Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi

Bölümleyici Kümeleme için Doğru Merkezi Noktaların Tayini

Determining Accurate Centroids in Partitional Clustering

Sibel Tarıyan Özyer

Ankara Medipol Üniversitesi, Bilgisayar Mühendisliği Bölümü, sibel.tariyan@ankaramedipol.edu.tr
ORCID: https://orcid.org/0000-0002-0312-9016

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 30 Ağustos 2023
Revizyon 17 Ekim 2023
Kabul 21 Ekim 2023
Online 30 Haziran 2024

Anahtar Kelimeler:

Fikir birliği, kümeleme, topluluk

ÖZ

Kümeleme, benzer nesneleri aynı kümede ve farklı nesneleri ayrı kümelerde sınıflandırmak için etiketlenmemiş veriler üzerine odaklanan denetimsiz bir veri madenciliği tekniğidir. Hemen hemen her alanda kullanılmaktadır. Özellikle bölümlü kümelemede kümelemedeki temel sorun, verilerin doğası ve küme sayısı hakkında herhangi bir bilgi olmadan, iyi ayrılmış doğal kümelerin elde edilmesidir. Farklı kümeleme süreçleri sonucunda birçok küme elde edilir. Fikir birliği kümelemesinden sonra daha doğru nihai kümeleme çözümüne ulaşılır.

Farklı yaklaşımlarla çeşitli kümeleme sonuçları elde edilebilmektedir. kümeleme algoritmaları, farklı performanslarla sonuç elde edilmesine yol açar. Bu çalışmada farklı çözümlerin daha iyi tek bir çözüme indirgendiği fikir birliği kümeleme sonucui sunulmuştur. Elde edilen sonuçlara dayanarak kıyaslama yapılmıştır. Geleneksel bölümsel kümeleme algoritması ve çeşitleriyle çalışmamıza temel oluşturan bir diğer çalışma birlikte incelenmiştir. Farklı fikir birliği kriterleri ile gerçekleştirilen kümeleme sonuçları çeşitli veri setleri üzerinde uygulanarak incelenmiştir. Doğal kümeleme sonucu olarak küme sayısı değeri de belirlenmeye çalışılmıştır. Son olarak sonuçlarımızın k-ortalama algoritmasından daha iyi performansla çalıştığı gösterilmiştir.

ARTICLE INFO

Article history:

Received 30 August 2023
Received in revised form 17 October 2023
Accepted 21 October 2023
Available online 30 June 2024

Keywords:

consensus, clustering, ensemble

ABSTRACT

Clustering is an unsupervised data mining technique dealing with unlabeled data to classify similar objects in the same cluster and different objects in separate clusters. It is almost used in all the fields. Particularly in partitional clustering, the fundamental problem in clustering is obtaining the well separated natural clusters without any knowledge of the inherent nature of the data as well as the number of clusters. Partitional clustering algorithms with different linkage options are available.

Different clustering processes provide many clusters. They boil down to better final clustering solution after the consensus clustering. They suggest several options for conducting experiments. In this study, several consensus clustering solutions have been presented and compared with each other. Traditional partitional clustering algorithm and another study that forms the basis of our study with varieties have been investigated. Different datasets with varying consensus criteria have been examined. Number of clusters value after getting the natural clusters has also been determined. Finally our results outperform the k-means algorithm.

Doi: 10.24012/dumf.1352625

* Sorumlu Yazar

Giriş

Kümeleme, benzer nesneleri kümeler olarak bilinen farklı gruplar halinde sınıflandırmak için denetimsiz bir öğrenme tekniğidir. Bu teknik farklı alanlardaki verilerin öbeklere ayrıştırılması için kullanılmaktadır. Kümeleme algoritmaları ihtiyaca, veri formatına göre farklılıklar gösterebilir.

Kümeleme algoritmaları bilgiye erişim, tıp bilimi, imalat, robot bilimi, finans, mahremiyetin korunması, yapay zeka, kentsel gelişim, havacılık, endüstriler, satış ve pazarlama gibi hemen hemen bir çok alanda kullanılmaktadır[4], [9], [18], [21]. Kümelemede daha iyi

bir servis ve daha fazla kar elde etme amacı öne çıkmaktadır. Hali hazırda, çok fazla uygulamaya sahip olmasına rağmen, tek bir evrensel algoritmanın tüm veri türlerini etkili bir şekilde sınıflandıramaması nedeniyle yeni kümeleme algoritmaları ortaya çıkmaya devam ediyor. Kümeleme algoritmaları, her bir örneğin özelliklerine veya karakteristiklerine dayalı olarak bir veri kümesi içindeki doğal grup sayısını keşfetmeye çalışır. İlk önce gerçek k değeri olarak adlandırılan doğru grup sayısını bilmeden, veri kümelerinden en uygun kümeleri elde etmek inanılmaz derecede zordur.

Küme toplulukları, bireysel kümeleme algoritmalarına göre daha fazla kararlılığa sahip, umut verici bir alternatif olarak ortaya çıkmıştır. Çalışmalar, rastgele şekil ve boyuttaki veri kümeleriyle uğraşırken küme topluluklarının avantajlarını göstermiştir. Bu faydalar, birçok bireysel kümenin rastgele hale getirilmesinden ve tekrarlanmasından kaynaklanır ve bunlar daha sonra optimize edilmiş bir küme oluşturmak için birleştirilir[3] [12].

Fikir birliği kümeleme, bir veya daha fazla kümeleme algoritması kullanılarak topluluk olarak bilinen birçok kümenin oluşturulmasını içerir. Her küme, örnekler içindeki ilişkileri temsil eden, ortak ilişkilendirme matrisi adı verilen bir sonuç üretir. Matrisler, bir fikir birliği fonksiyonu kullanılarak birleştirilir ve topluluk için bir benzerlik matrisi üretilir. Benzerlik matrisi üretimi tekrarlanarak farklı kümeleme sonuçlarının alındığı bir topluluk oluşturulur. [11], [15], [17], [22]. Ortaya çıkan matrisler yine bir fikir birliği fonksiyonu kullanılarak birleştirilir ve sonuçta tüm toplulukları temsil eden bir fikir birliği matrisi elde edilir. Her topluluğun benzerlik matrisleri ve nihai fikir birliği matrisinin her biri, toplayıcı kümeleme olarak bilinen hiyerarşik bir kümeleme algoritması kullanılarak analiz edilir. Bu matrislerin her birinin analizi, bir veri kümesi içindeki grupların iddia edilen sayısı olan önerilen bir k değerini üretir.

Kümeleme topluluklarının sonuçlarının değerlendirilmesiyle tek bir kümeleme sonucu elde edilebilir[8]. Fikir birliği kümelemenin farklı yaklaşımlarla uygulanmasına rağmen k-ortalama algoritması en temel kümeleme algoritmaları içinde en popüler seçeneklerden biri olmaya devam etmektedir. Basitliği ve performansının yeterliliği bakımından k-ortalama yaygın bir şekilde tatbik edilmektedir. Mevcut çalışma, önerilen algoritmanın farklı yorumlamalarını uygulayarak sonuçlarını kapsamlı bir şekilde incelemektedir. Bunlar arasında Bu yorumlamalardan bir tanesi de k-ortalama kümelemesi için başlangıç merkezlerini eniyileyen algoritmadır. Fikir birliği kümelemede geleneksel k-ortalama algoritması ve önerdiğimiz algoritmanın üç farklı yorumlaması karşılaştırılmıştır. Topluluk içerisinde her bir kümeleme sonucu birleştirirken farklı kümeleme sonucu topluluklarını birleştiren farklı fonksiyonların etkileri de incelenmiştir. Son olarak, her kümenin farazi k değerini seçmek için rastgele ve kontrollü yöntemler denenmiştir. Önerilen yöntemler, farklı fikir birliği fonksiyonları ve kümelemede küme sayısı seçimi seçenekleri ile çeşitlendirilmiştir. Buna göre farklı kombinasyonlarla beş farklı veri üzerinde sonuçlar elde edilmiştir.

Bu makale, giriş, temel yöntemler, önerilen yöntemin temeli, deney ve sonuçlar bölümlerinden oluşmuştur.

Temel Yöntemler

k-ortalama

K-ortalama kümeleme algoritması hakkında kapsamlı çalışmalar yapılmıştır [1], [2], [7], [10], [19]. Dolayısıyla ayrıntılı anlatımı yerine her küme için seçilen rastgele başlangıç merkezlerinin algoritmanın erken beklenmedik bir şekilde sonlanmasına odaklanılmıştır. Rassal bir şekilde seçili küme merkezlerinin iyileştirilmesi anlayışı üzerine ilerler, bu da yen algoritmanın zayıf noktalarından biridir, çünkü ilk tohumlamayla ortaya çıkan kümeler kümelerin son halini önemli ölçüde etkileyebilir. Bununla birlikte, rastgele seçilmiş başlangıç merkezlerinin, küme topluluklarını tartışırken, kaliteli bir küme topluluğunun istenen çeşitlilik temsilcisini yaratarak daha iyi sonuçlar alınabildiği gösterilmiştir[25]. Bu çeşitlilik yalnızca her bir topluluk yüksek sayıda küme içeriyorsa belirgindir, L. Bu durumda bariz bir dezavantaj, böyle bir algoritma kullanmanın artan maliyetidir. Benzersiz optimal ağırlık merkezlerini belirlemek ve rastgele seçimi tamamen ortadan kaldırmak için birçok başlangıç ağırlık merkezi seçim algoritması önerilmiştir. Böyle bir seçim algoritması, Literatürde başlangıç noktası seçimine ilişkin farklı yöntemler önerilmiştir[6][16].

Küme Merkezi Seçimlerinin Eniyilemesi

Önerilen yöntemimize temel teşkil eden çalışmada [6] herhangi bir veri kümesi için benzersiz bir başlangıç merkezleri kümesi üreten, optimize edilmiş bir başlangıç ağırlık merkezi seçilmesi sağlanmaktadır. Bu algoritma, herhangi bir binadaki sütunların dağılımından ilham almıştır. Sütunlar merkezi dayanak noktalarını temsil eder; çatının ağırlığına basıncına dayanacak şekilde birbirlerinden mümkün olduğunca uzağa yayılırlar [6], [5]. Algoritmanın zaman karmaşıklığı $O(n+(h_1xn_1)+\ldots+(h_kxn_k))'$dir, burada $n_k<\ldots<n_1<n$, n_i, i'ninci merkezi nokta hariç tutulduktan sonra veri öğelerinin geri kalanıdır. tasarlanan ilk ağırlık merkezinin komşuları ve h_i, i'ninci belirlenen merkezi nokta belirlenmeden önceki aykırı değerlerin sayısıdır. Aykırı değerlerin sayısı n'ye yakın olduğunda, algoritma yaklaşık kuadratik zaman karmaşıklığına ulaşır.

Temel Prensip: Başlangıç ağırlık merkezlerini seçmek için tüm noktaların ortalaması, veri dağılımının ağırlık merkezi olarak hesaplanır. Bu merkeze en uzak mesafeye sahip veri noktası birinci merkeze aday olarak seçilir. n_{min} noktalarının önemli bir kısmı, ağırlık merkezi adayının önceden belirlenmiş bir nb_{dis} mesafesi içerisine düşerse, aday gerçek bir ağırlık merkezine terfi ettirilir. Aksi halde aykırı değer olarak kabul edilir ve ağırlık merkezine en yakın bir sonraki nokta aday olarak seçilir.

Sonraki ağırlık merkezleri de benzer bir süreci takip eder, ancak mesafe, ağırlık merkezine ek olarak önceki tüm merkez noktalarından hesaplanır. Bu, tüm noktaların birbirinden mümkün olduğunca uzağa dağıtılmasını sağlar.

Eşik Değerine Bağlı Aykırı Noktaların Bulunması: Model alınan temel çalışmada, d_{max} değer, herhangi bir numuneden ağırlık merkezine olan maksimum mesafe olsun Buna göre, verilen referans(alfa ve beta parametre değerlerine bağlı kalınarak n_{min} metriği alfa x n ÷ k) olarak hesaplanır. Benzer şekilde, nb_{dis} metriği de beta x d_{max} ile bulunur; burada Eğer n_{min} komşu noktaları aday ağırlık merkezinin nb_{di}s mesafesi dahilindeyse, aday bir ağırlık merkezi haline gelir. Aksi halde aykırı bir duruma dönüşür. Bu aykırı değer tespit mekanizması, rastgele ağırlık merkezi başlatmanın doğasında olan bir sorun olan yerel minimumların ağırlık merkezi olarak seçilmesini önlemeye yardımcı olur.

Önerilen Yöntemin Temeli

Önerilen yöntemin kullandığı [6] numaralı çalışmada rastgele küme merkeziyle başlatılan çalışmaya göre üstünlük gösterirken aykırı nokta tespitine de olanak sağlar.

Aykırı nokta tespiti rastgele alfa ve beta değerleri atanarak farklı küme merkezleriyle kümeleme algoritmasının başlatılmasına olanak sağlayabilir. Bununla birlikte, bu rassallıkla birlikte fikir birliği kümelemesinin etkilerinden faydalanılarak aynı zamanda ağırlık merkezi seçim süreci sırasında bir dereceye kadar aykırı değer tespiti ile birlikte kümeleme sonucu da elde edilebilir.

Aykırı değer tespiti olmadan da uygun başlangıç küme merkezi bulunabilmektedir. [4]. Aykırı değer tespiti gerekmeyen durumda alfa =0 ve beta=1 değer ataması yapılır.

Fikir birliği Kümeleme

Fikir birliği kümelemede, ortak ilişkilendirme matrislerinin bir fikir birliği oluşturmak üzere birleştirildiği iki aşama vardır. Küme bağlantısı olarak adlandırdığımız ilk olay, bir topluluk içindeki her kümenin bir araya gelmesiyle gerçekleşir. İkinci olay, topluluk bağlantısı sırasında, her bir topluluğun nihai fikir birliği matrisini oluşturmak üzere birleştirildiği zamandır. Ortaya çıkan matrisleri analiz etmek ve bir k değerini belirlemek için hiyerarşik kümelemeyi kullanırız. Özellikle, toplayıcı kümeleme, fikir birliği kümelemesinde fikir birliği matrislerini analiz etmek için yaygın olarak kullanılan bir yöntemdir. Yığınsal hiyerarşik kümelemede her veri noktası tekil küme olarak ele alınır ve mevcut kümeler birbirine bağlanarak küme sayısı azaltılır. Daha sonra kümeleri birbirine bağlarken kümeler arasındaki mesafedeki en büyük sıçrama hesaplanarak optimum k değeri çıkarılır. En yaygın olanları seçtiğimiz çeşitli bağlantı alternatifleri vardır(tekli, tam, ortalama, ward metodu gibi) [18].

Kümelemede kümeler arası ilişkilerde, tek bağlantılıda, en yakın noktaları arasındaki en kısa mesafeye sahip iki küme birleştirilir. Hiyerarşik kümelemenin karmaşıklığı genellikle O(n3) olmakla birlikte, O(n2)'de de tek bağlantılı kümeleme yapılabilir[13] [18]. Tek bağlantı, kümeleme uygulama aşamasında oldukça basittir ve bu nedenle de yaygın olarak kullanılır. Ancak tek bağlantı, bir dizi kümenin birbirine beklenmedik bir şekilde uç uca eklenmesine yol açabilir. Kümelerin şekli ve boyutunun dikkate alınmamasından dolayı art arda birbirlerine eklenerek yeni kümeler oluşturabilirler. Beklenen doğal kümelerin oluşturulması mümkün olmayabilir; küme kalitesi ise genellikle küme içi varyans veya standart sapma geçerlilik indeksi gibi geçerlilik ölçümleri aracılığıyla kümenin sıkı ve yoğunluğuyla belirlenir.

Tam bağlantı, tek bağlantının tersine, iki kümenin en yakın noktaları arasındaki mesafe yerine, n uzak noktalarını kullanır. Tek bağlantıya benzer şekilde, tam bağlantı, CLINK algoritmasını kullanılarak O(n2) zaman karmaşıklığına indirgenebilir[18] [20]. Tek bağlantı küme zincirleri oluşturma eğilimindeyken, tam bağlantı küresel kümelere yol sebep olmaktadır. Yeni oluşturulan küme, her birleştirilmiş kümenin en uzak noktaları arasındaki mesafeye eşit bir çapa sahiptir ve diğer tüm veri noktaları, çapa karşılık gelen bir kürenin içine alınır. Tam bağlantının bir dezavantajı da potansiyel aykırı noktaların iki küme arasındaki çapın büyümesine yol açmasıdır. Aykırı nesnelere karsı hassasiyet artmaktadır.

Ortalama bağlantı, iki kümedeki her nokta arasındaki mesafelerin ortalamasını alarak tekli bağlantı ile tam bağlantı arasında bir konuma sahip olur. Tüm noktaların arasındaki ortalama mesafeye göre en yakın iki kümenin birleşiminin yeni bir küme oluşturması söz konusudur. Aykırı değerlere karşı tam bağlantıya göre daha az duyarlıdır ancak O(n2log n) maliyeti göreceli yüksektir.

Ward metodu hangi kümelerin birleştirileceğine karar vermek için sapma karesi kriterini kullanmaktadır. İki kümenin, her küme için bağlı oldukları merkezlere göre sapmaları hesaplanır. İki kümenin aynı küme içinde yeni bir küme oluşturulduğu varsayılarak toplam sapma hesaplanır. Bu toplam sapma daha sonra iki ayrı sapmanın toplamı ile karşılaştırılır. Sapmalardaki fark, veri setindeki her küme çifti için hesaplanır ve sapmada en küçük değişime sahip olan çift, birleştirme için seçilir. Ward metodu, noktaların küme merkezlerinden kare sapmasını en aza indirgemeye çalışan k-ortalama algoritmasına benzer. Zaman karmaşıklığı O(nlog n)'dir [18].

Genel ifadeyle farklı kümelerin sonucunda veri setindeki nesnelerin ikili bağlarının elde edilmesine dayalı adımlar aşağıdaki gibi özetlenebilir[8]:

D data seti N adet öğeye sahiptir. Buna göre,

1. Doğal küme sayısından daha büyük küme adeti için R adet farklı kümeleme sonucu oluşturulur.
2. R adet farklı kümeleme sonucu elde edilir.
3. Her bir sonuç için $S(t) = \{s_{ij}(t)\}$ i ve j=1..N ve t=1..R
4. Bir nesnenin en fazla bir adet kümeye ait olması durumunda, i ve j nesnelerinin aynı kümede olması durumunda $s_{ij}(t)$=1; aksi takdirde, 0 olarak değeri atanır. Buna göre, S(t) matrisi elde edilir.
5. t=1..R farklı kümeleme sonucuna göre elde edilen matrislerin toplamı nihai S matrisinde bir araya

getirilerek nesnelerin bir arada olma sayıları hesaplanır.

Küme Sayısı Değeri Seçimi

Her küme için farazi veya hedef k değeri seçilmelidir. Daha sonra her küme, bir ortak ilişkilendirme matrisinde temsil edilen aynı küme veri noktaları arasındaki ilişkiyle k sayıda kümeyi belirler. Varsayılan k küme sayısına göre veri noktalarının nasıl kümelendiğine bakarak, hangi veri noktalarının birbirine ait olduğu konusunda bir fikir birliği bulmaya çalışırız, bu da optimal küme sayısını sağlar.

Rastgele Küme Sayısı Değeri

Her küme için önerilen k değeri genellikle fikir birliği kümelemesinde rastgele seçilir. Daha önce de belirtildiği gibi, bunun nedeni, ortaya çıkan toplulukların artan çeşitliliğidir. Bu fayda iyi bir şekilde belgelenmiş olsa da, küme topluluklarının halihazırda yüksek hesaplama maliyeti nedeniyle gerçekte daha az pratiktir. İstatistiksel olarak konuşursak, daha fazla örnek alındığında çeşitlendirme etkileri artar, bu da topluluk başına daha fazla küme anlamına gelir.

Sıralı Küme Sayısı Değeri

Sıralı k seçimiyle, öncelikle veri kümesindeki farazi maksimum k küme sayısına bakılır. Beş veri setinin gerçek k değeri iki ile yedi arasında değişmektedir. Her veri setinde kümelerin var olduğu varsayılırsa k değeri ikiden başlatılır; daha sonra k sayısı sıralı arttırılarak yirmi adet küme sayısı değeri için farklı kümeleme sonuçları elde edilir. farklı k toplulukları elde edilir. Bu yöntemle topluluğun çeşitliliğini garanti eder ve benzersiz başlangıç ağırlık merkezi seçimine sahip algoritmalar için daha uygundur. Rastgele başlangıç merkezlerine sahip k-ortalama için aynı k değerine sahip iki küme çok farklı sonuçlara yol açabilir.

Deneyler

Test çalışmalarımızda Şarap, Cam, İyonosfer, Zambak ve Tiroit veri kümeleri kullanılmıştır. Çalışmalarda kullanılan veri setleri UCI ML veri seti havuzunda da yer almaktadır[2].

Dört farklı algoritma, fikir birliği fonksiyonları ve iki farklı küme değeri seçme yöntemi beş farklı veri seti üzerinde test edilmiştir. Elde edilen sonuçlara gör çalışmamıza temel teşkil eden yöntemin[6] tam bağlantıyla en kötü performansı göstermiştir. Sapma miktarı ortalamanın çok üstüne çıkmıştır. Tam bağlantının aykırı değerlere karşı hassasiyetini koruduğu düşünülen aykırı değer tespiti yüzünden bu durum aslında beklenmemekteydi.

Önermiş olduğumuz yöntemlerin hem aykırı değer olması hem de olmaması durumlarında [6] no'lu standart çalışmadan daha iyi performans göstermiştir. Tam bağlantı dışındaki üç varyasyon arasındaki tek fark aykırı değerlerin nasıl tespit edildiği olduğundan, [6] no'lu yöntemdeki aykırı değer tespitli halinin fikir birliği kümeleme için uygun olmadığı ortaya konmuştur. Bunun nedeni, veri kümeleri dikkate alınmadan alfa ve beta değerlerinin keyfi olarak seçilmesine ve aykırı değer tespitinin, her küme arasında sürekli değişen k'ye bağımlı olmasına bağlanabilir.

Elde edilen sonuçlara göre, Ward metodu, hem rastgele hem de sıralı k seçimiyle [6] no'lu çalışmanın aykırı değer tespiti kullanmayan haliyle birleştirildiğinde en güvenilir sonuçları üretmiştir. Rastgele alfa ve beta değerlerinin atanmasıyla çalışan temel yöntem tüm varyasyonlarda iyi performans göstererek k-ortalama algoritmasına yerine alternatif olabileceğini göstermiştir.

Karşılaştırıldığında, rastgele k seçimi sıralı k seçimine göre daha iyi sonuçlar vermiştir. Tüm veri kümelerinin ortalama gerçek k değeriyle karşılaştırıldığında yüksek değerdeki en fazla küme sayısı bunda rol oynamıştır.

Sonuçlar

K-ortalama kümeleme algoritması önermiş olduğumuz algoritmanın üç farklı yorumlamasıyla karşılaştırılmıştır. Algoritmalar dört farklı fikir birliği fonksiyonu ve değişik iki k seçim yöntemiyle birleştirilerek sonuçta 32 adet fikir birliği kümelemesi elde edilmiştir. Değişiklerle farklı varyasyonlar içeren bu yaklaşımlar beş adet veri setine uygulanmıştır. Elde edilen sonuçlara göre, aykırı nesne tespiti içermeden fikir birliği kümeleme tabanlı önerdiğimiz yöntemlerin en iyi performansı gösterdiğini ve rastgele merkezi nokta başlatarak çalışmaya başlayan k-ortalamanın yerine uygun bir alternatif olduğu gösterilmiştir. Önerilen algoritma, O(n2) zaman karmaşıklığına dönüşürken, aykırı değer tespitinin kaldırılması zaman karmaşıklığını O((k+1) n) seviyesine kadar iyileştirir. k-ortalama uygulamalarında ortaya çıkan ağırlık merkezlerinin eylemsizliği karşılaştırılarak birden fazla rastgele ağırlık merkezi başlatma işlemi yapılır. Sonuç olarak, zaman karmaşıklığındaki fark minimum düzeye inmiştir. Çalışılan fikir birliği fonksiyonları arasında Ward metodu en iyi sonuçları verirken aynı zamanda O(n log n) gibi en düşük zaman karmaşıklığına sahiptir.

Rastgele ve sıralı k seçimi karşılaştırıldığında rastgele yöntem daha iyi performans göstermiştir. Bununla birlikte, iki yöntem arasındaki standart sapma farkı yalnızca yaklaşık yüzde üçtür, bu da sıralı k seçimini, zaman karmaşıklığının önemli olduğu durumlarda geçerli bir yöntem haline getirir. Rastgele k seçimi, faydaları veya rassallığı görmek için topluluk başına yüksek miktarda küme gerektirirken sıralı k seçimi için küme sayısı

Tablo 1 Rastgele küme değeri seçerek doğru etiketlemeye göre sapma miktarı (Yüzde oranı cinsinden)

Kümeleme Yöntemi	Şarap	Cam	İyonosfer	Zambak	Tiroit	Ortalama
k-Ortalama + Tek Bağlantı	33.3	71.4	0	33.3	33.3	34.26
Temel Alınan Yöntem + Tek Bağlantı	33.3	71.4	0	33.3	33.3	34.26
Temel Alınan Yöntem (rastgele) + Tek Bağlantı	33.3	71.4	0	0	33.3	27.6
Temel Alınan Yöntem (aykırı noktasız) + Tek Bağlantı	33.3	71.4	0	33.3	0	27.6
k-ortalama Ortalama Bağlantı	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem + Ortalama Bağlantı	0	57.1	100	33.3	33.3	44.74
Temel Alınan Yöntem(rastgele) + Ortalama Bağlantı	0	57.1	0	33.3	33.3	24.74
Temel Alınan Yöntem (aykırı noktasız) + Ortalama Bağlantı	33.3	71.4	0	33.3	0	27.6
k-Ortalama + Tam Bağlantı	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem + Tam Bağlantı	66.7	71.4	100	33.3	33.3	60.94
Temel Alınan Yöntem (rastgele) + Tam Bağlantı	0	71.4	0	33.3	0	20.94
Temel Alınan Yöntem (aykırı noktasız) + Tam Bağlantı	0	71.4	0	33.3	0	20.94
k-Ortalama + Ward Metodu	0	71.4	0	33.3	33.3	27.6
Pi Temel Alınan Yöntem + Ward Metodu	0	71.4	100	33.3	33.3	47.6
Temel Alınan Yöntem (rastgele) + Ward Metodu	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem (aykırı noktasız) + Ward	0	57.1	0	33.3	0	18.08

Tablo 2 Sıralı küme değeri seçerek doğru etiketlemeye göre sapma miktarı (Yüzde oranı cinsinden)

Kümeleme Yöntemi	Şarap	Cam	İyonosfer	Zambak	Tiroit	Ortalama
k-Ortalama + Tek Bağlantı	33.3	71.4	0	33.3	33.3	34.26
Temel Alınan Yöntem + Tek Bağlantı	33.3	57.1	0	0	66.7	31.42
Temel Alınan Yöntem (rastgele) + Tek Bağlantı	33.3	71.4	0	33.3	33.3	34.26
Temel Alınan Yöntem (aykırı noktasız) + Tek Bağlantı	33.3	71.4	0	0	33.3	27.6
k-ortalama Ortalama Bağlantı	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem + Ortalama Bağlantı	200	57.1	0	33.3	33.3	64.74
Temel Alınan Yöntem(rastgele) + Ortalama Bağlantı	0	57.1	0	33.3	0	18.08
Temel Alınan Yöntem (aykırı noktasız) + Ortalama Bağlantı	33.3	71.4	0	33.3	0	27.6
k-Ortalama + Tam Bağlantı	0	71.4	0	33.3	66.7	34.28
Temel Alınan Yöntem + Tam Bağlantı	200	28.6	100	33.3	33.3	79.04
Temel Alınan Yöntem (rastgele) + Tam Bağlantı	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem (aykırı noktasız) + Tam Bağlantı	0	71.4	0	33.3	33.3	27.6
k-Ortalama + Ward	0	71.4	0	33.3	33.3	27.6
Pi Temel Alınan Yöntem + Ward	0	71.4	100	33.3	33.3	47.6
Temel Alınan Yöntem (rastgele) + Ward	0	71.4	0	33.3	33.3	27.6
Temel Alınan Yöntem (aykırı noktasız) + Wardu	0	57.1	0	33.3	0	18.08

kmax(belirlenen maksimum küme sayısı değeri) ile sınırlanır.

Aykırı değer tespiti yapılmadan önerilen yöntemi kullanan temel kümeleme, her topluluk içinde bir fikir birliği oluşturmak için Ward metodu kullanılır. Elde edilen sonuçlardan bu fikir birliği kümeleme algoritması öne çıkmaktadır. En çok tekrarlanan k değeri son k değeri olur. Çalışmalarda her bir topluluk için sapmada en fazla sıçramanın olduğu küme sayısı hesaplanmıştır. Çalışmada toplam 100 adet topluluk oluşturularak en fazla tekrarlanan k küme sayısı dikkate alınarak veri setinde gerçek etiketlemeye (sınıf) göre öğelerin standart sapma değişimi hesaplanmıştır. Düşük sapma değeri daha doğru kümeleme yapmıştır şeklinde değerlendirilmelidir. Önerilen yöntemlere temel teşkil eden [6] no'lu çalışmadır. Bu nedenle aşağıdaki tablolarda [6] no'lu çalışma Temel Alınan Yöntem olarak adlandırılmıştır. Çalışmalar fikir birliği fonksiyonları ve farklı küme sayısı değeri seçme gibi alternatiflerle çeşitlendirilmiştir

Kümeleme algoritmaları hali hazırda bir çok uygulama alanında tatbik edilebilir. Geleneksel bölümleyici algoritmaları açgözlü bir yaklaşım sergilerler. O nedenle başlangıçta seçilen ilk çözüm sonraki aşamalardaki iyileştirmelerin kalitesi açısından çök önemlidir. Aksi takdirde, erken yakınsayarak yanlış çözümler elde edilmesine yol açabilirler. O nedenle, hemen her alanda örneğin, bir firmadaki müşterilerin profil bilgilerinin elde edilmesinde veya bir tavsiye sisteminde yapılan önerilerin iyileştirilmesinde iyileştirmeler daha iyi bir servis sunulması açısından çok büyük önem taşır. Gelecekte, çalışmamızdaki veri setleri dışında yapılan çalışmaların ayrıca kümeleme algoritmalarının farklı uygulama alanlarına tatbiki üzerine çalışmalar yapılacaktır. Ayrıca, büyük veri üzerine uyarlanacaktır.

Kaynaklar

[1] Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation." Electronics 9.8 (2020): 1295.

[2] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007).

[3] Bai, Liang, Jiye Liang, and Fuyuan Cao. "A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters." Information Fusion 61 (2020): 36-47.

[4] Barakbah, Ali Ridho, and K. Arai. "A new algorithm for optimization of K-means clustering with determining maximum distance between centroids." IES 2006, Politeknik Elektronika Negeri Surabaya, ITS (2006).

[5] Barakbah, Ali Ridho, and Yasushi Kiyoki. "A fast algorithm for K-means optimization using Pillar algorithm." The 2nd International Workshop with Mentors on Database, Web and Information Management for Young Researchers. 2010.

[6] Barakbah, Ali Ridho, and Yasushi Kiyoki. "A pillar algorithm for k-means optimization by distance maximization for initial centroid designation." 2009 IEEE Symposium on Computational Intelligence and Data Mining. IEEE, 2009.

[7] Borlea, Ioan-Daniel, Radu-Emil Precup, and Alexandra-Bianca Borlea. "Improvement of K-means cluster quality by post processing resulted clusters." Procedia Computer Science 199 (2022): 63-70.

[8] Cano, José Ramón, et al. "A greedy randomized adaptive search procedure applied to the clustering problem as an initialization process using K-Means as a local search procedure." Journal of Intelligent & Fuzzy Systems 12.3-4 (2002): 235-242.

[9] Ezugwu, Absalom E., et al. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects." Engineering Applications of Artificial Intelligence 110 (2022): 104743.

[10] Ghazal, Taher M. "Performances of K-means clustering algorithm with different distance metrics." Intelligent Automation & Soft Computing 30.2 (2021): 735-742.

[11] Goder, Andrey, and Vladimir Filkov. "Consensus clustering algorithms: Comparison and refinement." 2008 Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX). Society for Industrial and Applied Mathematics, 2008.

[12] Golalipour, Keyvan, et al. "From clustering to clustering ensemble selection: A review." Engineering Applications of Artificial Intelligence 104 (2021): 104388.

[13] Goyal, Poonam, et al. "Spatial locality aware, fast, and scalable slink algorithm for commodity clusters." 2016 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2016.

[14] Ikotun, Abiodun M., et al. "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data." Information Sciences (2022).

[15] Jia, Yuheng, et al. "Ensemble Clustering via Co-Association Matrix Self-Enhancement." IEEE Transactions on Neural Networks and Learning Systems (2023).

[16] Jumadi Dehotman Sitompul, Bernad, Opim Salim Sitompul, and Poltak Sihombing. "Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm." Journal of Physics: Conference Series. Vol. 1235. No. 1. IOP Publishing, 2019.

[17] Lancichinetti, Andrea, and Santo Fortunato. "Consensus clustering in complex networks." Scientific reports 2.1 (2012): 336.

[18] Sharma, Shweta, and Neha Batra. "Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering." 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019.

[19] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." IEEE access 8 (2020): 80716-80727.

[20] Sreedhar Kumar, S., et al. "A brief survey of unsupervised agglomerative hierarchical clustering schemes." Int J Eng Technol (UAE) 8.1 (2019): 29-37.

[21] Turgut, Emre, Murat Taşyürek, and Nuh AZGINOĞLU. "Kentsel Dönüşüm Sürecinde Binaların Mekânsal Veri Madenciliği Yöntemleri ile Tespiti." Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi 13.2 (2022): 161-167.

[22] Zhang, Mimi. "Weighted clustering ensemble: A review." Pattern Recognition 124 (2022): 108428.