Spotify Müzik Parçalarını K-Means Yöntemi İle Kümeleme

Muharrem Küçükyılmaz
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
muharrem.kucukyilmaz@gmail.com

***Resumen*—Bu makalede spotify müziklerini K-Means yöntemi ile sınıflandırmak amaçlanmıştır.Korelasyon matrisine bakılarak ve PCA (Principal Component Analysis) ile özellik seçimleri yapılmıştır. İki farklı özellik kümesi için K-Means ile farklı k değerleri için kümeleme işlemi yapılmıştır. Elbow yöntemiyle de en iyi k değeri belirlenmiştir.**

***Index Terms*—K-Means, Elbow yöntemi, PCA, korelasyon matrisi**

I. Giriş

Günümüz dünyasında büyük verilerin artmasıysa birlikte bu verilerin işlenmeside zorlaşmıştır. Bu verileri en iyi şekilde kullanmak için farklı Makine Öğrenmesi modelleri geliştirilmiştir. Ancak bu modeller etiketli veri ile oluşturulmaktadır. Büyük verileri etiketlemek zor olduğu ve insan faktörüyle birlikte objektiflik sağlanamadığı için etiketsiz verileri etiketleme işlemide makinelere bırakılmıştır. Bunun için birçok farklı yöntemler geliştirilmiştir. K-means kümeleme algoritması bunlardan bir tanesidir. Bu makalede K-means kümeleme yöntemiyle spotify müziklerinin benzer müziklerin aynı kümede olduğu gruplar oluşturulmaya çalışılacaktır.

II. Sistem Tasarımı

II-A. Ön İşlemler

19 özellikten(valence, year, acousticness, artists, danceability, duration_ms, energy, explicit, id, instrumentalness, key, liveness, loudness, mode, name, popularity, release_date, speechiness, tempo) ve 170653 örnekten oluşan spotify veri kümesine aşağıdaki ön işlemler uygulanmıştır.

- herhangi bir özelliği boş olan örnekler silinmiştir.
- 'id'özelliği unqiue oldugundan silinmiştir.
- 'name', 'artists', 'release_date'kategorik özelliklerdir.Kategorik özellikler One-hot encoding yöntemi ile nümerik özelliklere dönüştürülebilirler.Bu veri kümesindeki kategorik özelliklerin farklı örnek sayısı çok fazla 'name':133638, 'artists':34088, 'release_date':11244) olduğundan ve sparse matris oluşacağından bu özelliklerde silinmiştir.
- veri kümesi 20000é düşürülmüştür.
- son olarak veri kümesi normalize edilmiştir.

Ön işlemlerden sonra veri kümesi (20000, 15) boyutundadır.

II-B. Özellik Seçimi

Model başarısını arttırmak ve işlem yükünü azaltarak daha başarılı sonuçlar elde etmek için özellik seçimi çok önemlidir.

II-B1. Korelasyon Matrisine Bakılarak Özellik Seçimi: Korelasyon matrisi, herbir özelliğin diğerlerine göre Pearson korelasyon katsayısı

$$p_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

formulü ile elde edilen matristir. 15 özellik için elde edilen korelasyon matrisi Şekil 1'deki gibidir. Şekil 1 'de

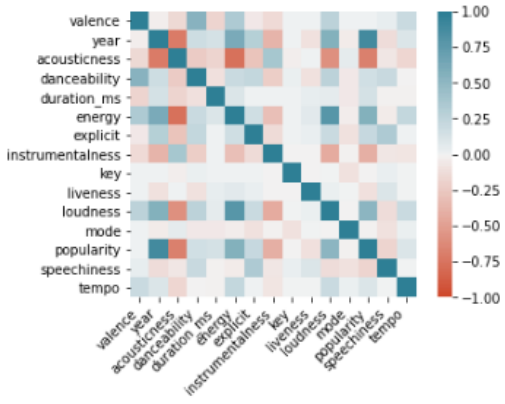

Figura 1. Özellikler arası korelasyon matrisi

görüldüğü üzere; 'year'& 'popularity', 'loudness'& 'energy', 'valence & 'danceabilityárasında doğru ilişki ve 'energy'& 'acousticness', 'acousticness'& 'year','acousticness'& 'loudness', 'acousticness'& 'popularityárasında ters ilişki olduğu için 'energy','acousticness've 'year'özellikleri de veri kümesinden çıkartılmıştır. Şekil 2 'de yeni korelasyon matrisi görselleştirilmiş ve özellikler arasında ilişki olmadığı gözlemlenmiştir. Bu yöntem ile özellik uzayı 11 özelliğe düşürülmüştür.

II-B2. PCA ile Özellik Seçimi: PCA yöntemi, matrislerin eigen değerlerine bakarak karakteristik özelliklerini elde etmek için kullanılır. Veri kümesinin özellikleri m(pca özellik sayısı)«n(veri kümesi özellik sayısı) olmak üzere m tane yeni özelliğe dönüştürülür. Böylece farklı karakteristikte yeni özellikler elde edilmiş olur. Şekil 3 'de görüldüğü gibi bu veri kümesi için iki boyut yeterlidir. Bu yöntem ile özellik uzayı 2 yeni özelliğe düşürülmüştür.

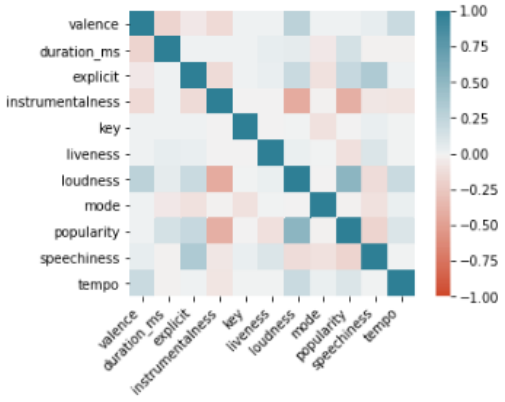

Figura 2. Azaltılmış özellikler arası korelasyon matrisi

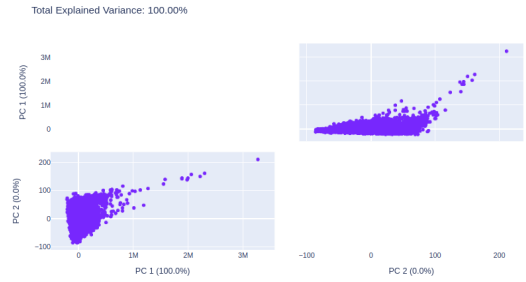

Figura 3. PCA ile Özellik Uzayı

II-C. K-Değerinin Elbow Yöntemi İle Seçimi

Elbow Yöntemi Kmeans Kümeleme algoritmasında farklı k değerleri için elde edilen WCSS(within-cluster sums of squares) değerlerinin elde edilip optimum k değerini bulmayı amaçlamaktadır. En iyi k değeri;

- başlangıç nokta: (en küçük k değeri, bu k ñın WCSS değeri)
- bitiş nokta: (en büyük k değeri, bu k ñın WCSS değeri)
- ara nokta: (ara adımdaki k değeri, bu k ñın WCSS değeri)
- a = distance(son nokta, ara nokta)
- b = distance(başlangıç nokta, ara nokta)
- c = distance(başlangıç nokta, bitiş nokta)

olmak üzere

$$h = \sqrt{b^2 - (\frac{c^2 + b^2 - a^2}{2c})^2}$$

formulü ile her bir k değeri için başlangıç ve bitiş noktaları arasına çizilen doğruya dik uzaklığı hesaplanmış ve en büyük olan optimum k olarak belirlenmiştir.Bu projede maksimum k değeri 14 seçilmiştir. Korelasyon matrisi ile özellikleri azaltılmış veri kümesi için optimum k değeri şeki4 'te görüldüğü gibi 5 ve PCA ile özellikleri azaltılmış veri kümesi için optimum k değeri 5 'te görüldüğü gibi 6 olarak belirlenmiştir.

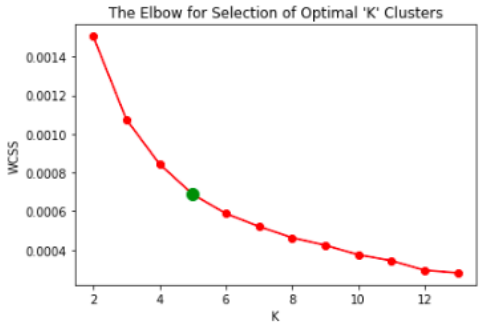

Figura 4. Korelasyon matrisi ile özellikleri azaltılmış veri kümesi için kmeans yönteminde farklı k değerleri-WCSS grafiği

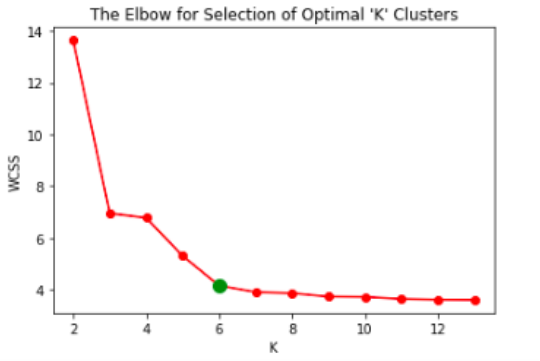

Figura 5. PCA ile özellikleri azaltılmış veri kümesi için kmeans yönteminde farklı k değerleri-WCSS grafiği

III. Deneysel Analiz

Veri Kümeleri kmeans yöntemi ile gruplandıktan sonra şekil 6 ve şekil 7 'deki PCA ile 2 boyutta görselleştirilmiştir.

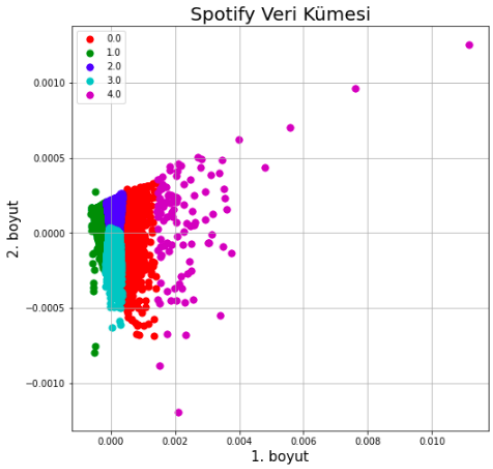

Figura 6. Korelasyon ile özellik azaltılmış veri kümesi

Korelasyon ile özellik azaltılmış veri kümesi için en iyi k değeri 5 olarak bulunmuştur ve her bir grupta bulunan şarkı sayısı tablo I 'deki gibidir.

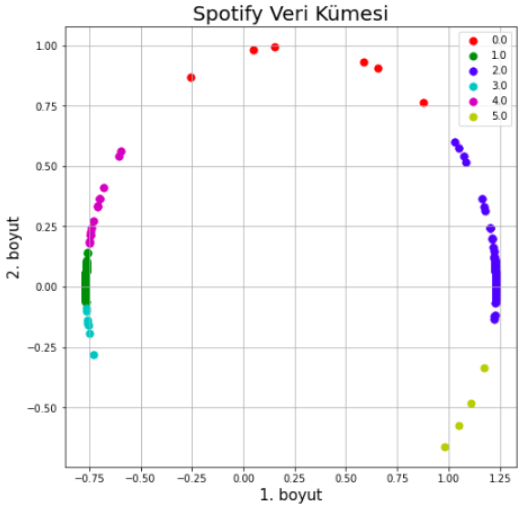

Figura 7. PCA ile yeni özellikli veri kümesi

C1	C2	C3	C4	C5
1900.	8006.	3617.	6375.	102.

Cuadro I

KORELASYON ILE ÖZELLIK AZALTILMIŞ VERI SETI IÇIN K=5 GRUPLAMA

PCA ile yeni özellikli veri kümesi için en iyi k değeri 6 olarak bulunmuştur ve her bir grupta bulunan şarkı sayısı tablo II 'deki gibidir.

C1	C2	C3	C4	C5	C6
6	12292	7677	7	14	4

Cuadro II

PCA ILE ÖZELLIK AZALTILMIŞ VERI SETI IÇIN K=6 GRUPLAMA

Tablo I incelendiğinde gruplar arası örnek dağılımının iyi olduğu gözlenmektedir. Son gruptaki örnekler Şekil 6 daki mor renklere karşılık gelmektedir. Son gruptaki örnek sayısı az olsa da varyansı yüksek olduğu için ayrı bir grup oluşturmuştur.

Tablo II incelendiğinde gruplar arası örnek dağılımının iyi olmadığı gözlemlenmektedir. 1. ve 6. gruptaki örnek(şekil 7 deki kırmızı ve sarı renkli örnekler) sayıları az olsa da diğer grup merkezlerine çok uzak ve benzerlikleri az olduğu için yeni grup oluşturmuşlardır. 4. ve 5. gruplar(7 deki açık mavi ve mor renkli örnekler) yeşil gruba dahil olabilirmiş.

IV. SONUÇ

Bu projede 20000 spotify şarkısı bir dizi ön işlemden geçirilerek kmeans ile en iyi gruplama yapılması amaçlanmıştır. Şekil 7 'dan da görüleceği üzere benzer özellikler aynı grupta olacak şekilde en iyi gruplama yapılmıştır. Bazı örnekleri gruplamada başarılı olmasa da genel olarak benzer özellikli şarkılar aynı gruptadır. Şekil 6 ya bakıldığında ise örnekler birbirine çok yakındır. Yani iyi bir gruplama yapılmış olsa bile tek bir grup olarakta değerlendirilebilirmiş.