Mehmet Akif Ersoy Üniversitesi
Sosyal Bilimler Enstitüsü Dergisi
ISSN: 1309-1387
Cilt: 8 Sayı: 15 2016 Haziran (s. 85-94)

TEMEL BİLEŞENLER ANALİZİ VE K-ORTALAMA KÜMELEME YÖNTEMİNİN BİRLİKTE KULLANIMI: BİR ÖRNEK UYGULAMA

COMBINED USE OF PRINCIPAL COMPONENT ANALYSIS AND K-CLUSTERING METHOD: A CASE STUDY

Nilgün ŞENGÖZ[1], Gültekin ÖZDEMİR[2]

ÖZ

Bu çalışmada, veri setlerinin kümelenmesi için kullanılan yöntemlerden biri olan K-ortalama yöntemi incelenmiştir. Buna istinaden büyük ölçekte verilen veri setlerini kümelemekte bir takım zorluklar yaşandığından ötürü boyut indirgemede yaygın olarak kullanılan Temel Bileşenler Analizi yöntemi kullanılmıştır. 3 farklı kümeye ayrılmak istenen veri seti için öncelikle, k-ortalama yöntemi uygulanmış olup, toplamdaki hata sayısı 16 olarak görülmüştür. Sonrasında temel bileşenler analizi kullanılarak boyut indirgenmiş ve böylelikle 16 olan hata sayısı 13'e düşürülmüştür.

Jel Kodu: C380, C000, C400

Anahtar Kelimeler: Temel Bileşenler Analizi, K-Ortalama, Veri Boyut İndirgeme, Sınıflandırma, Kümeleme Yöntemi

ABSTRACT

In this study, data sets, one of the methods used to cluster K-means method is studied. Consequently, the large-scale clusters of the data set are due to a number of difficulties that are widely used in dimensionality reduction of the Principal Component Analysis method is used. Divided into three different sets priorities for the desired data set, k-means method has been applied; a total of 16 in the number of errors were seen as. Size after using principal component analysis and thus reduced the number of errors decreased to 13, which is 16.

Jel Code: C380, C000, C400

Keywords: Primcipal Component Analysis, K-Means, Data Dimension Reduction,Classification , Clustering Method

[1] Uzman, Mehmet Akif Ersoy Üniversitesi, Stratejik İşbirliği Proje Danışmanlık Eğitim Uygulama ve Araştırma Merkezi,nilgunsengoz@mehmetakif.edu.tr
[2] Doç.Dr., Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği, gultekinozdemir@sdu.edu.tr

Başvuru Tarihi: 29.12.2014
Yayına Kabul tarihi: 02.06.2016
DOI: 10.20875/sb.97841

1. GİRİŞ

Günümüzde bilgisayarların yardımı ile büyük miktarda veriler toplanmakta ve dünya genelinde her yerde depolanmaktadır. Ve bu eğilim her geçen yıl daha büyük boyutlara ulaşmaktadır. Büyük terabaytlar içine sığdırılmaya çalışılan veri içerisinde aranılan ve istenen 'bilgi' ye otomatik metotlar olmadan ulaşmak neredeyse imkânsızdır. Yıllar boyunca büyük veri setlerini ayıklamak için birçok algoritma geliştirilmiş olup bunlardan bazıları sınıflandırma, kümeleme, ilişkisel kural (associational rule) vb. Bu bağlamda k-ortalama kümeleme yöntemi veri madenciliğinde en çok kullanılan denetimsiz öğrenme metodudur. *(Hartigan & Wang, 1979; Lloyd, 1957;MacQueen, 1967)* K-ortalama kümeleme analizinin temel amacı, verilen bir popülasyonu, benzer nesneler aynı grupta toplanıncaya kadar, gruplara veya kümelere parçalamaktır. Sonuç olarak, ilgi düzeyleri aynı olan yeni kategoriler keşfedilir. *(Mirkin B, 2005)* Diğer yandan yüksek boyutlu veri yığınlarının boyutlarını indirgemek için kullanılan metot ise Temel Bileşenler Analizidir. *(Jolliffe, 2002)* Temel Bileşenler Analizi yönteminin çalışma prensibi, büyük boyutlu veri setleriyle en büyük varyanslarını birlikte almasıdır. Bu denetimsiz boyut küçültme metodu çok farklı alanlarda kullanılmaktadır, örneğin; meteoroloji, görüntü işleme, gen analizleri gibi.

Bu çalışmada üç farklı buğday türüne ait geometrik şekillerine göre yedi kategorili, her bir elementten 69 adet bulunan 207 tane örnek seti mevcuttur. Sırasıyla Kama, Rosa ve Canadian diye adlandırılan bu değişik buğday türlerini ilk öncelikle temel bileşenler analizi kullanarak varyansların maksimizasyonu sağlanmış olup, sonrasında ise k-kümeleme yöntemi ile veri setlerinin kümelenmesi sağlanacaktır.

2. TEMEL BİLEŞENLER ANALİZİ

Temel Bileşenler Analizi (TBA), bir boyut azaltma işlemidir. Eğer bir dizi değişken mevcut ise (muhtemelen çok sayıda) ve değişkenlerden bazılarını fazlalık ki burada bahsedilen aynı yapı içerisinde birbirleriyle ilişkili olan değişkenler olduğuna inanılıyorsa TBA'nin kullanımına uygundur. Bu fazlalık durumundan ötürü, gözlemlenen değişkenlerdeki varyansların çoğu elde edilen değişkenlerin temel bileşenlerinin (yapay değişkenlerin) daha küçük sayıya azaltma işlemi mümkün olmaktadır. Temel bileşenler daha sonra takip eden analizlerde belirleyici ve ölçüt değişken olarak kullanılabilmektedir**.** *(Hatcher, L. (1994)*

Bu yöntem, en yalın anlamıyla veri kümesini basitleştirmek için kullanılır. Buradaki önemli nokta, elde edilen verilere doğru açıdan bakarak birbirleriyle ilişkilerini daha iyi açıklamaya çalışmaktır. Bu analiz sonucunda, *p* boyutlu uzayı çok iyi tanımlayan *p* tane yeni dik değişken (temel bileşen veya özvektör) elde edilir. Elde edilen temel bileşenlerin birimi yoktur. *p* tane değişkenin taşıdığı bilginin k tane (k<=*p*) yeni değişkenle açıklanması ise temel bileşenlerin ana amacını oluşturur (Alpar, R (2011)). Bu sistem veri seti için yeni bir koordinat sistemi seçip en büyük varyansa sahip olanı ilk eksene yerleştirir, ikinci en büyük varyansa sahip olanı ikinci eksene yerleştirerek devam eden bir süreç izler. x_i , i= 1,...,n'e kadar bir veri seti olsun. Birim vektörü *u* denilirse eğer veri setinin varyansı olarak bulunduktan sonra *u*'yu maksimize edecek şekilde yansıtılır. Genel bir kayıp olmaksızın veri seti 'sıfır ortalama' olarak kabul edilir (Jang vd 1997) ;

$$\sum_{i=1}^{n} x_i = 0$$

İç çarpım yöntemiyle x_i'nin u üzerine yansıtılmasıyla:

$$p_i = x_i.u = x_i^T u = u^T x_i$$

u birim vektörü kısıtı altında:

$$\|u\| = \sqrt{u^T u} = 1$$

x_i sıfır ortalama olduktan sonra, u ise:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} u_i^T x_i = u^T \sum_{i=1}^{n} x_i = u^T.0 = 0$$

p_i'nin karesi şu şekilde ifade edilir:

$$p_i^2 = (\mathrm{u}^T \mathrm{x}_i)(\mathrm{x}_i^T \mathrm{u}) = \mathrm{u}^T (\mathrm{x}_i \mathrm{x}_i^T)\,\mathrm{u}$$

Böylece p_i'nin varyansı ise:

$$\sigma_p^2(\mathrm{u}) = \frac{1}{n}\sum_{i=1}^{n} p_i^2$$

$$= u^T \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) u$$

$$= u^T R u$$

Burada ifade edilen simetri matrisindeki R, veri setinin korelasyon matrisidir. Lagrange çarpımı kullanılarak yeni bir amaç fonksiyonu tanımlayarak, yansıtılan varyans $\sigma_p^2(\mathrm{u})$ birim vektör kısıtı minimize edilebilir.

$$J = u^T R u + \lambda(1 - u^T u)$$

Yukarıdaki denklemle sıfıra eşitlenirse eğer:

$$\nabla_u J = 2Ru - 2\lambda u = 0$$

veya

$$Ru = \lambda u$$

R ve u_i vektörlerine tekabül eden λ korelasyon matrisinde bir özvektördür. Yukarıdaki durumu istinaden, yansıtılan varyans ise:

$$\sigma_p^2(\mathrm{u}) = u^T R u = u^T \lambda u = \lambda u^T u = \lambda$$

R korelasyon matrisinin en büyük özvektör değerine yansıtılan varyans $\sigma_p^2(\mathrm{u})$ sahiptir. Bu durum yansıtılan vektör u'nun özvektöre eşit olduğu durumda görülür.

Korelasyon matrisi simetrik ve onun özvektörleri ise birbirine diktir. R'nin n tane özvektörleri kullanılarak verilen x vektörü şöyle açıklanabilir:

$$x = \sum_{i=1}^{n} q_i u_i$$

u_i üzerine x'in yansıması $q_i (= x.u_i)$'dir ve u_i ise R'nin i'nci birim özvektörüdür (Jang, Jyh-Shing Roger. vd 1997).

3. K-ORTALAMA KÜMELEME YÖNTEMİ

K-ortalama kümeleme yöntemi ki aynı zamanda C-ortalama kümeleme yöntemi olarak da bilinmekte olup, eğiticisiz bir yöntem, yani veriler sisteme yüklendikten sonra algoritmanın çeşitli sonuçlar çıkarması olarak bilinmektedir. K-ortalama yöntemi kümelerin merkezlerini kullanarak verileri karakterize eder. Bunlar, karesi alınmış hataların toplamının en aza düşürülmesiyle belirlenir.

$$\sum_{k=1}^{K} \sum_{k, x_k \in G_i} (x_i - m_k)^2$$

Burada, $X=(x_1,..., x_n)$ veri matrisini, $m_k = \sum_{k, x_k \in G_i}^{n} x_i / n_k$ C_k kümenin merkezi ve n_k ise C_k'daki noktaların sayısıdır. (Ding ve He, 2004)

K-ortalamalar yönteminde yakınsama garanti edilene kadar iki önemli adım gerçekleştirilir; ilki tüm veri noktaları üzerinden geçmek ve onların en yakın merkezleri için bunları yeniden atamak, sonra ki adım ise kendilerine verilen puan ortalaması olarak merkezlerini yeniden hesaplamaktır. (Arthur & Vassilvitskii, 2007)

4. ÖRNEK UYGULAMA

Bu çalışmanın amacına yönelik olarak, örüntü sınıflandırma için UCI Machine Learning internet sitesinden güncel ve daha önce yapay sinir ağları üzerinde çalışılmamış olan tohum (seed) veri kümesi üzerinde üç farklı buğday türüne ait geometrik şekillerine göre yedi kategorili, her bir elementten 69 adet bulunan 207 tane örnek seti mevcuttur (). Sırasıyla Kama, Rosa ve Canadian diye adlandırılan bu değişik buğday türlerini ilk öncelikle temel bileşenler analizi kullanarak varyansların maksimizasyonu sağlanmış olup, sonrasında ise k-kümeleme yöntemi ile veri setlerinin k-ortalama yöntemi ile kümelenmiştir. Tüm bu işlemler için IBM SPSS Statistics 20 program kullanılmıştır.

Geometrik şekillerine göre ölçülmüş 7 parametre mevcuttur. Bunlar; kernel buğdayının alanı, çevresi, uzunluğu, genişliği, oyuğu, yoğunluğu ve asimetrik katsayısına göre 207 tane örnek mevcuttur. Buradaki önemli ayrıntı ise, ilk 69 veri Kama, sonraki 69 veri Rosa ve son 69 veri ise Canadian'dır.

İlk öncelikle elde olan veri setini k-ortalama yöntemiyle kümeleme analizi gerçekleştirildi. 3 tane buğday çeşidi olması nedeniyle, k-ortalama yönteminin küme sayısı 3 olarak belirlenmiştir.

Tablo 1: İlk Küme Merkezleri

	Küme		
	1	2	3
Alan	20	13	11
	,160	,200	,230
Cevre	17	13	12
	,030	,660	,630
Yogunluk	,8	,8	,8
	74	88	84
Uzunluk	6,	5,	4,
	513	236	902
Genislik	3,	3,	2,
	773	232	879
Asimetrik_Katsayisi	1,	8,	2,
	910	315	269
Oyuk	6,	5,	4,
	185	056	703

Yukarıdaki Tablo 1'de k-ortalama yöntemi kullanılarak ilk küme merkezleri hesaplanmıştır. Algoritma üzerinde yapılan denemeler neticesinde 10 tekrarın iyi sonuçlar verildiği görülmüştür. Tekrarlar sonrasında ilk merkezler ile arasındaki minimum uzaklık ise 6,470'tır. Küme merkezlerinin son durumu ise Tablo 2'de görülmektedir.

Son olarak her bir kümenin eleman sayıları Tablo 3'te verilmiştir. Burada 1. küme Rosa'ya, 2. küme Canadian'a ve 3. küme ise Kama'ya tekabül etmektedir. Her biri 69 veriden oluşmaktayken, k-ortalama yöntemine göre 1. kümede 61, ikinci kümede 74 ve 3. kümede ise 72 veri mevcuttur. Sonuç olarak, Rosa kümesi için 8 tane veriyi, Canadian kümesi için 5 tane veriyi, ve Kama kümesi için ise 3 tane veriyi doğru hesaplayamamıştır. Toplamda ise 16 tane veriyi doğru kümeleyememiştir.

Tablo 2: Son Küme Merkezleri

	Küme		
	1	2	3
Alan	18	11	14
Cevre	,722	,982	,648
Yogunluk	16	13	14
Uzunluk	,297	,281	,460
Genislik	,8	,8	,8
Asimetrik_Katsayisi	85	53	79
Oyuk	6,	5,	5,
	209	230	564
	3,	2,	3,
	723	876	278
	3,	4,	2,
	604	738	649
	6,	5,	5,
	066	088	192

Tablo 3: Her bir Kümenin Eleman Sayısı

Küme	1	1
	2	4
	3	2

Ham veriler k-ortalama yöntemiyle kümelendikten sonra, bu verilere Temel Bileşenler Analizi uygulanıldı. Tablo 4'te görüleceği üzere, yedi kategorili 207 adet veriye Temel Bileşenler Analizi uygulandıktan sonra ilk bileşen 5,019 ikinci bileşen ise 1,201 özvektör değerini almış, ve bu iki bileşen tüm veri setinin kümülatif olarak %88,846'sını açıklamaktadır.

Tablo 4: Temel Bileşenler Analizi

Bileşen	Özvektörler		
	Toplam	Varyans Yüzdesi	Kümülatif Yüzde
1	5,019	71,694	71,694
2	1,201	17,152	88,846
3	0,687	9,817	98,663
4	0,069	,983	99,646
5	0,019	,267	99,913
6	0,005	,076	99,988
7	0,001	,012	100,000

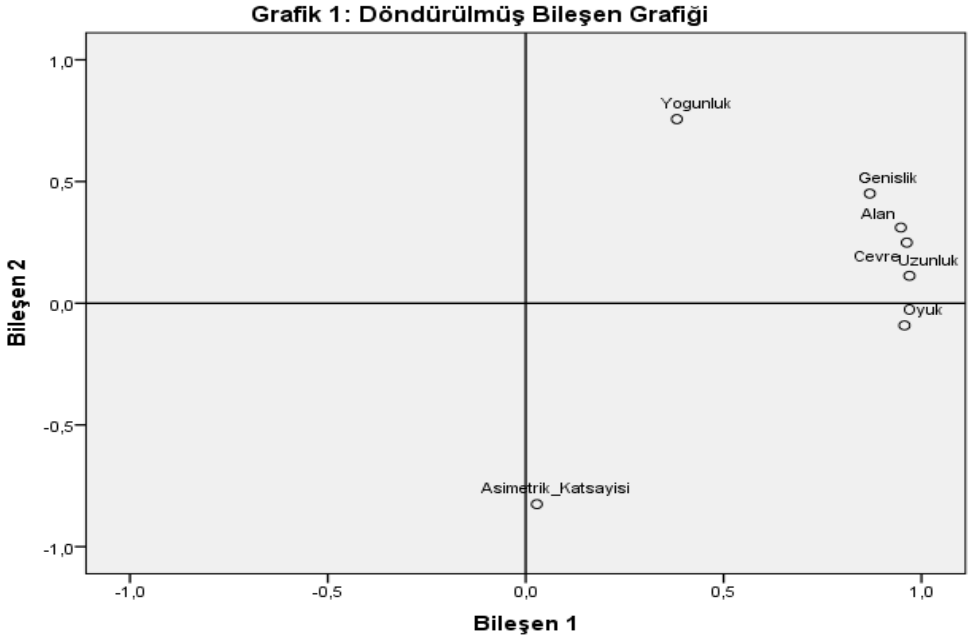
Tablo 5'te rotasyon olmadan önce her bileşenin aldığı değerler görülebilmektedir. Varimax (varyansın maksimizasyonu) yöntemine göre döndürülen bileşenlerin değeri ise Tablo 6'da gösterilmektedir.

Tablo 5: Bileşen Matrisi

	Bileşen	
	1	2
Alan	,997	,028
Cevre	,990	,091
Yogunluk	,615	-,583
Uzunluk	,951	,221
Genislik	,970	-,131
Asimetrik_Katsayisi	-,252	,786
Oyuk	,870	,409

Tablo 6: Döndürülmüş Bileşen Matrisi

	Bileşen	
	1	2
Alan	,948	,311
Cevre	,963	,249
Yogunluk	,382	,756
Uzunluk	,969	,112
Genislik	,869	,451
Asimetrik_Katsayisi	,028	-,825
Oyuk	,957	-,091



Grafik 1'de ise döndürülmüş (rotated) bileşenlerin bileşen 1 ve bileşen 2'ye olan konumları yer almaktadır. Grafikte de görüldüğü üzere; Alan, Genişlik, Çevre, Uzunluk ve Oyuk Bileşen 1'e daha yakınken, Asimetrik_Katsayısı ise Bileşen 2'ye daha yakınlık göstermektedir. Yoğunluk ise Tablo 6'ya baktığımız zaman Bileşen 2'ye daha fazla yakınlık göstermektedir.

Tablo 7: Korelasyon Matrisi

		Alan	Cevre	Yogunluk	Uzunluk	Genislik	Asimetrik_ Katsayisi	Oyuk
Korelasyon	Alan	1,000	,994	,602	,950	,970	-,217	,865
	Cevre	,994	1,000	,522	,973	,944	-,205	,892
	Yogunluk	,602	,522	1,000	,361	,758	-,321	,222
	Uzunluk	,950	,973	,361	1,000	,860	-,163	,933
	Genislik	,970	,944	,758	,860	1,000	-,246	,749
	Asimetrik_Katsayisi	-,217	-,205	-,321	-,163	-,246	1,000	-,002
	Oyuk	,865	,892	,222	,933	,749	-,002	1,000

Tablo 7'de ise her bir elementin birbirleriyle olan ilişkileri yer almaktadır. Tabloya göre, 'Alan' elementi ile 'Çevre' elementi arasında 0,994'lük bir ilişki mevcut iken, 'Alan' ile 'Asimetrik_Katsayısı' ile -0,217' lik negatif bir ilişki mevcuttur.

5. TARTIŞMA VE SONUÇ

Temel Bileşenler Analizinde ki temel felsefe, aralarındaki korelasyonu en az olan birbirlerine dik, veri setinin büyük çoğunluğunu kapsayacak yani onu ifade edecek iki bileşeni bulmaktır. Tablo 5 ve Tablo 6'ya ayrı ayrı bakıldığında bileşen 1'de en büyük değer ile bileşen 2'nin en büyük değeri Alan elementi ile Asimetrik_Katsayı elementi olarak görülebilmektedir. Diğer nazaran birbirlerine diktirler. Böylelikle aralarındaki korelasyon yoktur veya çok azdır denilebilir.

Bu nedenledir ki, k-ortalama yöntemini başta elde edilen 7 elementli veri seti yerine sadece Alan ve Asimetrik_Katsayı elementleri kullanılarak yani diğer 5 elementi göz ardı ederek oluşturulmuş olan sonuçlar aşağıdaki Tablo 8'daki görülebilmektedir.

Tablo 8: Her bir Kümenin Eleman Sayısı

	1	**62,000**
Küme	**2**	**71,000**
	3	**74,000**

Tablo 8'de açıkça görülüyor ki, temel bileşenler analizi yöntemi kullanıldıktan sonra k-ortalama yöntemi uygulanırsa eğer, doğru kümelenme sayısı artmaktadır. Tablo 3 ile Tablo 8 karşılaştırıldığında, sırasıyla 1., 2., ve 3., kümelenme sayısı 61, 74 ve 72 iken, TBA analizi sonucunda boyutu indirgendikten sonra k-ortalama yöntemi kullanıldığında sırasıyla 1., 2., ve 3., kümelenme sayısı 62, 71 ve 74 olmaktadır. Bu durumda önceden 16 tane veri yanlış kümelendirilirken, şimdi ise 13 tane veri yanlış kümelendirilmiştir. Bu da Temel bileşenler analizinin etkisini açıkça ortaya koymaktadır. Keza element sayısı fazla olduğu zaman algoritma yanlış kümelendirme yapabilmektedir. Eğer eldeki veri kümesinin boyutu küçültülüp, en önemli öznitelikler belirlenirse, bu örnekte de olduğu gibi algoritmanın daha iyi kümelendirmesi sağlanabilmektedir.

Temel Bileşenler Analizi istatistikte yaygın olarak kullanılan denetimsiz boyut küçültme tekniğidir. K-ortalama yöntemi ise denetimsiz öğrenmede yer alan veri kümeleme metodudur. Ding ve He'nin 2004 yılında yaptığı çalışma sonucu, büyük veri setlerini kümelemekte temel bileşenler analizi yöntemi yardımıyla daha iyi sonuçlar alınabilmektedir. Bu çalışmada tek yöntem kullanılmış olup, diğer sınıflandırma yöntemleri ile karşılaştırılması yöntemin etkinlik düzeyini ölçmeye yardımcı olacaktır. Bu nedenle çalışma ileride yapılacak çalışmalara örnek teşkil edebilmektedir.

6. REFERANSLAR

Alpar, Reha (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. Detay Yayıncılık, 3. baskı

Arthur, David & Vassilvitskii, Sergei (2007) *k-means++: The Advantages of Careful Seeding*

Ding, Chris, He, Xiaofeng (2004), "K-means Clustering via Principal Component Analysis", Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

Hartigan, John., & Wang, M. (1979). *A K-means clustering algorithm*. *Applied Statistics*, 28, 100–108.

Hatcher, Larry (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, Cary, *NC: The SAS Institute*. Review pp. 325-339.

Jang, Jyh-Shing Roger. Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, (1997) "*Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*"

Jolliffe, I. (2002). *Principal Component Analysis*. *Springer*. 2nd edition.

Lloyd, Sarah. (1957). *Least squares quantization in pcm*. *Bell Telephone Laboratories Paper*, Marray Hill.

MacQueen, James. (1967). Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symposium, 281–297.

Pinkowski, Brain, (1997*). Principal component analysis of speech spectrogram images*. *Pattern Recogn*, 30, 777–787.

Ramsay, James, Munhall KG, Gracco VL, Ostry DJ. (1996). *Functional data analyses of lip motion*. *J Acoust Soc Am*., 99, 3718-3727.