X-MEANS KÜMELEME ALGORİTMASI WEKA UYGULAMASI

Öğr. Gör. Serpil SEVİMLİ DENİZ

Van Yüzüncü Yıl Üniversitesi, Bilgisayar Programlama Bölümü, Van, Türkiye

sdeniz@yyu.edu.tr

Prof. Dr. H. Eray ÇELİK

Van Yüzüncü Yıl Üniversitesi, Ekonometri Bölümü, Van, Türkiye

herayçelik@yyu.edu.tr

Özet

Kümeleme veri dizisinde yer alan benzer nesnelerin aynı gruplarda yer alacak biçimde ayrıştırılmasıdır. Bölümleyici küme analizi algoritmalarından k-means, kümeleme konusunda popüler olmasına rağmen, yetersiz hesaplama yapması, küme sayısının kullanıcı tarafından tanımlanması ve aramada yerel minimuma eğilimli olması açısından üç temel eksikliğe sahiptir. X-means algoritması, k-means ın her çalışmasından sonra, mevcut merkezlerin hangi alt kümelerinin, verilere daha iyi uyacak şekilde bölünmesi gerektiği konusunda yerel kararlar verebilen, k sayısını belirtilen aralıkta bulabilen alternatif bir algoritmadır. X-means ile Bayes Bilgilendirme Ölçütünü (BIC) optimize ederek kümelenme yerlerinin ve kümelenmelerin sayısını verimli bir şekilde araştıran yeni bir algoritma önerilmiştir. K-means ile yapılan çalışma Silhoutte katsayısı, x-means ile yapılan da BIC ile test edilerek x-means kümeleme algoritmasının k-means kümeleme algoritmasından daha efektif sonuçlar ürettiği görülmektedir. Bir kez çalıştırılarak kullanılan k-means algoritmsı geliştirilerek, x-means ile model seçimi için yeni bir k-means tabanlı algoritma sunulmaktadır. İstatistiksel temelli kriterler kullanan bu model üzerinde yapılan uygulamalar k-means'tan hızlı ve daha iyi performanslı olduğunu göstermektedir. X-means; küme sayısını kendi belirlemektedir. Verileri analiz ederek min/max küme sayısını belirleyebilir. Dictance metric ve veri yapısı kendisine özeldir. Nominal veri alamaz. Bu çalışmada WEKA (Waikato Environment for Knowledge Analysis) yardımıyla aynı veri setine k-means ve x-means kümeleme uygulaması yapılmıştır aynı veri setini deneysel olarak 2 den 6 ya kadar kümelere ayırarak gerekli değerleri belirledik. Matlab kullanarak, Silhuette indeksine göre en uygun sonucu 2 kümenin verdiğini gözlemledik. X-means ile aralık değerini 2-6 girerek optimum sonuca 2 küme sayısıyla varıldığını gördük. X-means sınırsız bir Gaussian EM algoritmasında bir model araştırmasını yönlendirmek için BIC'nin uygulanması olarak tanımlanabilir. Çok büyük veri setlerinde çok büyük ölçekli gözlemler kullanarak yapılan çalışmalara yardımcı olacak geniş bir algoritma sınıfı için bir fırsat sunmaktadır.

Anahtar sözcükler: Kümeleme, k-means, x-means

ABSTRACT

Clustering is the decomposition of similar objects in the data sequence to take place in the same groups. Although k-means from partitioned cluster analysis algorithms is popular in clustering, it has three fundamental deficiencies in terms of insufficient computation, user-defined number of clusters and local minimum in search. The X-means algorithm is an alternative algorithm that, after every run of the k-means, can find the number k in the specified range, giving local decisions about which subclusters of the existing centers should be better fit to fit the given value. A new algorithm has been proposed to optimize the Bayesian Informatization (BIC) with X-means and to efficiently search the number of clusters and clusters. In this study, k-means and x-means clustering were applied to the same dataset with the help of WEKA. It is seen that the work done with K-means is tested with Silhoutte coefficient, x-means with BIC, and x-means clustering algorithm produces more effective results than k-means clustering algorithm.

Keywords: Clustering , k-means, x-means

1-Giriş

Kümeleme algoritmaları, birbirine benzeyen nesnelerin bir araya gelmesini sağlayan veri madenciliği yöntemleridir (Yılmaz ve Patır 2011). Kümeleme analizi gizli örüntülerin denetimsiz öğrenme yoluyla aranmasıdır. Makine öğreniminde denetimsiz öğrenme araçlarından biri olan kümeleme analizi, nesneleri benzerlik ilişkilerine göre gruplandırması ile insan beyninin tipik bir akıl yürütme işlevini taklit etmektedir(Akpınar, 2014). Kümeleme algoritması genel olarak dört aşamalı bir süreçtir. İlk olarak veri matrisinin oluşturulması ikinci aşamada benzerlik ya da uzaklık matrislerinin oluşturulması, üçüncü aşama kümeleme yönteminin belirlenmesi ve kümelerin oluşturulması ve son olarak da sonuçların yorumlanması(Alpar, 2011). Kümeleme algoritmaları Hard ve soft algoritmalar olarak ikiye ayrılmaktadır. Hard kümelemede her nesne sadece bir kümenin üyesidir. Fuzzy mantığının bir uygulaması olan soft kümelemede ise her nesne belirli bir düzeyde birden fazla kümenin üyesi olabilir. Soft kümeleme araçları Fuzzy-C Means ve FLAME(Fuzzy clustering by Local Approximation of Memberships) önemlidir. Algoritmaların büyük bir kısmı hard sınıfınındadır. Hard sınıfındaki algoritmalar da Hiyerarşik temelli, Bölümleyici, yoğunluk temelli, ızgara temelli ve Alt uzay arama algoritmaları olarak ayırılmaktadır. Konumuz olan k-means ve x-means algoritmaları bölümleyici küme analizi algoritmalarındandır (Akpınar,2014).

2-K-means ve X-means

K-means, verileri sayısı araştırıcı tarafından belirlenen kümelere ayırmakta kullanılan bir algoritmadır. Veriler kümeler arasındaki değişkenliğin maksimum, kümeler içi değişkenliğin minimum olması amaçlanarak bölütlere ayrılır. K-means ve K-medoids araçları (Duda & Hart, 1973; Bishop, 1995) uzun zamandır kullanılan kümeleme yöntemleridir. K- means yöntemi büyük veri dizilerinin ölçeklenmesinde oldukça etkindir. Yöntem bir lokal optima noktasında son bulmaktadır. Ayrıca sıradışı değerler aritmetik ortalamayı çok kolay etkileyebileceği için bu tür durumlarda hassastır (Jiavei Han, 2006).Yöntemin en önemli sorunlarından biri küme sayısı olan k değerinin önceden belirtilmesidir. k-means kümeleme yöntemi, iteratif bir yöntemdir ve çok sayıda birimden elde edilen p değişkenli veri setlerini küme içi kareler toplamını minimize edecek biçimde k kümeye ayırmayı amaçlar. Veriler her bir iterasyonda farklı kümelere atanarak optimal çözüm permutasyonel bir yöntem kullanılarak belirlenir (Özdamar, 2004). Her bir iterasyonda olusan kümede, değişkenlerinin ortalamaları alınarak yeni küme merkezleri belirlendiği için k-means yönteminin uygulanabilmesi için veri setindeki değişkenlerin en azından aralık ölçekte olması gerekir (Bilen, 2004).k-means kümeleme yönteminin algoritması şu biçimdedir:

1. Ayırmak istedigimiz k küme sayısı belirlenir.
2. k tane veri başlangıçta küme merkezi olması amaçlanarak rastgele veya özel belirlenir.
3. Küme merkezi olmayan verilerin belirlenen uzaklık ölçülerine göre küme merkezlerine uzaklıkları hesaplanır.
4. Her veri en yakın olduğu küme merkezine atanır.
5. Yeni küme merkezleri, k adet baslangıç kümesindeki değişkenlerin ortalamaları alınarak oluşturulur.
6. Veriler belirlenen uzaklık ölçülerine göre en yakın oldukları oluşturulan yeni küme merkezlerine atanırlar.
7. Verilerin, yeni oluşturulan küme merkezlerine olan uzaklıkları öncekilerle kıyaslanır.
8. Uzaklık uygun oranda azalmış ise 6. adıma dönülür.

9. Eğer çok büyük bir değişiklik yok ise iterasyon sona erdirilir (Bilen 2004; Martinez ve Martinez 2005).

k-means kümeleme algoritmasında iterasyonun durdurulmasını gerektiren kısıtlardan birisi, kareli hata ölçüsüdür. Belirtilen ölçüt, belirlenen uzaklık ölçülerine göre, her birimin en yakın olduğu küme merkezlerine olan kareli toplam uzaklıkları (SSE Sum of Square Eroor) minimize etmeyi amaçlar. SSE' lerin küçük olması küme merkezlerinin kümeleri iyi temsil ettiginin bir göstergesidir. k-means kümeleme yönteminde baslangıç küme merkezleri genelde tesadüfi olarak seçilir. Bunun neticesinde aynı veri seti için k-means kümeleme yöntemi farklı zamanlarda uygulandıgında farklı SSE toplamları elde edilebilir. Bu yüzden başlangıç küme merkezlerini tesadüfî olarak seçmek her zaman elverisli olmayabilir.

K-means metodunun bazı eksiklikleri vardır bunlar; Birincisi, yavaştır ve her bir yinelemeyi tamamlamak için gereken süreye göre yetersiz ölçeklenir. İkincisi, kullanıcı tarafından k kümelenmelerinin sayısının tanımlanması gerekir. Üçüncüsü, sabit bir k değeriyle çalışacak şekilde sınırlandığında, dinamik olarak k 'yi değiştirebileceğinden daha kötü bir yerel optima bulabilme riski vardır. Kümeleme modeli geliştirme işinin önemli bir parçası da kümeleme algoritmalarından elde edilen kümelerin değerlendirilmesidir. Veri kümesinde küme yapısı olmasa bile kümeleme algoritmalarının veri seti içerisinde istenilen sayıda küme bulacağı bilinmektedir. Bu riske karşı kümeleme algoritmalarının sonuçlarının değerlendirilmesi için küme doğrulama (cluster validity) yöntemleri geliştirilmiştir. Bu sayede kümeleme çalısmalarında küme kalitesi ve uygun küme sayısı belirlenerek kümeleme işlemleri başarıyla tamamlanabilir(Toledo, 2005). Silhouette (S), Davies-Bouldin (DB), Dunn (D), Calinski ve Harabasz (CH), Krzanowski Lai (KL) ve Hartigan (H) küme dogruluk (cluster validity) endeksleriyle uygun küme sayısının bulunması için matematiksel sonuçlar üretilebilir. Çalışmamızda, k-means kümeleme algoritmasının sonuçlarını Silhoutte katsayısı ile yorumlayacağız. Bu katsayı hesaplanırken, ilk olarak kümeye atanan ve *i* olarak ifade edilecek bir nesnenin, aynı kümede bulunan diğer nesneler ile benzemezlik düzeyi hesaplanır ve aritmetik ortalaması alınır. Elde edilen ve *a(i)* ile gösterilen bu değer küçüldükçe, kümeleme sonuçlarının daha geçerşi olduğu sonucuna varılacaktır. *i n*esnesi ile diğerler arasındaki ortalama benzemezlik bulunur. *i* nesnesi ile ortalama benzemezliği en küçük olan kümenin ortalama benzemezlik düzeyi *b(i)* ile gösterilir ve bu küme i nesnesinin komşu kümesi olarak nitelendirilir. Bu değerlerden Silhoutte katsayısı:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} eğer & a(i) < b(i) & 1 - \frac{a_i}{b_i} \\ eğer & a(i) = b(i) & 0 \\ eğer & a(i) > b(i) & \frac{b(i)}{a(i)} - 1 \end{cases}$$

s(i) değeri -1 ile +1 arasında yer alır ve +1 değerine yaklaşması *i* nesnesinin iyi kümelendiğini, -1 e yaklaşması ise başka bir kümede olmasının daha uygun olacağı sonucunu verir. Sıfıra yaklaşması ise kararsızlıktır. Kümedeki tüm *s(i)* değerlerinin ortalaması, kümenin homojenlik düzeyini gösterecektir. Tüm nesnelerin *s(i)* değeri ise k sayısının doğru seçilip seçilmediği hakkında bilgi verir. Şimdiye kadar tanımlanan algoritma, sadece k sayısının kullanıcı tarafından sağlandığı durumlarda kullanılır. X-means algoritması optimum k sayısını nasıl belirleyebileceğimiz sorusuna yanıt verir. X-means algoritması ile k 'nın mantıklı bir aralıkta tanımlanmasını sağlanılır. Bu aralıktaki k, BIC(Bayesyian Information Criteria) gibi bir model seçim sorumlusu tarafından en iyi şekilde puanlanır. Özünde, algoritma k ile verilen aralığın alt sınırına eşit olarak başlar ve üst sınıra ulaşılana kadar ihtiyaç duyulduğu yerde merkez eklemeye devam eder. Bu süreçte, en iyi skoru kaydeden centroid seti kaydedilir.

Algoritma, tamamlanana kadar tekrarlanan işlemeler aşağıdaki şekildedir.
1. Parametreler geliştirin
2. Yapı geliştirin
3. K> K_{max} ise arama sırasında bulunan en iyi skorlama modelini bulursa durur. Bulamazsa 1. adıma geri gider
Birinci adımda yakınsama için geleneksel k-means algoritması kullanılır. 2. adımda yeni merkezlerin bulunup bulunamayacağı ve nerede bulunacağı öğrenilir. Bu, ikiye bölünmüş bazı merkezlere izin vererek elde edilir. X-means algoritmasında bölütleme için iki yaklaşım sözkonusudur. k sayısı 1 den başlayarak arttırılır her yenilendiğinde merkezi değiştirerek yeni eklemeler yapar ve sistemi test eder başarı oranı artıyorsa buna göre karar verir. Verileri yakınlık derecesine göre bölütlediğinizde k sayısını algoritma kendisi bulur. İkinci yöntem ise merkezleri bölmek şeklinde gerçekleşir. Her bir merkezden yeni merkezler çıkarılır. Verilerin yapısına göre belirlenen belli bir açı vardır. Arama alanımız olası tüm 2^k ayrıştırma konfigürasyonlarını kapsamakta ve her bir bölgede BIC'yi yerel olarak değiştirerek, optimumunu belirlemeyi amaçlamaktadır.

2.1. BIC kriteri

Schwarz Bayes Bilgi Ölçütü (BIC) bir model seçim aracıdır. BIC, aşağıdaki formüle göre verilir D ailesinden aldığımız verileri, alternatif M_j modeller k 'nın farklı değerlerine sahip çözümlere karşılık gelir. Burada modeller k-araçları tarafından üstlenilen tipin tamamıdır.

En iyi modeli seçmek için posterior olasılıklarını Pr [MjID] kullanacağız. Posteriorlara yaklaşmak için normalleşmeye kadar, Kass ve Wasserman'dan (1995) aşağıdaki formülü önermiştir.

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2}\log R$$

burada $\hat{l}_j(D)$ kümelemeye göre verilere ait log-olabilirlik olup maksimum olabilirlik noktasından alınmıştır ve p_j : M_j'deki parametrelerin sayısıdır. Bu aynı zamanda Schwarz kriteri olarak da bilinir. Aynı küresel Gauss dağılımı altında, varyans için maksimum olabilirlik tahmini (MLE) şöyledir:

$$\hat{\sigma}^2 = \frac{1}{R-K}\sum(x_i - \mu_{(i)})^2$$

Nokta olasılıkları şunlardır:

$$\hat{P}(x_i) = \frac{R_{(i)}}{R}\frac{1}{\sqrt{2\pi}\,\hat{\sigma}^M}\exp(-\frac{1}{2\hat{\sigma}^2}\left\|x_i - \mu_{(i)}\right\|^2)$$

Verilerin gerçekleşme olasılığı

$$l(D) = \log\prod_i P(x_i) = \sum_i(\log\frac{1}{\sqrt{2\pi\sigma^M}} - \frac{1}{2\sigma^2}\left\|x_i - \mu_{(i)}\right\|^2 + \log\frac{R_i}{R})$$

$1 \le n \le K$. Sadece set D_ne odaklanarak, n merkekli ve maksimum olabilirlik tahminlerine takılan noktalar;

$$\hat{l}(D_n) = -\frac{R_n}{2}\log(2\pi) - \frac{R_n.M}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n lpgR_n - R_n\log R$$

p_j serbest parametrelerinin sayısı, $K-1$ sınıfı olasılıklarının, $M.K$ sentroid koordinatlarının ve bir varyans tahmininin toplamıdır. X-means karşılaştığı en iyi modeli seçtiğinde ve aynı zamanda tüm merkezleme testleri için yerel olarak BIC formülünü kullanır.

3- Uygulama

Heritage Foundation adlı bir araştırma şirketi Wall Street Journal ile birlikte "Ekonomik Özgürlük Endeksi" isimli yıllık bir endeks yayınlamaktadır. Bu endeks iş özgürlüğü, ticaret özgürlüğü, mali özgürlük, kamu harcaması, parasal özgürlük, yatırım özgürlüğü, finansal özgürlük, mülkiyet hakları, yolsuzlukla mücadele ve işgücü özgürlüğü olmak üzere on değişken tarafından belirlenir. Endekste yer alan kriterlere göre Birleşmiş Milletlere üye olan ülkeler sıralanmaktadır. Bu çalışmada, OECD üyesi ülkelerin 2018 yılına ait ekonomik özgürlük endeksinde yer alan veriler kullanılmıştır. Kümeleme analizleri WEKA (Waikato Environment for Knowledge Analysis) yardımıyla yapılmış olup, silhuette indexi hesaplamaları için MATLAB kullanılmıştır.

x-means sonuçları;

Küme sayısı	***İterasyon sayısı ve Hata kareler ortalaması***	***Kümeler***	***Silhuette endeks değeri***
K=2	Number of iterations: 10 Within cluster sum of squared errors: 53.78410117805115	Time taken to build model (full training data) : 0.01 seconds === Model and evaluation on training set === Clustered Instances 0 65 (36%) 1 115 (64%)	0.3229
K=3	Number of iterations: 12 Within cluster sum of squared errors: 45.97131956468816	Time taken to build model (full training data) : 0 seconds === Model and evaluation on training set === Clustered Instances 0 26 (14%) 1 79 (44%) 2 75 (42%)	0.2253
K=4	Number of iterations: 8 Within cluster sum of squared errors: 36.57792251577094	Time taken to build model (full training data) : 0 seconds === Model and evaluation on training set === Clustered Instances 0 25 (14%) 1 57 (32%) 2 58 (32%) 3 40 (22%)	0.1883
K=5	Number of iterations: 5 Within cluster sum of squared errors: 33.11289406475867	Time taken to build model (full training data) : 0 seconds === Model and evaluation on training set === Clustered Instances 0 25 (14%) 1 50 (28%) 2 51 (28%) 3 18 (10%) 4 36 (20%)	0.2068
k=6	Number of iterations: 8 Within cluster sum of squared errors: 31.76595165337176	Time taken to build model (full training data) : 0.01 seconds === Model and evaluation on training set === Clustered Instances 0 8 (4%) 1 52 (29%) 2 49 (27%) 3 18 (10%) 4 36 (20%) 5 17 (9%)	0.2124


```
XMeans
======
Requested iterations          : 1
Iterations performed          : 1
Splits prepared               : 2
Splits performed              : 1
Cutoff factor                 : 0.5
Percentage of splits accepted
by cutoff factor              : 100 %
------
Cutoff factor                 : 0.5
------

Cluster centers               : 2 centers

Cluster 0
          40.668695652173916 36.494782608695665 32.31565217391306 77.77130434782609 69.15130434782607 60.562608695652166 72.80347826086954 71.1304347826087 47.17391304347826 39.21739130434783
Cluster 1
          72.66923076923077 67.4476923076923 60.86307692307693 74.55846153846153 54.56307692307691 79.81076923076925 82.46923076923076 84.48000000000002 76.46153846153847 65.07692307692308

Distortion: 92.448707
BIC-Value : 159.869395


Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      115 ( 64%)
1       65 ( 36%)
```


4- Tartışma ve Sonuç

Bir kez çalıştırılarak kullanılan k-means algoritmsı geliştirilerek, x-means ile model seçimi için yeni bir k-means tabanlı algoritma sunulmaktadır. İstatistiksel temelli kriterler kullanan bu model üzerinde yapılan uygulamalar k-means'tan hızlı ve daha iyi performanslı olduğunu göstermektedir. X-means; küme sayısını kendi belirlemektedir. Verileri analiz ederek min/max küme sayısını belirleyebilir. Dictance metric ve veri yapısı kendisine özeldir. Nominal veri alamaz. Aynı veri setini deneysel olarak 2 den 6 ya kadar kümelere ayırarak gerekli değerleri belirledik. Silhuette indeksine göre en uygun sonucu 2 kümenin verdiğini gözlemledik. X-means ile aralık değerini 2-6 girerek optimum sonuca 2 küme sayısıyla varıldığını gördük. x-means sınırsız bir Gaussian EM algoritmasında bir model araştırmasını yönlendirmek için BIC'nin uygulanması olarak tanımlanabilir. Çok büyük veri setlerinde çok büyük ölçekli gözlemler kullanarak yapılan çalışmalara yardımcı olacak geniş bir algoritma sınıfı için bir fırsat sunmaktadır.

5- Kaynaklar

Alpar, R. (2011). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Üçüncü Baskı, Detay Yayıncılık, Ankara.

Alparslan, E. (2015). Veri Madenciliği Kümeleme Algoritmaları ve Model Değerlendirme. Erişim tarihi 08.05.2015

Bilen, Ö. (2004), ÖSS Sınav Sonuçlarının Okul Bazında Veri Madenciligi ile incelenmesi, Yüksek Lisans Tezi, FEN Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi, İstanbul.

D. Pelleg, and A. Moore: Accelerating exact X-means algorithms with geometric reasoning, *KDD-99*, 1999, 277-281.

D. Pelleg, and A. Moore, _-means: Extending X- means with efficient estimation of the number of clusters, *17th International Conf. on Machine Learning*, 2000, 727–734

Han, J. and Kamber, M. (2001). Data Mining: ConceptandTechniques. USA: Morgan Kaufmann Publishers.

Kuo, R.J.,Ho, L.M. and Hu, C.M. (2002). Integration of Self- OrganizingFeatureMapand K-meansAlgorithmFor Market Segmentation, Computers&OperationsResearch, Vol: 29, Issue:11.

Martinez, W.L. ve Martinez A. R. (2005), Exploratory Data Analysis with MATLAB, Boca Raton: CRC Press, USA.

Özdamar, K. (2004), Paket Programlar ile İstatistiksel Veri Analizi 2, Kaan Kitabevi, Eskisehir.

Toledo, M.D.G. (2005), A Comparison in Cluster Validation Techniques, Yüksek Lisans Tezi, University of Puerto Rico Matematics Department, Puerto Rico

Yılmaz, Ş., Patır, S. "Kümeleme Analizi ve Pazarlamada Kullanımı", Akademik Yaklaşımlar Dergisi, 2(1): 91-113, 2011.

Zontul, M., Kaynar, O. ve Bircan, H., (2004), "SOM Tipimde Yapay Sinir Aglarını Kullanarak Türkiye' nin _thalat Yaptıgı Ülkelerin Kümelenmesi Üzerine Bir Çalısma",Cumhuriyet Üniversitesi, iktisadi ve _dari Bilimler Dergisi, 5(2):47-68.