Süleyman Demirel Üniversitesi,
Fen Bilimleri Enstitüsü Dergisi,
15-3(2011),192-202

Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması

Güncel SARIMAN*
Süleyman Demirel Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü/ ISPARTA
Alınış Tarihi:17.05.2011, Kabul Tarihi:10.10.2011

Özet: Günümüzde bilişim alanındaki gelişmeler sayesinde kullanıcılar daha gelişmiş bilgisayar teknolojilerine hâkim olmaktadır ve bu gelişmeler beraberinde sayısal verilerin birikmesine neden olmaktadır. Bilişim teknolojileri biriken verileri saklayabilir. Tüm bu saklanan verilerin anlamlandırılması için veri madenciliği kullanılmaktadır. Veri madenciliği, verilerden önceden bilinmeyen anlamlı bilgileri tanımlama ya da tahmin etme tekniklerini içermektedir. Veri madenciliğinde verileri ortak özellikleri ile gruplamak için kümeleme algoritmaları yardımıyla ilginç desenler bulunabilir. Bu çalışmada UCI Machine Learning Repository veritabanından "Flags" veri seti alınarak k-means ve k-medoids bölümlemeli kümeleme algoritmalarıyla ülkelerin özelliklerine göre kümelere ayrılması hedeflenmiştir. Uygulama Asp.Net ile web ara yüzünde geliştirilerek internet ortamında kullanıcılara sunulmuştur. Çalışmanın sonunda veri seti K-Means ve K-Medoids algoritmalarıyla çalıştırılmış ve elde edilen analiz sonuçları karşılaştırılmalı olarak incelenmiş ve kullanım yerlerine yönelik öneriler sunulmuştur

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Analizi, Asp.Net, K-Means, K-Medoids

A Study of Clustering Techniques in Data Mining: Comparison of The K-Means and K-Medoids Clustering Algorithms

Abstract: Nowadays, thanks to advances in the field of IT, users dominate the more advanced computer Technologies and all these developments leads to the accumulation of numerical datas. Information Technologies can store accumulated datas. Data mining is used for all datas that are stored for giving meaning. Data mining contains estimation techniques or identify previously unknown meaningful techniques. Interesting patterns can be found to cluster with common features via clustering algorithms in data mining. In this study, "Flags" data set that taken from UCI Machine Learning Repository database, based k-means and k-medoids partitioned clustering algorithms aimed at the separation of clusters according to their country features. Application developed and presented to the users with Asp.Net in web user guide. At the end of the study, the k-means and k-medoids algorithms were checked by comparing the performances and presented suggestions for their place of use.

KeyWords: Data Mining, Cluster Analysis, Asp.Net, K-Means, K-Medoids

Giriş

Günümüzde başta is dünyası olmak üzere, birçok farklı alanda kullanılan veri madenciliği, MIT (Massachusets Institue of Technology) tarafından 2001 yılında yayınlanan bildirgeye göre dünyayı değiştirecek 10 teknoloji arasında gösterilmiştir (Mit, 2005). Gelecekte daha çok önem kazanacak olan veri madenciliği üzerinde yapılan çalışmalara her geçen gün yenileri eklenmektedir, tıptan uzay bilimlerine kadar birçok farklı sektörde kullanılan veri madenciliğinin kullanım alanlarına her gün yenileri eklendiği düşünülürse, konunun önemi daha iyi anlaşılır (Dinçer, 2006).

Veri madenciliğinin amaçları genellikle sınıflandırma, kümeleme, tahmin öngörü ve benzer gruplama olarak sıralanmaktadır. Amaçlardan biri olan kümeleme, istatistiksel veri analizi, örüntü tanıma vb. birçok alanda sık kullanılmaktadır. Veri tabanlarındaki verilerin gruplar veya kümeler altında toplanarak, benzer özelliklere sahip nesnelerin bir araya gelmesini sağlayan kümeleme algoritmaları veri madenciliği alanında büyük bir öneme sahiptir.

* guncelsariman@sdu.edu.tr

Ancak kümeleme algoritmalarının seçimi hem mevcut veri türüne hem de amaç ve uygulamaya bağlıdır. Şayet kümeleme analizi tanımlayıcı ya da açıklayıcı bir araç olarak kullanılacaksa bu durumda aynı veri seti üzerinde birkaç algoritmanın uygulanması mümkün olmaktadır (Akın, 2008).

Bu çalışmada, öncelikle veri madenciliğinin ne olduğu ve yöntemleri kısaca özetlenecek daha sonra ise çalışmanın yapılış aşamaları anlatılarak ayrıntılı sonuçlar verilecektir. Son bölümde elde edilmiş anlamlı sonuçlardan yola çıkılarak kümeleme algoritmaları karşılaştırılmış ve optimum sonuç veren algoritma belirtilmiştir.

Çalışmada, veri madenciliğinde kullanılan kümeleme algoritmalarından yola çıkarak bölümlemeli algoritmalar arasındaki farklar ortaya koyulmaya çalışılmıştır. Bölümlemeli kümeleme algoritmaları kullanılarak "Flags" veri setinin 194 ülke bayrağına ait 30 özelliği içerisinden en uygun kümelemeyi yapacak 3 özelliği(ülke alanı, ülke nüfusu, ülke dini) alınmış ve benzer ülkelerin aynı kümelerde olması amaçlanmıştır. Uygulama web ara yüzünde geliştirilerek kullanıcıların internet üzerinden kümeleme işlemlerini gerçekleştirmeleri sağlanmıştır.

Veri Madenciliği ve Kümeleme Analizi

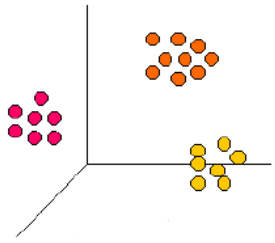
Veri tabanı sistemlerinin artan kullanımı ve veri depolama ünitelerinin hacimlerindeki olağanüstü artış geleneksel sorgulama ve raporlama araçlarının dev veri yığınları karşısında etkisiz kalmasına yol açmıştır. Bunun sonucunda veri tabanlarında bilgi keşfi (VTBK) (KDD-Knowledge Discovery in Databases) adı altında yeni arayışlar ortaya çıkmıştır.(Dinçer, 2006).

Basit bir tanım yapmak gerekirse veri madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi madenleme işidir. Diğer bir deyişle, veri madenciliği tek başına bir şey ifade etmeyen veriler içindeki gizli örüntüleri ve ilişkileri ortaya çıkarmak için istatistik, yapay zekâ ve makine öğrenmesi gibi yöntemlerin ileri veri çözümleme araçlarıyla kullanılmasını kapsayan süreçler topluluğudur. Veri madenciliği uygulamaları başta pazarlama, bankacılık, tıp, mühendislik, endüstri, borsa analizleri ve ulusal güvenlik alanlarında kullanılmaktadır. Veri madenciliği çalışmaları yapmak üzere birçok ticari yazılım üretilmiştir. Oracle DM, Microsoft SQL Server Analysis Services, SPSS Clementine, SAS Enterprise Miner bu ürünlerden sadece birkaçıdır (Bozkır vd., 2009).

Veri madenciliği büyük veri yığınlarında gizli olan örüntüleri ve ilişkileri ortaya çıkarmak için istatistik ve yapay zekâ kökenli çok sayıda ileri veri çözümleme yönteminin tercihen görsel bir programlama ara yüzü üzerinden kullanıldığı bir süreçtir (Dolgun vd., 2009). Veri Madenciliği çok büyük veri yığınlarından kritik bilgileri elde etmeyi sağlar. Böylelikle normal şartlar altında uzun zaman süren araştırmalarla doğruluğu kesin olmayacak şekilde elde edilen bilgi veri madenciliği ile kısa sürede ve kesin olarak elde edilir. Elde edilen bu bilgi objektif değerlendirmeler yapılmasında ya da stratejik kararlar almada kullanılır. Bu bilgiler kurumsal veri kaynaklarının iyi analiz edilmesine ve iş dünyasındaki yaklaşımlara ilişkin tahminlerde bulunulmasına yardımcı olur. Kısaca veri madenciliği sayesinde şirketler stratejik adımlar atarken çok büyük veri yığınları arasından kendilerine yol gösterecek kritik verileri ayıklayarak analiz edebilir (Alpaydın, 2000). Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir (Arslan, 2008).

Veri madenciliğinde kullanılan teknikler eldeki veri türüne ve elde edilen sonuçların kullanım amacına göre modellere ayrılabilir. Bu modeller iki başlık altında toplanabilir. Bunlar tahmin edici (Predictive) ve tanımlayıcı (Descriptive) modellerdir. Tanımlayıcı modeller veri setinin içinden ilişkileri çıkarır. Tanımlayıcı modellerde kullanılan veri madenciliği teknikleri ise, kümeleme, özetleme, birliktelik kuralları, sıralı dizilerdir. Tahmin edici modeller ise, sonuçları önceden bilinen durumlardan bir model geliştirir ve bu model ile sonuçları bilinmeyen veri kümelerinden yeni sonuçlar elde etmektedir. Tahmin edici modellerde kullanılan veri madenciliği teknikleri sınıflandırma, eğri uydurma, zaman serileridir. Kümeleme analizi ise bir veri kümesindeki bilgileri belirli yakınlık kriterlerine göre gruplara ayırma işlemidir. Bu grupların her birine "küme" adı verilir. Kümeleme analizine kısaca "kümeleme" adı verilir. Kümeleme işleminde küme içindeki elemanların benzerliği fazla, kümeler arası benzerlik ise az olmalıdır. Kümeleme veri madenciliği tekniklerinden tanımlayıcı modellere yani gözetimsiz sınıflandırmaya girer. Gözetimsiz sınıflamada amaç, başlangıçta verilen ve henüz sınıflandırılmamış bir küme, veriyi anlamlı alt kümeler oluşturacak şekilde öbeklemektir. Kümeleme işlemi tamamen gelen verinin özelliklerine göre yapılır (Dinçer, 2006).

Kümeleme analizinin kullanılmasında benzer uzaklıklar dikkate alınarak yararlanılabilecek alternatif ölçü ve yöntemler bulunmaktadır. Birimler arası uzaklıklar için Euclidyen, Standardize Euclidyen, Manhattan Mahalanobis, Kareli Euclidyen, Minkowski veya Canberra ölçüleri kullanılabilmektedir. Bu da kümeleme analizinin uygulamada kullanılmasında dikkatli davranmayı zorunlu kılmaktadır. Kümeleme algoritması veritabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dâhil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptir. Kümeleme modellerinde amaç, Şekil 1' de görüldüğü gibi küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir (Arslan, 2008).

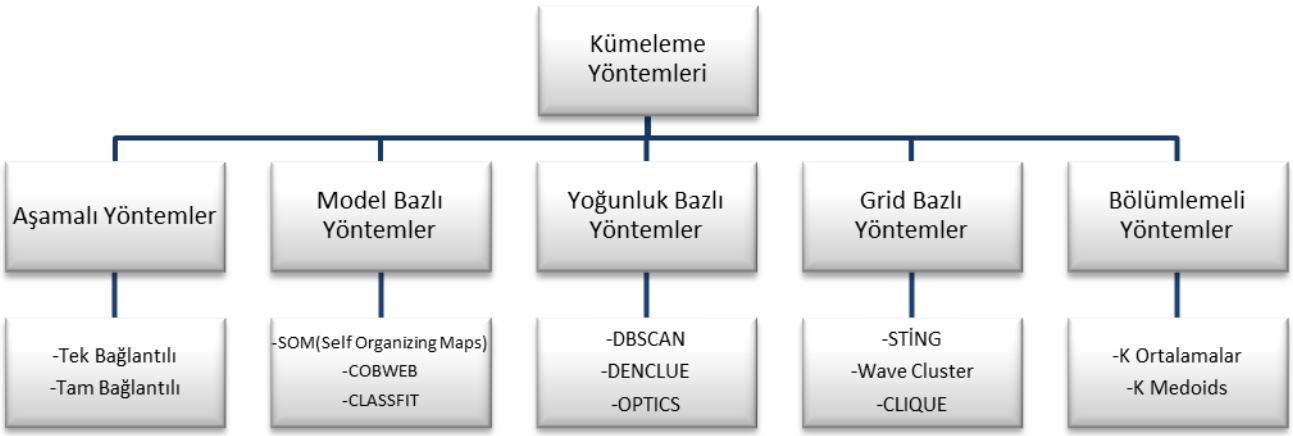


Şekil 1. *Koordinat Düzleminde Kümeleme Örneği*

Tahmin edici modeller kümeleme modelini, homojen veri grupları oluşturması için veri ön işleme aşaması olarak da kullanmaktadırlar.

Kümeleme analizi, bireylerin ya da nesnelerin sınıflandırılmasını ayrıntılı bir şekilde açıklamak amacıyla geliştirilmiştir. Bu amaca yönelik olarak, ele alınan örnekte yer alan varlıklar aralarındaki benzerliklere göre gruplara ayrılır, daha sonra bu gruplara dâhil edilen bireylerin profili ortaya konur. Bir başka ifade ile kümelemenin amacı, öncelikle ele alınan örnekte gerçekte var olduğu bilinen, varlıklar (birey ya da nesne) arasındaki benzerliklere dayanan az sayıdaki karşılıklı özel grupları oluşturmak, daha sonra bu gruplara giren varlıkların profilini ortaya koymaktır. Diğer bir hedef ise benzer elemanların gruplanmasıyla veri setini küçültmektir (Pang-Ning Tan vd., 2006). Satış hareketleri veya çağrı merkezi kayıtları gibi çok fazla parametre içeren çok büyük miktarlardaki verileri analiz etmede en uygun yöntemlerden biri kümelemedir (Işık, 2006).

Kümeleme yöntemleri hiyerarşik ve hiyerarşik olmayan (bölümlemeli) kümeleme yöntemleri olmak üzere iki bölüme ayrılırken bu konuda yapılan araştırmalar bu algoritmaların daha alt bölümlere ayrılabileceğini göstermektedir (Berkhin, 2004). Hiyerarşik kümeleme yönteminde özellikle işleyişin daha kolay anlaşılabilmesi için dendogram (ağaç grafiği)dan yararlanılır. Hiyerarşik kümeleme yönteminde anlatılan işlemlere dayalı olarak kullanılan hiyerarşik metotlardan en çok kullanılanları; Tek bağlantılı, Tam bağlantılı, Ortalama bağlantı, Merkezi ve Ward metotlarıdır. Hiyerarşik olmayan kümeleme yöntemi ise küme sayısı hakkında bir ön bilginin olması ya da araştırmacının anlamlı olacak küme sayısına karar vermiş olması durumunda tercih edilmektedir. Hiyerarşik olmayan kümeleme yönteminde en çok tercih edilen iki yöntem Mac Queen tarafından geliştirilen k-ortalama tekniği ve en çok olabilirlik tekniğidir. Bu çalışmada ise hiyerarşik olmayan kümeleme algoritmaları üzerine bir uygulama geliştirilmiştir. Veri Madenciliği ile ilgili kaynaklarda kümeleme yöntemleri Şekil 2' deki gibi sınıflandırılmaktadır (Han ve Kamber, 2000).



Şekil 2. *Kümeleme Yöntemleri*

Materyal ve Metot

Çalışmada UCI Machine Learning Repository veri tabanından "Flags" veri seti alınarak ülkelerin belirli özelliklerine göre gruplara ayrılması için kümeleme algoritmalarıyla bir çalışma gerçekleştirilmiştir. Çalışmanın ilk aşamasında metin dosyasından (*.txt) alınan veri kümesi ön işlemden geçirilerek MsSql veri tabanına, geliştirilen yazılım ile aktarılmıştır. Ön işlem aşamasında bayrak verileri algoritmalara uygun hale getirilmiştir. K-means ve K-Medoids algoritmaları sayısal verilerle çalıştığı için sözel tanımlar rakamsal verilere dönüştürülmüştür. Veri kümesinde ülkelere ait 30 ayrı özellik bulunmaktadır. Bu özellikler Çizelge 1' de gösterilmiştir.

Çizelge 1. *Bayrak Veri setinin özellikleri*

ÖZELLİK	AD
Bayrak Verileri	Name, LandMass, Zone, Area, Population, Language, Religion, Bars, Stripes, Colours, Red, Green, Blue, Gold, White, Black, Orange, MainHue, Circles, Crosses, Saltires, Quarters, Sunstars, Crescent, Triangle, İcon, Animate, Text, Topleft, BotRight

Veri tabanına aktarılan verilerden "var-yok" sözel verileri içeren özellikler "1-0" tipinde sayısallaştırılmıştır. Çalışmada ülkelere ait alan, nüfus ve din arasındaki ilişkilere bakılarak en iyi kümeleme sonuçlarının verilebileceği düşünülmüştür. Bu yüzden 3 parametre üzerinden analiz işlemleri gerçekleştirilmiştir. Veri tabanında ülke dinlerini kullanarak işlem yapabilmek için sözel veriler sayısallaştırılmıştır. Çizelge 2' de sayısallaştırılan veriler ve numaraları bulunmaktadır.

Çizelge 2. *Ülke Dinlerinin Sözel ve Sayısal Karşılıkları.*

DİN ADI	AĞIRLIK
Catholic	0
Other Christian	1
Muslim	2
Buddhist	3
Hindu	4
Ethnic	5
Marxist	6
Others	7

Bölümlemeli Kümeleme Algoritmaları

Bölünmeli kümeleme algoritmaları k giriş parametresini alarak n tane nesneyi k tane kümeye böler. Bu teknikler, dendogram gibi iç içe bir kümeleme yapısı üzerinde çalışmak yerine tek-seviyeli kümeleri bulan işlemler gerçekleştirir(Jain vd., 1999). Bütün teknikler merkez noktanın kümeyi temsil etmesi esasına dayanır. Bölümlemeli yöntemler hem uygulanabilirliğinin kolay hem de verimli olması nedeniyle iyi sonuçlar üretirler. Çalışmada kullanılan k-means ve k-medoids bölümlemeli kümeleme algoritmaları aşağıda açıklanmıştır.(Işık, 2006)

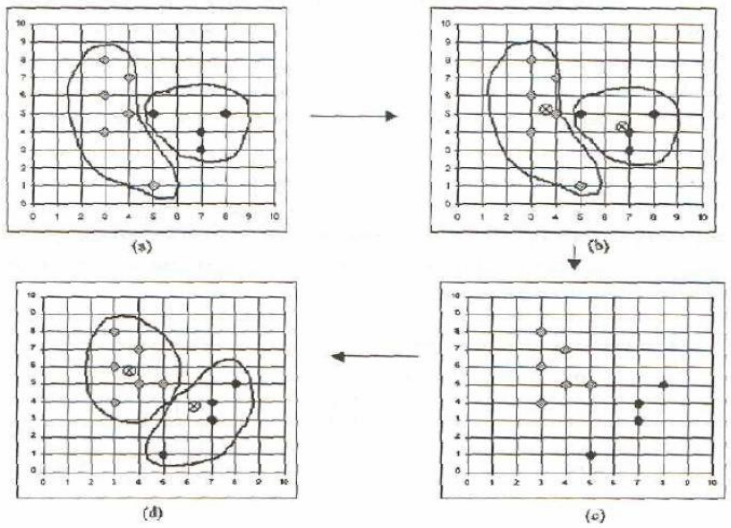
K-Means

En yaygın kullanılan gözetimsiz öğrenme yöntemlerinden biridir. K-means'in atama mekanizması her verinin sadece bir kümeye ait olabilmesine izin verir (Evans, 2005). Bu nedenle, keskin bir kümeleme algoritmasıdır(bölümlemeli kümeleme). K-means algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri kümesini, giriş parametresi olarak verilen k adet kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Çalışma yönteminde, (1) numaralı Öklit uzaklığı formülü temel alınarak kümeleme yapılmaktadır (Dinçer, 2006).

$$p = (p_1, p_2, \ldots, p_n) \text{ ve } q = (q_1, q_2, \ldots, q_n)$$

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (1)$$

K-Means algoritması rastgele seçilen K (küme sayısı) adet merkez noktayla başlar. Veri kümesindeki her nokta kendisine en yakın merkez noktanın kümesine atanır. Küme merkezinin değeri kendine ait noktaların ortalaması alınarak hesaplanır. Bu işlem merkezlerin değerleri değişmeyinceye kadar devam eder (Amasyalı vd., 2008).



Şekil 3. *K-means kümeleme adımları*

K-means algoritmasının işlem basamakları aşağıdaki gibidir.

1.Adım: k adet nesneyi rastgele seç. Seçilen k adet nesne küme merkezlerini temsil eder. $M_1,M_2,..M_k$ Örnek orta nokta aşağıdaki gibi hesaplanır(Gersho, Gray, 1991).

$$M_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik} \qquad (2)$$

2.Adım: Küme içi değişmeleri (3). Karesel Hata Formulu formüldeki gibihesaplanır.$e_1,e_2,...e_k$ (Linde vd., 1980).

$$e_i^2 = \sum_{i=1}^{n_k}(x_{ik} - M_k)^2 \qquad (3)$$

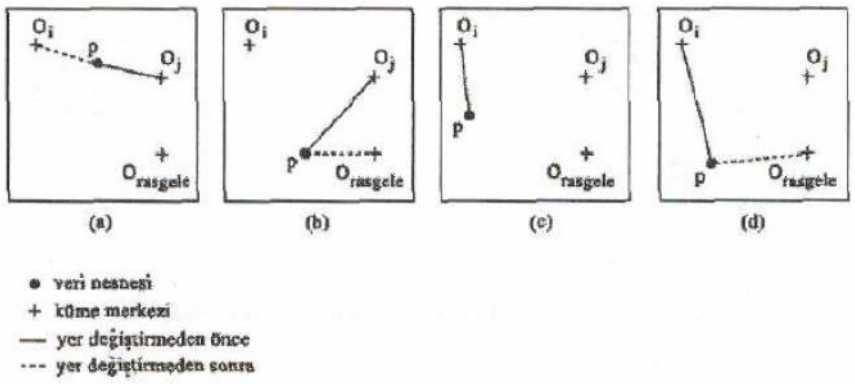
K kümesini içeren bütün kümeler uzayı için kare-hata, küme içindeki değişmelerin toplamıdır. O halde söz konusu kare-hata değeri formül (4) deki gibi hesaplanır:

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \qquad (4)$$

3.Adım: Her bir veriyi kendisine en yakın kümeye ata.
4.Adım: Verilerin hepsi en yakın kümelere atandığında tekrar k tane küme için merkezleri hesapla
5.Adım: Küme Merkezlerinde bir değişiklik olmayıncaya kadar 2. ve 3. Adımları tekrarla.

K-Medoids

K-medoids algoritmasının temeli verinin çeşitli yapısal özelliklerini temsil eden k tane temsilci nesneyi bulma esasına dayanır. En yaygın kullanılan k-medoids algoritması 1987 yılında Kaufman and Rousseeuw tarafından geliştirilmiştir. Temsilci nesne diğer nesnelere olan ortalama uzaklığı minimum yapan kümenin en merkezi nesnesidir. Bu nedenle, bu bölünme metodu her bir nesne ve onun referans noktası arasındaki benzersizliklerin toplamını küçültme mantığı esas alınarak uygulanır. Kümeleme literatüründe temsilci nesnelere çoğunlukla merkez tipler (centrotypes) denilmektedir (Işık, 2006).



Şekil 4. *K-medoids algoritmasının kümeleme adımları*

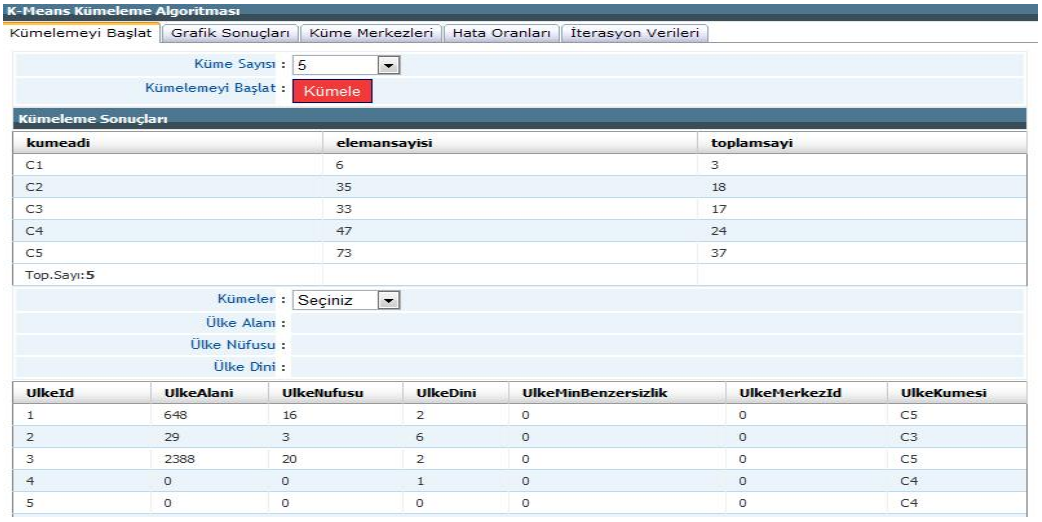
K-medoids algoritmasının birçok farklı türevi bulunmaktadır. PAM (Partitioning Around Medoids) ilk ortaya atılan K-medoids algoritmasıdır. PAM, öncelikle k-means algoritmasında olduğu gibi rastgele seçtiği k adet sayıyı küme merkezi olarak alır. Kümeye her yeni eleman katıldığında kümenin elemanlarını deneyerek kümenin gelişmesine en fazla katkıda bulunabilecek noktayı tespit edince bulduğu noktayı yeni merkez, eski merkezi ise sıradan küme elemanı olacak şekilde yer değiştirme (swap) işlemi yapar (Dinçer, 2006). K-Medoids algoritmasının işlem basamakları aşağıdaki gibidir:

Adım 1: K küme sayısının belirlenmesi.
Adım 2: Başlangıç medoidleri olarak k nesnelerinin seçimi.
Adım 3: En yakın medoid x'e sahip kümeye, kalan nesneleri atamak
Adım 4: Amaç fonksiyonunu hesaplamak. (Hata kareler kriteri: en yakın medoidler için bütün nesnelerin uzaklıklarının toplamı)
Adım 5: Tesadüfî olarak medoid olmayan y noktasının seçimi.
Adım 6: Eğer x ile y'nin yer değiştirmesi amaç fonksiyonunu minimize edecekse bu iki noktanın (x ile y) yerini değiştirmek.
Adım 7: Değişiklik olmayana kadar Adım 3 ile Adım 6 arası işlemler tekrarlanır (Akın, 2008).

Bulgular

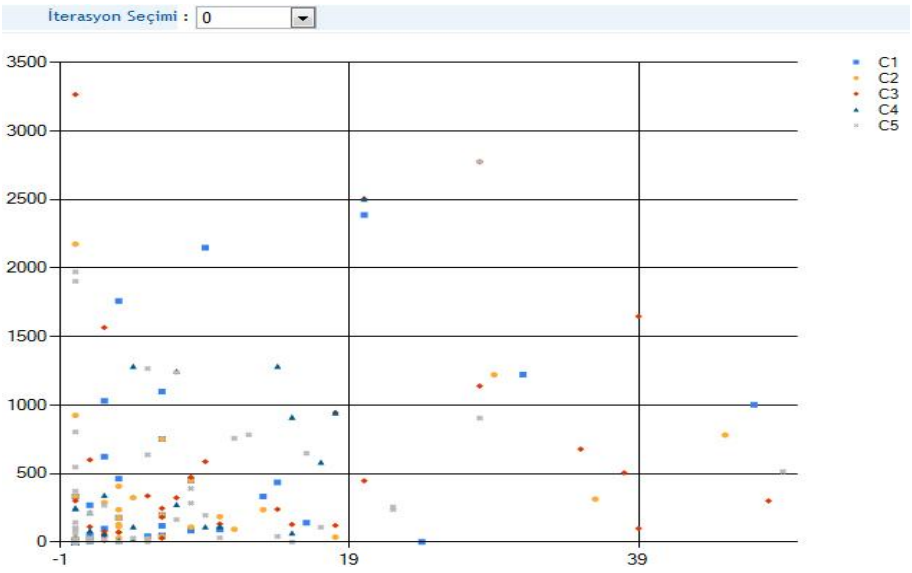
"Flags" veri seti kullanılarak geliştirilen uygulamayı eş zamanlı ve birden çok kullanıcıya sunmak için web ara yüzü geliştirilmiştir. Web ara yüzü ile veriler kümelere k-means ve k-medoids bölümlemeli algoritmalar ile ayrılmıştır. Uygulama Microsoft Visual Studio kullanılarak .Net ortamında geliştirilirken C# programlama dili kullanılmış ve asp.net ile tasarlanmış olup uygulama için k-means ve k-medoids algoritmalarını çalıştırabileceğimiz iki sayfa tasarlanmıştır. Algoritmalar için veri setindeki 30 farklı özellik listelenir, fakat

uygulama için tüm özellikler yerine, doğru kümelere ayrılabilecek 3 önemli özellik kullanılmış ve bunlar ülke dini, ülke nüfusu, ülkenin kapladığı alandır. Geliştirilen uygulamada her iki sayfada da algoritmaları karşılaştırabilmek için ara yüzler benzer şekilde tasarlanmıştır. Bölümlemeli algoritmalar başlangıçtaki küme sayısına göre çalıştırılmaktadır. Çalışmada uygun küme sayısını bulmak için çeşitli denemeler yapılmış ve en iyi küme sayısı olarak 5 seçilmiştir. Şekil 5' de veri tabanından çekilen ülke verilerine göre k-means algoritmasının kümeleme sonuçları ve her bir kümede bulunan üye yüzdeleri görülmektedir.

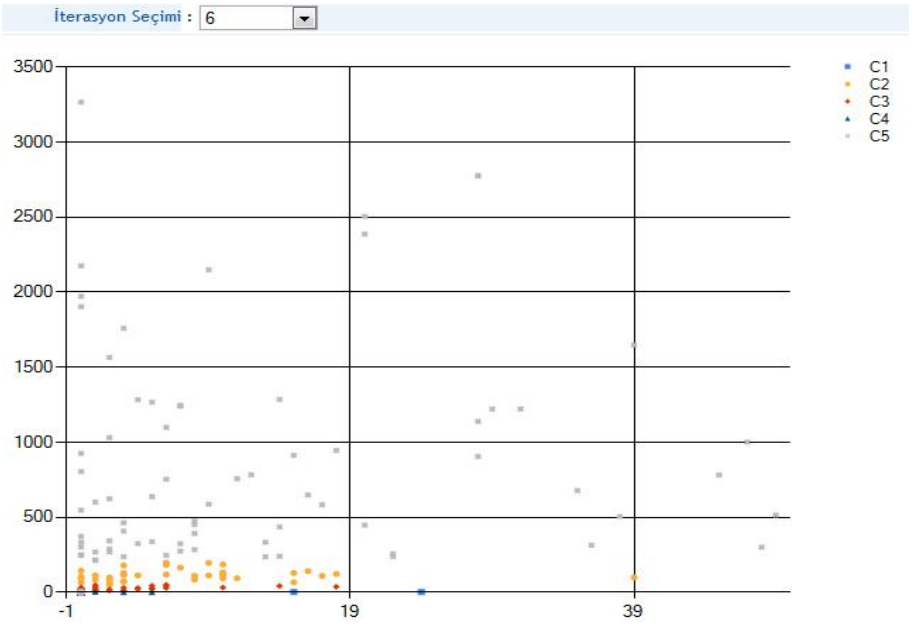


Şekil 5. *K-means algoritmasının k=5 için küme yüzdeleri*

Küme sayısına göre elemanların kümelere dağılımı algoritma sonucunda ortaya çıkmaktadır. Web ara yüzü ile grafik sonuçları sekmesinde her bir adımdaki küme dağılımları gözlenebilmektedir. Böylece algoritmanın performansı adım adım gözlenebilmektedir. Algoritmalar için küme sayısını K=5 vererek performans analizi yapılmıştır. Buna göre k-means algoritmasında 1. ve 6. (son) adım sonundaki grafik gösterimi şekil 6 ve şekil 7' deki gibi olmaktadır.



Şekil 6. *K-means algoritmasının k=5 için 0. adımdaki küme dağılımları*



Şekil 7. *K-means algoritmasının k=5 için 6.adımdaki küme dağılımları*

Uygulamanın son bulması için her adım(iterasyon) bitiminde kümelerde değişiklik olup olmadığına bakılır. Uygulama sonunda küme içi değişme oranına bakılır. Küme içi değişim oranları şekil 8' deki gibidir.

iterasyonsayisi	hataoran
1	905976103,703255
2	634229753,078598
3	471470162,6793
4	258818033,311307
5	212214857,054297
6	212325727,404425
7	212310438,694004
Top.Sayı:7	

Şekil 8. *K-means algoritmasının k=5 için her adımdaki hata oranları*

K-means kümeleme algoritmasını tekrarlayarak en iyi sonucu elde edebiliriz. Kümeleme sonucunda kümeler arasındaki ilişki çizelge 3'deki gibidir.

Çizelge 3. *K-means algoritmasıyla ülkelerin kümeler arasındaki dağılımları*

Kümeler	Alan	Nüfus	Din
C1	7690 - 22402	15 - 1008	0-1-6
C2	57-648	0-90	0-1-2-3-4-5-6-7
C3	6-51	0-18	0-1-2-3-5-6-7
C4	0-5	0-5	0-1-2-3-4
C5	678-3268	0-684	0-1-2-3-4-5-6-7

K-means algoritmasının avantajları uygulanabilirliğinin kolay olması ve büyük veri kümelerinde hızlı çalışabilmesidir. K-means, küme sayısının başlangıçta tanımlanmasını gerektiren bir metottur. Kullanıcının başlangıçta K(küme) sayısını belirleme zorunluluğu vardır. K-means kategorisel özellikler içeren uygulamalarda gerçekleştirilememektedir. Ayrıca k-means metodu küresel olmayan, yoğunlukları farklı olan ve farklı büyüklüklere sahip kümeleri içeren veri kümelerini keşfetmede uygun bir yöntem değildir. Aynı zamanda gürültülü ve sıra dışı verilere duyarlı olmaması k-means algoritmasının diğer bir zayıflığıdır (Işık, 2006).

K-medoids algoritmasında ise k-means algoritmasından farklı olarak başlangıçta rastgele atanan küme merkezleri ile kümenin diğer elemanları swap işlemi yapılarak yer değiştirir. Maliyet hesabı daha iyi oluncaya kadar uygulama devam eder. Çalışmada ara yüz k-means algoritmasında olduğu gibidir. Küme sayısını 5 seçerek elde edilen sonuçlar ve yüzde oranları şekil 9' daki gibidir.

Kümeleme Sonuçları

kumeadi	elemansayisi	toplamsayi
C1	53	%27
C2	33	%17
C3	63	%32
C4	25	%12
C5	20	%10
Top.Sayı:**5**		

Şekil 9. *K-medoids algoritmasının k=5 için küme yüzdeleri*

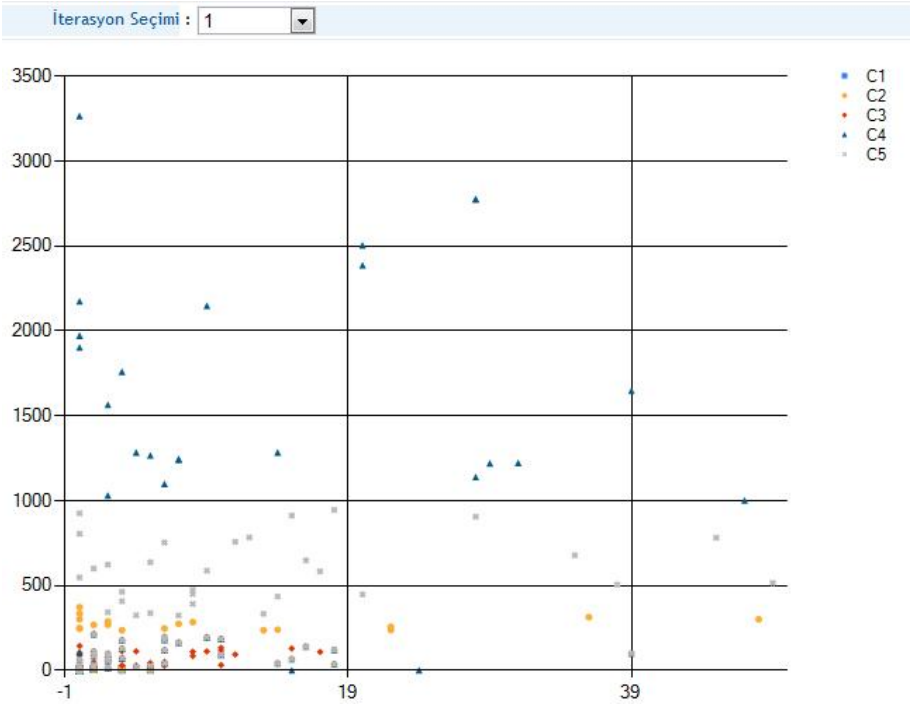
Uygulamanın son bulması için küme içi toplam maliyet hesabı yapılarak en iyi sonuca ulaşılır. Bir sonraki adımda, toplam maliyet bir öncekinden fazla ise yani "0" dan büyükse uygulama son bulur. Şekil 10' da uygulamada ki aşamalar ve her adımdaki toplam maliyet sayıları verilmektedir.

Hata Oranları

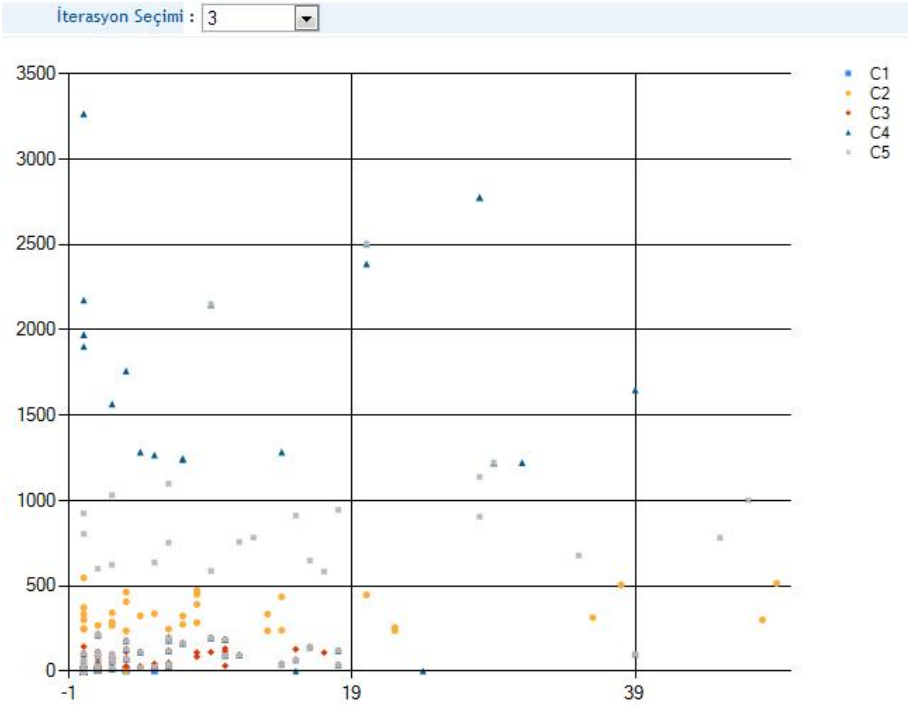
iterasyonsayisi	maliyetsayisi
1	428,214171408665
2	425,169986275008
3	413,323463940511
4	413,870260724655
Top.Sayı:**4**	

Şekil 10. *K-medoids algoritmasının k=5 için her adımdaki hata oranları*

Algoritma toplam 4 adımda en iyi sonuca ulaşmaktadır. Sonuçların 1.ve 3. adımdaki grafiksel gösterimi şekil 11 ve şekil 12' deki gibidir.



Şekil 11. *K-medoids algoritmasının k=5 için 1.adımdaki küme dağılımları*



Şekil 12. *K-medoids algoritmasının k=5 için 3.adımdaki küme dağılımları*

4.adımda 413,870-413,323>0 olduğu için uygulama son bulur. Programı defalarca çalıştırarak en iyi sonuca ulaşılır. Çizelge 4' de k-medoids algoritmasının veri seti üzerindeki sonuçları ve kümelerin dağılımı verilmiştir.

Çizelge 4 *K-medoids algoritmasıyla ülkelerin kümeler arasındaki dağılımları*

Kümeler	Alan	Nüfus	Din
C1	0-11	0-5	0-1-2-3-4-5
C2	236-547	1-118	0-1-2-3-5-6-7
C3	11-215	0-90	0-1-2-3-4-5-6-7
C4	1221-22402	0-1008	0-1-2-4-5-6
C5	583-1139	1-84	0-1-2-3-5

K-medoids, k-means algoritmasında olduğu gibi küme sayısının başlangıçta tanımlanmasını gerektiren bir metottur. Veriye uygun k sayısının belirlenmesi için birden fazla denemenin yapılması gerekir. Kategorisel özellikler içeren veriler gibi bazı uygulamalarda gerçekleştirilememektedir. Ayrıca k-medoids algoritması da farklı büyüklüklerde ve küresel olmayan kümeleri doğru tespit edemez. K-medoids algoritmasının sadece bir öteleme için işlemsel karmaşıklığı O(k (n-k)2)'dir. Bu nedenle, sadece küçük veri setleri için uygun bir yöntemdir (Işık, 2006).

Küme sayısının 5 olarak belirlenmesiyle k-means ve k-medoids algoritmalarının zaman, adım(iterasyon) ve performansları çizelge 5 deki gibi gösterilmiştir. Yapılan denemelerden sonra her iki algoritmada kümeleme başarımları benzerlik gösterse de k-means algoritması gürültülü ve dağınık verileri kümelemede k-medoids algoritması kadar hassas davranamamıştır. En büyük fark ise algoritmaların çalışma zamanları ve adımlar(iterasyon) arasındadır. K-means algoritması daha önce belirtildiği üzere daha hızlı sonuç vermesiyle tercih edilebilen bir algoritmadır. Çalışmalar sonucunda k-means algoritması genel olarak 194 ülke bayrağına ait 3 özelliği 0,030 sn ile 0,069 sn arasında kümelemiştir. Toplam adım(iterasyon) sayıları ise 7 ile 19 arasında değişmektedir. Ülkelerin alanlarını ilk kümede 7690 – 22402, ikinci kümede 3268 – 678 üçüncü kümede 57-648 dördüncü kümede 51 – 6 beşinci kümede 5 - 0 olarak bulunmuştur.

K-medoids kümeleme algoritması ise daha uzun sürede ve kısa adım(iterasyon) sayıları ile sonuca ulaşmaktadır. Aynı veriler için yapılan çalışmalarda küme sayısının 5 seçilmesiyle beraber süre olarak 5,973 sn ile 15,739 sn arasında kümeleme gerçekleşmiştir. Toplam adım(iterasyon) sayıları ise 2-7 arasında değişmektedir. Algoritma ülkelerin alanlarına ait özelliği kümelerken birinci kümede 22402 – 1221, ikinci kümede 1139 – 583, üçüncü kümede 547-236, dördüncü kümede ise 215 – 11, beşinci kümede 11 – 0 olarak bulmuştur.

Çizelge 5. *K-means ve K-medoids algoritmalarının performansları*

		K-Means		*K-Medoids*	
Deney No	**Küme Sayısı**	**Zaman(sn)**	**Adım Sayısı**	**Zaman(sn)**	**Adım Sayısı**
1	5	0,030	7	5,973	2
2	5	0,037	9	11,239	4
3	5	0,039	10	9,581	4
4	5	0,054	13	12,893	5
5	5	0,069	19	15,739	7

Tartışma ve sonuç

Bölümlemeli kümeleme algoritmaları k giriş parametresini alarak n tane nesneyi k tane kümeye böler. Nesneler birbirlerine benziyorlarsa ve başka kümelerdeki nesnelere benzemiyorlarsa, aynı kümeye alınır. Bölümlemeli kümeleme algoritmalarında yöntemler hem uygulanabilirliğinin kolay olması hem de verimli olması nedeniyle iyi sonuçlar üretir (Işık, 2006). Çalışmada ülke bayraklarının gruplara ayrılması kümeleme algoritmaları kullanılarak gerçekleştirildi. Bölümlemeli kümeleme algoritmalarından k-means ve k-medoids algoritmaları kullanılarak performansları karşılaştırılırken her iki algoritma için başlangıç küme sayısı k=5 verilerek kümeler arası ilişkiler incelendi ve k-medoids algoritmasında C1-C2-C3-C4-C5 kümelerinin birbirlerinden daha iyi ayrıldığı gözlemlendi. Kümeler arası kayıplar k-means kümelerine e göre k-medoids algoritmasıyla minimuma indirgemiştir. Bayrak analizi yapılarak ülkelerin yüz ölçümleriyle nüfusları arasında doğru orantı olduğu, benzer yüz ölçümüne sahip olan ülkelerin aynı kümelerde yer aldığı yukarıdaki çizelgelerden anlaşılmaktadır. Çalışmada küme içi benzerlikler maksimum, kümeler arası benzerlikler ise minimuma indirgenmiştir. Her iki algoritma da defalarca çalıştırılarak optimum sonuca ulaşmak mümkündür. Çalışmada 194 veri kümelenmeye çalışılmıştır. K-medoids algoritması bu tür veri setlerinde en uygun algoritma olmasına rağmen büyük veri setlerinde yerini Clara algoritmasına bırakmıştır. Clara algoritması büyük veritabanlarını tarayarak temsilci noktalar seçmek yerine, veritabanından rastgele bir örnek kümeyi alarak k-medoids algoritmasını bu örnek küme üzerine uygular(Kauffman ve Rousseeuw, 1990). Bu çalışma ile k-means ve k-medoids algoritmalarının çalışma zamanlarına bakılarak bir karşılaştırma yapılmış ve k-medoid algoritmasının k-means'e göre daha yavaş çalıştığı gözlemlenmiştir. Aynı zamanda k-medoids algoritmasının ise kümeleme başarımında k-means algoritmasına göre daha etkili olduğu da yapılan çalışmalar sonucunda elde edilmiştir. Uygulamalarda adım(iterasyon) sayıları da zamana bağlı olarak doğru bir orantı göstermektedir. Sonuç olarak k-medoids kümeleme algoritması dağınık verileri kümelede k-means algoritmasına göre zaman açısından bakıldığında daha yavaş çalışsa bile daha iyi sonuç verdiği gözlemlenmiştir. Geliştirilen sistem web ara yüzü ile internet erişiminin olduğu her yerden tek bir bilgisayara bağlı kalmadan analiz edilebilmektedir.

Kaynaklar

Akın, K. Y. 2008. Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi. Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 164s.

Alpaydın, E. Zeki veri madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Bilişim 2000 Eğitim Semineri, 2000.

Amasyalı, F. M., Ersoy, O. 2008. Kümeleyici Topluluklarının Başarısını Etkileyen Faktörler, IEEE 16th Signal Processing and Communication Applications Conference, SIU 2008, Aydın.

Arslan, H. 2008. Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya, 80s.

Berkhin, Pavel. 2004. Survey of Clustering Data Mining Techniques. http://citeseer.nj.nec.com/berkhin02survey.html. (Erişim Tarihi: 07.04.2004).

Bozkır, S. A., Gök, B., Sezer, E. 2009. 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük.

Bülbül, Ş. 2009. Propensity Skor Uygulamalarında Kümeleme Analizinin Test Amaçlı Kullanımı. 10. Ekonometri ve İstatistik Sempozyumu, Atatürk Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, 27-29 Mayıs, Palandöken-Erzurum.

Dinçer, E. 2006. Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması. Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli, 101s.

Dolgun, Ö. M., Özdemir, G. T., Oğuz, D. 2009. Veri Madenciliğinde Yapısal Olmayan Verinin Analizi:Metin ve Web Madenciliği. İstatikçiler Dergisi, 2, 48-58.

Evans, S., Lloyd, J., Stoddard, G., Nekeber, J., Samone, M. 2005. Risk Factors For Adverse Drug Events. The Annals of Pharmacotherapy, 39, 1161-1168.

Gersho, A., Gray, R.M. 1991. Vector Quantization and Signal Compression. Kluwer Academic Publishers Norwell, USA, 738pp.

Işık, M. 2006. Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları. Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, İstanbul, 76.

Jain, A.K. Murty, M.N. Flynn, P.J. 1999. Data Clustering: A Review, *ACM Computing Surveys*. 3, 31.

Kauffman, L., Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, 342 pp.

Linde, Y., Buzo, A. 1980. An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, January, 702-710.

Mit, 2005. www.eecs.mit.edu.tr (Erişim tarihi: 18 mayıs 2005)

Pang-Ning Tan, P. N., Steinbach, M., Kumar, V. 2006. Introduction to Data Mining. Addison Wesley, 769 pp.