0.187
			%20	1	0.011	0.52	0.65	0.031	0.128
		8	**%0**	1	0.087	0.31	0.95	0.020	0.095
			%10	0.99	0.090	0.49	0.90	0.022	0.105
			%20	0.868	0.0045	0.53	0.67	0.017	0.240
	3	**2**	**%0**	0.98	0.01	0.43	0.93	0.027	0.076
			%10	0.96	0.013	0.66	0.91	0.034	0.051
			%20	0.62	0.06	0.65	0.88	0.051	0.123
		8	**%0**	1	0.08	0.26	0.98	0.043	0.065
			%10	0.96	0.008	0.62	0.94	0.025	0.093
			%20	0.94	0.13	0.65	0.92	0.082	0.086

DURUM1: **Eşit kovaryans varsayımı altında**

Klasik ve önerilen sınıflandırma yöntemi için elde edilen sonuçlar incelendiğinde önerilen yöntemin po ve miss% değerlerinin istenilene çok yakın ve klasik yönteme göre çok daha iyi değerler olduğu görülmektedir. Bu durum grup sayısının 2 olması durumunda daha da belirgin bir şekilde göze çarpmaktadır.

Düşük boyutlu veri kümelerinde: grup sayısının artması; klasik yöntemde **po** değerlerini ciddi oranda arttırırken önerilen yöntemin po değerleri önemsenmeyecek kadar az bir oranda artış göstermiştir. **Missclass%** değerleri her iki yöntemde de artmış ama en fazla artış klasik yöntemde olmuştur. Kontaminasyon oranının artması durumunda her iki yöntemde de performans kaybı yaşanmış ancak en büyük kayıp yine klasik yöntemde olmuştur. Ortamalama vektörünün kontaminasyon yapılması durumunda klasik yöntem de özellikle po ve missclass değerleri istemeyen ölçüde büyümüş ancak önerilen yöntem performası hiç değişmemiş ve düşük kontaminasyon uygulamasına hemen hemen eş değer olan benzer bir sonuçla karşımıza çıkmıştır. **Yüksek boyutlu veri kümelerinde ise;** düşük boyutlu için elde edilen sonuçlar aynen korunmaktadır. Genel olarak grup sayısının artması, kontaminasyon oranının yükselmesi klasik yöntemde kötü sonuçlar verirken önerilen yöntem çoğu şartlara karşı dayanıklı olup sonuçlarında farklılık yapmamasının yanında pp, po ve missclass% değerleri hemen hemen tüm konfigürasyonlarda iyi sonuçlar vermektedir.

Tablo6: Eşit olmayan kovaryans varsayım altında elde edilen simülasyon sonuçları

Durum2:Eşit olmayan varyans-kovaryans									
				Klasik Yöntem			**CBRD2**		
n,p	**k**	**m***	**Kont.oranı**	**pp**	**po**	**Missclass%**	**pp**	**po**	**Missclass%**
250,10	**2**	**2**	**%0**	0.99	0.52	0.40	0.94	0.003	0.034
			%10	0.99	0.73	0.508	0.866	0.007	0.049
			%20	0.99	0.78	0.607	0.857	0.006	0.073
		8	**%0**	1	0.54	0.432	1	0.048	0.056
			%10	1	0.87	0.504	1	0.052	0.062
			%20	1	0.99	0.720	0.924	0.063	0.076
	3	**2**	**%0**	0.98	0.56	0.54	0.823	0.012	0.12
			%10	0.97	0.67	0.660	0.766	0.011	0.14
			%20	0.96	0.75	0.690	0.645	0.025	0.23
		8	**%0**	0.99	0.56	0.45	1	0.041	0.05
			%10	0.98	0.82	0.670	1	0.058	0.07
			%20	0.99	0.89	0.675	0.99	0.049	0.12
50,250	**2**	**2**	**%0**	1	0.011	0.41	0.90	0.021	0.065
			%10	1	0.015	0.52	0.849	0.018	0.087
			%20	0.99	0.005	0.62	0.64	0.041	0.140
		8	**%0**	1	0.004	0.49	0.90	0.017	0.051
			%10	1	0.008	0.53	0.89	0.028	0.074
			%20	0.876	0.007	0.61	0.64	0.024	0.113
	3	**2**	**%0**	0.96	0.021	0.59	0.92	0.021	0.095
			%10	0.95	0.013	0.66	0.88	0.042	0.123
			%20	0.655	0.6	0.76	0.72	0.071	0.181
		8	**%0**	0.99	0.007	0.452	0.96	0.004	0.095
			%10	0.96	0.008	0.726	0.90	0.051	0.101
			%20	0.72	0.11	0.625	0.89	0.098	0.152

DURUM 2: Eşit olmayan kovaryans varsayımı altında

Düşük boyutlu veri kümelerinde klasik yöntemin po ve missclass% değerleri özellikle kontaminasyon oranının arttırılması durumunda çok yüksek elde edilmiş. Önerilen yöntemde herhangi bir sorun yok. Tüm ölçüm değerleri istenilene yakın. Kontaminasyon değerinin artması ve grup sayısının artması sonuçları önemsenmeyecek kadar az bir oranda arttırmış. **Yüksek boyutlu veri kümelerinde** pp ve po değerleri hemen hemen aynı düzeyde iken en büyük problemin klasik yöntemin missclass değerlerinde olduğunu söyleyebiliriz. Değerlerin hemen hepsi 0.5 den büyük çıkmıştır. Önerilen yöntemin missclass değerlerinde ciddi bir problem yok. En büyüğü 0.15 değerini almıştır. Genel olarak; simülasyon sonuçlarını inceleyecek olursak; kontaminasyon oranının ve grup sayısının artmasının yanlışlıkla sapan değer olarak belirleme oranı (po) ile yanlış sınıflandırma oranında ciddi problemlere yol açmış olduğu ancak önerilen yöntemin bu artıştan pek etkilenmediği ve pek çok durumda dayanıklı doğru sonuçlar verdiğini gözlemlenmiştir.

7. SONUÇ

Çalışmada kümeleme ve temel bileşenler analizi yardımı ile yüksek boyutlu veri kümelerinde çoklu sapan değer bulmaya yardımcı olacak yeni bir yöntem önerildi. Yöntemin performansı gerçek veri kümeleri ve simülasyon çalışması üzerinde gösterildi.

Robust kümeleme yöntemi olarak tanımladığımız bu yöntem yüksek boyutlu veri kümelerinde etkin olarak kullanılabilmektedir. Yöntem gruplandırılacak verileri, benzerliklerine göre alt sınıflara ayırarak açıklamaktadır. Analizde tüm özellikler bakımından toplu değerlendirmeye müsaade edildiğinden benzer yöntemlere göre üstünlük sağlamaktadır. Yöntem tam ranklı olmama durumunda da çalışmaktadır Veride birden fazla grup olması durumunda daha gerçekçi, dayanıklı ve kesin sonuçlar vermektedir.

Kaynaklar

- Acuna E. and Rodriguez C., (2004), A Meta Analysis Study of Outlier Detection Methods in Classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, available at academic.uprm.edu/~eacuna/paperout.pdf. In proceedings IPSI 2004, Venice
- Angiulli, F. and C. Pizzuti, Outlier Mining in Large High-Dimensional Data Sets, (2005). IEEE
- Barnett, V. & Lewis, T. (1994) Outliers in Statistical Data, 3rd edn (Chichester:Wiley).
- Billor, N., Hadi, A. S. and Velleman, P. F. (2000), "BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators," *Computational Statistics and Data Analysis*, 34, pp. 279-298.

- Kıral, G.,Billor N. , Turkmen A.(24-26 Mayıs 2012)"Robust Sınıflandırma yöntemi ile grup sapan değerlerin belirlenmesi". Eastern Mediterranean University. Gazimağusa, Kıbrıs.
- Billor N., Kıral G.,Turkmen A.S.(2012)"Clustering Based Robust Multivariate Outlier Detection" ,poster, *Joint Statistical Meetings (JSM) ,July 28-August 2, 2012 San Diego,USA*
- Breunig M. M. (2001). *"Quality Driven Database Mining"*, Ph.D. thesis, Computer Science Department, University of Munich, Munich, Germany.
- Caroni, C. and Billor, N.(2007) 'Robust Detection of Multiple Outliers in Grouped Multivariate Data', Journal of Applied Statistics, 34: 10, 1241 — 1250
- Cutsem, B and I. Gath, (1993). Detection of Outliers and Robust Estimation using Fuzzy Clustering, Computational Statistics & Data Analyses 15, pp. 47-61.
- Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics A Practical Approach*, London: Chapman and Hall.
- Gath, I and A. Geva, (1989). Fuzzy Clustering for the Estimation of the Parameters of the Components of Mixtures of Normal Distribution, Pattern Recognition Letters, 9, pp. 77-86.
- Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, Vol. 286:531-537. Kondylis, A.2006
- *Hartigan and Wong (1979) "A K-Means Clustering Algorithm" Journal of the Royal Statistical Society. Series C (Applied Statistics) Vol. 57, No. 1 - Vol. 60, No. 5.*
- Hartigan, J., (1975). Clustering algorithms. John Wiley & Sons, New York
- Hawkins, D., (1980). Identifications of Outliers, Chapman and Hall, London.
- Jiang, M., S. Tseng and C. Su, (2001). Two-phase Clustering Process for Outlier Detection, Pattern Recognition Letters, 22: 691-700.
- Kaufman, L. & Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis (NewYork: JohnWiley).
- Knorr, E. and R. Ng, Algorithms for Mining Distance-based Outliers in Large Data Sets, (1998) .Proc. the 24th International Conference on Very Large Databases (VLDB), pp. 392-403.
- Knorr, E., R. Ng, and V. Tucakov, (2000). Distance-based Outliers: Algorithms and Applications. VLDB Journal, 8(3-4): 237-253.
- Kondylis A, Hadi AS (2006) Derived components regression using the BACON algorithm.Computational Statistics and Data Analysis, 51: 556 -569
- Loureiro,A., L. Torgo and C. Soares, (2004). Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.
- MacQueen, J., (1967). Some methods for classification and analysis of multivariate observations. In:Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66). Univ.California Press, Berkeley, Calif., pp. Vol. I: Statistics, pp. 281–297.

- Papadimitriou, S., H. Kitawaga, P. Gibbons, and C. Faloutsos, (2003). LOCI: Fast outlier detection using the local correlation integral. Proc. of the International Conference on Data Engineering, pp. 315-326.
- Ramaswami, S., R. Rastogi and K. Shim, (2000). Efficient Algorithm for Mining Outliers from Large Data Sets. Proc. ACM SIGMOD, pp. 427-438
- Rousseeuw, P. and A. Leroy, (1996). Robust Regression and Outlier Detection, 3rd ed.. John Wiley & Sons.
- Transactions on Knowledge and Data Engineering, 17(2): 203-215.
- Wang, S.,Woodward,W.A., Gray, H.L.,Wiechecki, S. & Sain, S.R. (1997) A new test for outlier detection from a multivariate mixture distribution, *Journal of Computational and Graphical Statistics*, 6, pp. 285–299.
- Willems G, Joe H and Zamar R (2009). Diagnosing multivariate outliers detected by robust estimators. *J Comput Graphical Statist*, 18, 73-91.
- Zhang, J. and H. Wang, (2006). Detecting outlying subspaces for high-dimensional data: the new Task, Algorithms, and Performance, Knowledge and Information Systems, 10(3): 333-355.