Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü
Sayılsal Yöntemler Anabilim Dalı

VERİ MADENCİLİĞİNDE KÜMELEME ANALİZİ VE SAĞLIK SEKTÖRÜNDE BİR UYGULAMASI

Selim ÇAM

Yüksek Lisans

Tez Danışmanı
Doç.Dr. Hüdaverdi BİRCAN

Sivas
Şubat - 2014

VERİ MADENCİLİĞİNDE KÜMELEME ANALİZİ VE SAĞLIK SEKTÖRÜNDE BİR UYGULAMASI

Selim ÇAM

Cumhuriyet Üniversitesi
Sosyal Bilimler Enstitüsü

Lisansüstü Eğitim, Öğretim ve Sınav Yönetmeliğinin Sayısal Yöntemler Anabilim Dalı İçin Öngörüdüğü

YÜKSEK LİSANS TEZİ
Olarak Hazırlanmıştır.

Sivas
Şubat - 2014

KABUL VE ONAY

Üniversite:	**: Cumhuriyet Üniversitesi**
Enstitü	**: Sosyal Bilimler Enstitüsü**
Ana Bilim Dalı	**: İşletme Ana Bilim Dalı**
Bilim Dalı	**: Sayısal Yöntemler**
Tezin Başlığı	**: Veri Madenciliğinde Kümeleme Analizi ve Sağlık Sektöründe Bir Uygulaması**
Savunma Tarihi	**: 17.02.2014**
Danışmanı	**: Doç. Dr. Hüdaverdi BİRCAN**

Unvanı - Adı Soyadı		**İmza**
Jüri Başkanı:	Prof. Dr. Mahmut KARTAL	
Üye	: Doç. Dr. Hüdaverdi BİRCAN	
Üye	: Yrd. Doç. Dr. Mehmet Ali ALAN	

Oy Birliği X

Oy Çokluğu ☐

Selim ÇAM tarafından hazırlanan Veri Madenciliğinde Kümeleme Analizi ve Sağlık Sektöründe Bir Uygulaması başlıklı tez, kabul edilmiştir. 26/02/2014

Prof. Dr. Alim YILDIZ
Enstitü Müdürü

TEŞEKKÜR

Yüksek Lisans çalışmamın her aşamasında bilgi ve deneyimleri ile beni yönlendiren, beni araştırmaya yönelten ve hiçbir yardımını benden esirgemeyen danışman hocamDoç.Dr. Hüdaverdi BİRCAN'a teşekkürlerimi sunarım. Ayrıca, tüm bu süre içerisinde beni destekleyen ve daima yanımda olarak başarıya ulaşmamı sağlayan Ailem' e, Sayın Arş.Gör. Murat Fatih TUNA'ya sonsuz teşekkürlerimi sunarım.

ÖZET

ÇAM, Selim,. Veri Madenciliğinde Kümeleme Analizi ve Sağlık Sektöründe Bir Uygulama, Yüksek Lisans Tezi, Sivas, 2014.

Veri madenciliği, büyük ölçekteki veriler arasından üstü kapalı, çok net olmayan fakat potansiyel olarak kullanışlı olabilecek bilgilerin ortaya çıkartılmasında önemli rol oynamaktadır. Veri madenciliği günümüzde, teknolojinin gelişmesiyle birlikte yaygınlaşarak birçok disiplin içerisinde yer alan ve kabul gören bir metot halini almıştır

Bu çalışmada, Cumhuriyet Üniversitesi Hastanesi'ne 2011 yılında başvurmuş olan hastaların 2006-2011 arasındaki kayıtlar, hasta başvuru davranışlarının belirlenmesi amacıyla incelenmiştir. Bu bağlamda, iş gücü potansiyeli olan 18-65 yaş arasındaki hastaların verilerine veri madenciliği kümeleme analizi yöntemlerinden olan K-Ortalamalar ve Yoğunluk Tabanlı Kümeleme Algoritması yöntemleri uygulanmıştır. Veri tabanından alınan hasta verileri analiz sonucunda hastaların demografik verileri Ki-Kare, Kruskall-Wallis H ve Mann-Whitney U testleri ile incelenerek, demografik verilerin oluşturulan kümelerle etkileşimde olup olmadığı ortaya konulmuştur.

Yapılan analizler sonucunda oluşturulan kümelerde, hastaların yaşlarının kümelere göre farklılık gösterdiği görülmüştür. Ayrıca genç hastaların daha farklı teşhisler ile ileri yaşlardaki hastaların ise daha az çeşitli hastalıklarla hastaneye başvurdukları ortaya çıkartılmıştır. Hastaların Sivas ili sınırlarından gelmesi farklılaşma için bir etken olabildiği gibi, yaşanılan yerler de (köy, kasaba, ilçe, şehir merkezi vb.) kümelere göre farklılık göstermektedir.

Bu çalışma esas itibariyle hastaneye başvuru yapan bir hasta için tedavi prosedürleri, personel tahsisi, ilaç ve tıbbi malzeme ihtiyaçlarının düzenlenmesi gibi hususlarda yardımcı bir nitelik taşımaktadır.

Anahtar Kelimeler: Veri, Bilgi, Veri Madenciliği, Yoğunluk Taban Kümeleme Algoritması, K-Ortalamalar Yöntemi

ABSTRACT

ÇAM, Selim,. Clustering Analysis in Data Mining and An Application in Health Sector.,Master's Thesis, Sivas, 2014.

Data mining plays an important role in extracting veiled, obscure, but potentially available knowledge from large data stacks. Nowadays, data mining has become widespread accompanied by technological advancement, following its entity as a multidisciplinary and well accepted method.

In order to determination patients' appeal behaviors, records of the year 2011 belonging to the patients appealed to Cumhuriyet University Hospital were analyzed in this study. In this context, K-Means and Density Based Clustering Algorithms which are from data mining methods of clustering analysis were conducted. In consequence of analysis of patients' data derived from demographic parameters were statistically examined by Chi-Square, Kruskall Wallis H and Mann Whitney U test and ultimately, it was determined that whether the demographic parameters interact with created clusters or not.

As a result of the analysis, it was seen that patients' ages differentiate into clusters. Additionally, it was revealed that young patients apply to hospital with more different diagnosis as well as old patients with less various diseases. Patients' application from provincial border of Sivas can be a factor for differentiation and residential places (village, town, district, city center, etc.) differentiate as well.

This study has helping characteristics in point of providing treatment procedures for an appealed patient, proper personal allocation and organization of medical material needs.

Key Words: Data, Knowledge, Data Mining, Density Based Clustering, K-Means Method.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

Veri madenciliği, geniş veri yığınları içerisinden, ihtiyaç duyulan noktalarda kullanılabilme ihtimali olan, ortaya çıkmamış bilgilerin ve ilişkilerin anlaşılır ve kullanılabilir bir şekilde araştırılması yöntemlerine verilen genel bir analiz yöntemidir. Başka bir bakışla: "daha önceden bilinemeyen geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu kararların işletme kararları verirken kullanılmasıdır" (Silahtaroğlu, 2008, s.10).

Veri madenciliği incelendiğinde yeni bir kavramla karşılaştığımızı söyleyemeyiz. Milattan önce yaşamış ve varlığı bilinen gelişmiş devlet modellerinde (Antik Çin, Yunan) askeri ve mali alanlarda yöneticilerin basit anlamda istatistiği kullandığı görülmektedir (Nisbet vd., 2009, s.4). Aynı durum 1600-1700'lü yıllarda gelişmekte olan Avrupa ülkelerinde şans oyunlarının anlaşılması ve kazanılması bağlamında olasılık teorisinin ortaya çıkmasına sebep olmuştur (Korkmaz, 2005). Son dönemlerde ise veri işleme 1950'lerin devasa boyutlardaki ilk bilgisayarlarıyla başlayan, günümüzde cebimize sığabilecek kadar küçültülmüş bilgisayarlarla yapılabilmektedir.

Temelde, istatistik yöntemler olmadan veri madenciliğinden bahsedemeyiz, ancak tek başına da istatistik günümüz veri madenciliği için yetersizdir. Veri madenciliğiningelişmesindeki diğer yardımcı faktörler; yapay zeka (artificialintelligence) ve makine öğrenimidir (machinelearning).

İşletmelerin ortak yönlerinden birisi de başarılı bir yönetim için bilgiye gereksinim duymalarıdır (Tuna, 2013, s.6). Veri madenciliğini uygulamadaki amaç, bir firmanın, işletmenin ya da kurumun sahip olduğu veriden faydalı olabilecek bilgilere ulaşması böylelikle kaynakları daha etkin kullanmak, tedarik noktalarını ya da potansiyel yatırımları, faydaları, durumları ortaya çıkartarak mikro ekonomi oluşturmaktır. Dolayısıyla, pazar payı, müşteri memnuniyeti, karlılık gibi bir şirketteki kritik noktalar hakkında bilgi sahibi olunabilir ve daha esnek, etkin kararlar alınabilir.

Veri madenciliğinin temel tanımından yolaçıkacak olunursa, veri madenciliğinin bir tahminleme ya da ispatlama yöntemi değil, daha önceden bilinmeyeni ortaya çıkarmasıyla diğer benzetilen analiz yöntemlerinden farklı olduğu görülür (Silahtaroğlu, 2008, s.10).

Bilişim ve teknoloji sektörlerindeki gelişmeyi ele alacak olursak diğer sektörlerdeki durumunda pek farklı olmayacağı anlaşılır. TÜİK verilerine göre Türkiye'de 2011 yılında 143.706 milyon $'lık ihracat rakamına ulaşılmıştır. Bu rakamdan yola çıkarak, satılan mallar, satış yolları, bu malların üretim sırasındaki hammadde ve işgücü v.b. parametreler düşünüldüğünde klasik yollarla hesaplanamayacak bir veri yığınıyla karşılaşırız. Ulaşılan teknoloji düzeyiyle daha büyük veri tabanları oluşturup, daha hızlı ve az maliyetle bilgiye ulaşabilmekteyiz. Ancak her ne kadar gelişmiş sistemlere sahip olunsa da veri artışına yetişilememektedir. Gmail şirketi, üye sayısı olarak 8 milyonu aşmış durumdadır ve herbir üye için 7673,464572 megabayt (27.01.2012 itibariyle) kullanım alanı açmıştır.

Dünyada veri madenciliğini kullanan birçok şirket vardır ve bazıları şunlardır: Amazon.Com, APEX, Barclays, Carrefour, CopenhagenEnergy, FORTIS BANK, IMS Health, JCB CO, Laurentian Bank, Moscow City Telephone Network, Kuzey Carolina Eyalet Hazine Dairesi, Octopus, Office Depot, OneWorldHealth, Siemens, Sovereign Bank, Vodafone, Hepsiburada.com, Migros v.b.

Türkiye'de ise veri madenciliğini kullanan şirketlerden bazıları şunlardır: Aviva Sigorta, AXA Sigorta, Garanti Bankası, Eureko Sigorta, HSBC Bank A.Ş., Kıraça Şirketler Topluluğu, Türk Ekonomi Bankası, Turkcell, Yapı Kredi, v.b.

Özel sektörde her nekadar veri madenciliği kullanılan bir karar verme yöntemi olarak uygulansa da kamu kurumlarında bu tarz bir çalışmanın etkili bir şekilde kullanıldığından bahsetmek zordur. Özellikle çalışmanın merkezinde yer alan sağlık sektöründe veri madenciliği yöntemlerinin kilit bir öneme sahip olduğu düşünülmektedir. Uluslararası hastalık sınıflandırma sistemine (ICD-10, International

Statistical Classification of DiseasesandRelatedHealthProblems) göre 11332 adet (http://hastane.dicle.edu.tr/index.php?option=com_content&view=article&id=125:icd, 11.08.2012) hastalık tanısı vardır. Bu hastalıklara karşılık gelen hastalar ve demografik bilgiler gözönüne alınacak olunursa elde edilen verilerin analizi veri madenciliği yöntemleriyle hesaplama yapmayı zorunlu kılmaktadır. Türkiye'nin nüfus, iklimi v.b. özellikleri bakımından bölgesel olarak hastalıkların çeşidi, kronik olup olmaması değişiklik göstermektedir. Dolayısıyla tedavide yer alacak hekim ve nitelikli personel sayısının, ilaç ve tıbbi malzeme bulundurulmasının etkili bir şekilde planlanması, ayrıca ülke ekonomisine de yük getirmeyecek bir düzenlemenin yapılması gerekliliği doğmaktadır.

Veri madenciliği uygulamaları sağlık sektöründe farklı yöntemler ve konu başlıkları altında incelenmiştir. Literatürde hastaların teşhis, tedavi süreçleriyle ilgili verileri toplanarak hastalık tahminleri; hastaların hastaneye başvuruları ile hasta davranış tahminleri yapılmıştır. Örneğin Kaur ve Wasan. 2006 yılında yaptıkları bir çalışmada hastaların yaş, cinsiyet, belirti sıklığı, kan tahlili sonuçları ve vücut kitle endeksi gibi verilerini alarak hastalarda herhangi bir hastalık olup olmadığını araştırmışlardır.

Nagadevara 2004 yılında Uluslararası e-Yönetim Konferansı'ndaki bildirisinde hastaların demografik ve çevresel bilgileri analizi sonuçlarını sunmuştur. Çalışmasında sinir ağlarını kullanarak sonuçlara ulaşmıştır. Benzer şekilde Ertuğrul v.d 2013 yılında. Pamukkale Üniversitesi Hastanesi'ne gelen hastaların hastaneye geldikleri zamanı, demografik verileri, poliklinik verilerini alarak hastaneye gelen 2009-2011 dönemindeki hastaların hasta profilini belirlemişlerdir.

Mullins vd. (2006), 667.000 hastaya ait verilerle yaptıkları çalışmalarında, tıbbi akademik bir veri madenciliği yaklaşımı olan HealthMiner® adını verdikleri tescilli bir sistem aracılığıyla uyguladıkları CliniMiner® isimli denetimsiz bir veri madenciliği metodu kullanmışlardır. Demografik, sosyo-ekonomik ve klinik verilerle, seçilen vakalardaki biyolojik çıktıların değerlendirdiği çalışmanın sonucunda, farklı klinik

hastalıklara yönelik tahmin analizleri aracılığıyla bilgi ve örüntüleri ortaya çıkarmışlardır.

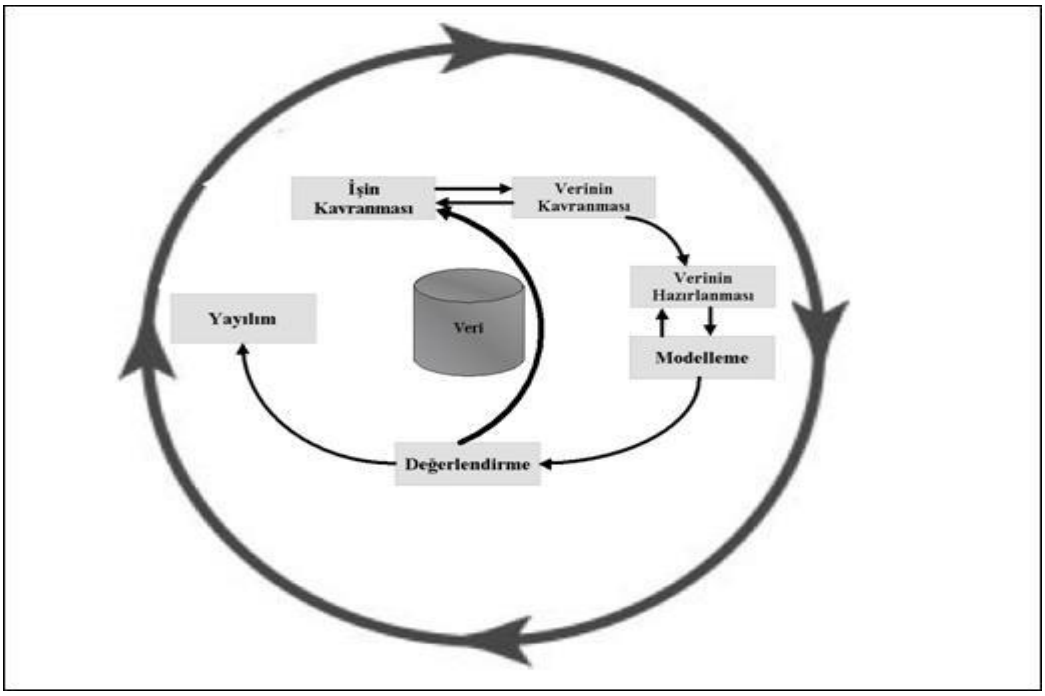
Koh ve Tan (2011), yaş, beden kitle endeksi, bel-kalça ölçüsü, haftada yapılan egzersiz sayısı değişkenlerini kullanarak veri madenciliği yöntemlerinden birisi olan karar ağaçları yönteminden faydalanarak analizler yapmışlardır. Çalışmanın sonuçları ile geliştirilen müşteri ilişkileri yönetim metodu, elde edilen sınıflara ait farklı hizmet tipleri geliştirilmelerine olanak tanımıştır.

Bu çalışmada Sivas ilindekiuygulama ve araştırma hastanesinin veri tabanlarından alınan bilgilerle, bu Sivas ilinden ve komşu illerden hastaneye başvurmuş olan hastaların bilgileri çıkartılması amaçlanmıştır. Çalışmada öncelikle veri madenciliği hakkındaki temel kavramlardan bahsedilerek veri madenciliği yöntemleri, teknikleri hakkında bilgiler verilecek ve ilerleyen konularda veri madenciliğinde kümeleme analizi teorik bir çerçevede incelenip, veri tabanından alınan veriler farklı tekniklerle analiz edilecektir.

2.GENEL BİLGİLER

Günümüzde her insan bir veri yığını tarafından çevrelenmiştir(Witten,2011, s.3). Ancak insanlar bu durumun o kadar da farkında gözükmemektedir. Bir ürün yada hizmet aldığında, kredi kartı kullanıldığında, havale işlemi gerçekleştirildiğinde, elektronik posta adresleri kullanıldığında, internet üzerinden bir bilgiyi paylaşırken, v.b. birçok işlemde bilgiler veri tabanlarına işlenmektedir.

Hesap yapma konusunda insalara yardımcı olan ve bilgisayarın atası kabul edilen abaküs'ten buyana teknolojik gelişmeler baş döndürücü bir hal almıştır. Üstelik bu gelişmelerin sonucu, kullanıcıları bunaltmaktadır. Elektronik posta adreslerine gönderilen reklamlar, cep telefonlarına gönderilen mesajlar bunlara verilebilecek sadece iki örnektir. Ancak şu durum da unutulmamalıdır ki; veri işleme tekniklerinin gelişmesiyle; hastalık riskleri tahminedilebilir, dolandırıcılık faaliyetleri azaltılabilir, terör olayları arasındaki bağlantılar ortaya çıkartılarak terörist eylemlerin önlenmesi için harekete geçilebilir, internet ortamında kötü amaçlı yazılımlar ortaya çıkartılabilir(Awad, 2009, s.1).



Şekil 1.Veri Madenciliği Süreci Yaşam Döngüsü

(Kaynak: http://okul.selyam.net/pars_docs/refs/25/24645/24645_html_m4a91cb91.png, 15.08.2012)

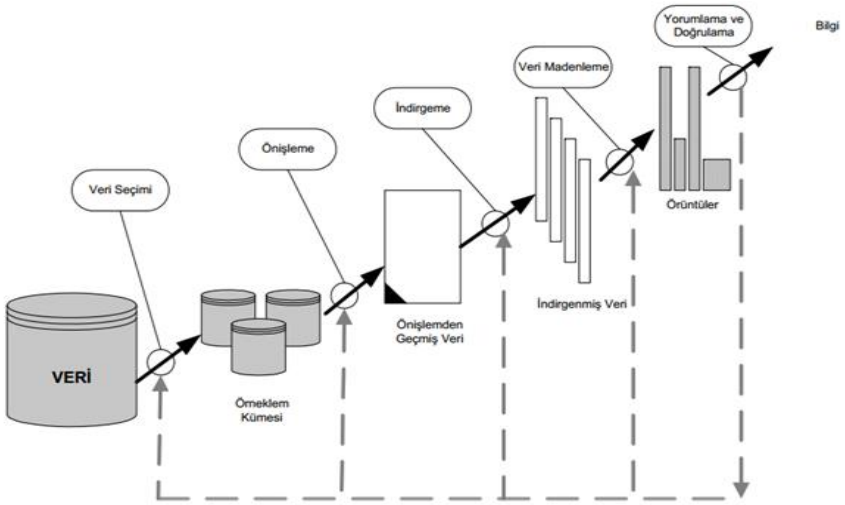
Şekil 1'de de görüldüğü gibi veri madenciliği kavramının merkezinde veri vardır. Veri işleme işin kavranması adımıyla başlamaktadır. Analizi gerçekleştirecek olan çalışmacılar ya da araştırmacılar öncelikle işi ve veriyi kavramalıdırlar. Hazırlanma ve modelleme adımları sadece uygulama esaslarını içeren ara basamaklardır. Değerlendirme adımında sonuçlar değerlendirilir ve uygun sonuçlar elde edilmesi durumunda, sonuçlar işletmeler için kullanılabilecek bireylere iletilir.

Veri madenciliğini konusunu teorik çerçevede incelemeden önce sadece veri ve bilgi kavramlarını değil teknolojik gelişmelerle birlikte ortaya çıkan veri tabanı, veri ambarı gibi temel kavramların ve özelliklerinin de bilinmesi gerekmektedir. En kısa haliyle: gerçekleştirilen bir işlem sonucu elde edilen "veri", "veri ambarı"ndaki bir "veri tabanı"na kaydedilir; matematiksel ve istatistiksel işlemlerle "bilgi" halini alır ve oluşturulan bilgi istenirse tekrardan veri olarak bu sürece dahil edilir.

2.1.VERİ, BİLGİ VE ÖĞRENME

Günlük yaşantıda karşılaşılan olaylar sonucu sayısal, metinsel, görsel verilerle karşılaşırız. Ancak karşılaşılan veriler bilgi olarak kullanılabilir yada kullanılamaz. Bu yönüyle veriye ham gerçek de denilebilir. Ancak veriler bilgelik sürecindeki temel taştır. Kayıt altına alınan veriler analiz edilebilir, işlenebilir ve bilgiye dönüştürülebilir.

Verinin kullanılışı ve işlenmesi açısından günümüz teknolojileri birçok açıdan yardıma koşmaktadır.Buradaki amaç klasik yöntemlerle yani elle yapılan hesaplamalarla, veri madenciliğinin yapılmasının güçlüğüdür. Tezin konusu itibariyle kişilerin aldığı hizmete karşılık veri tabanındakigirdiler düşünülecek olunursa, herhangi bir hastanede bir kişiye yapılan hizmet en azından: hastanın yaşı, doğum yeri, adresi, sağlık güvencesinin türü, anne ve baba adı, kimlik numarası, telefon numarası, hastaneye geliş tarihi, muayene olunan poliklinik, muayene eden doktor bilgilerini içermektedir. Bu noktada tek kişinin verilerinin bir şey ifade etmediği görülür.



Şekil 2. Verinin Bilgiye Dönüşümü

(Kaynak: http://binedir.com/blogs/veri-madenciligi/image_6e9f15ed.png, 26.08.2012)

Tek başlarına anlam ifade etmeyen verilerin, matematiksel ve istatistiksel yöntemlerle işlenerek kurumlar ya da şirketler için kullanılabilir hale getirilmiş haline

de bilgi denmektedir. Kavram olarak bilgi; bir şeyin bilinmesi, kabul edilmesi yada belgelenmesi şeklindedir. Bu adımlar bilginin üç faktörü olduğunu ortaya koymaktadır(Grünberg, 2005). Üç faktörü veri madenciliğine uyarlarsak ham gerçeklerin doğruluğunu kabul edip işleme tabi tutarak doğruluk noktasında ispatlanmış bilgilere ulaşmaktayız (Şekil 2). Ancak şu durum unutulmamalıdır ki, elde edilen bilgi kullanışlı olabilir yada olmayabilir.

2.2. VERİTABANI

Verilerin kayıt altına alınabilmesi günümüz teknolojisiyle daha kolaylaşmıştır. Buna rağmen verinin düzgün kaydedilebilmesi, insan hatalarının enaza indirilebilmesi amacıyla ek sistemlere, yöntemlere gereksinim doğmuştur. Bu noktadaveritabanı sistemleri geliştirilmiş ve veriyi yönetmek için Veritabanı Yönetim Sistemleri yaklaşımı ortaya çıkmıştır. Veritabanı sistemlerinin temel amacı olarakşu maddeler söylenebilir(Özkan, 2008, 14):

- ✓ Girdilerin tamamının görülebilmesi
- ✓ Kullanıcılar için yapıların ve bilgilerin kolay anlaşılabilmesi
- ✓ İleriki analiz adımlarında ilişkisel durumların ortaya çıkarılabilmesi
- ✓ Süreç işlemlerinin ihtiyaçlarının ve sonuçlarının etkinliğinin arttırılması
- ✓ Sorgu ve işlem formülasyonunun mantıksal bağımsızlığının elde edilebilmesi (Thalheim,2000, s.1)

Veritabanı sistemleri genelde, kullanıcılara insan hatasının azaldığını göstermiştir. Elle yapılan işlemlerde yada elektronik ortamda olsa dahi birçok tablonun olması insan hatasını arttırmaktadır. Bunun yanısıra şirketlerin etkin ve hızlı karar alması gerektiğini düşünürsek, zaman kaybının ve verisayısı fazlalığınıntolere edilmesi kolay olmamaktadır. Veritabanı sistemlerinin uygulanmasıyla (Özkan, 2008, s.15):

- ✓ Veri tekrarıyla karşılanmaz; veriler veritabanında tek kayıttan kullanılmak üzere tasarlanmıştır. Böylelikle yüksek hacimli kayıt ortamlarına ihtiyaç önlenmiş olur.
- ✓ Tutarlılık sağlar; farklı tablolara yapılan aynı girdiler insan hatasına sebep olabilmektedir. Örneğin hastaneye gelen bir hastaya farklı polikliniklerden kayıt açılması durumunda;hastaya ait verilerin birinin bile yanlış girilmesi hastanın başka bir hasta olarak tanımlanması yani mükerrer kayıt olmasıyada işlem açılamaması demektir. Ayrıca geriye dönük yapılacak analizlerde sonuçların yanlış yorumlanmasına sebep olmaktadır. Bu durum özel sektördeki şirketlerde de karar vericiler için kritik hatalara sebep olmaktadır.
- ✓ Denetim sağlar; veriler ortak kullanıma açık olduklarından sorgu denetimleriyle veriler üzerindeki hatalar ortadan kaldırılabilir. Tekrardan bir hastanın kimlik numarası örneğinin verecek olursak: hastanın 11 haneli kimlik numarasının az yada daha fazla olarak girilmesi durumunda sistem otomatik olarak uyarı verir veya işlem yapmaz. Denetleme yöntemleri olarak pazarlama, stok kontrolde örnek gösterilebilir.

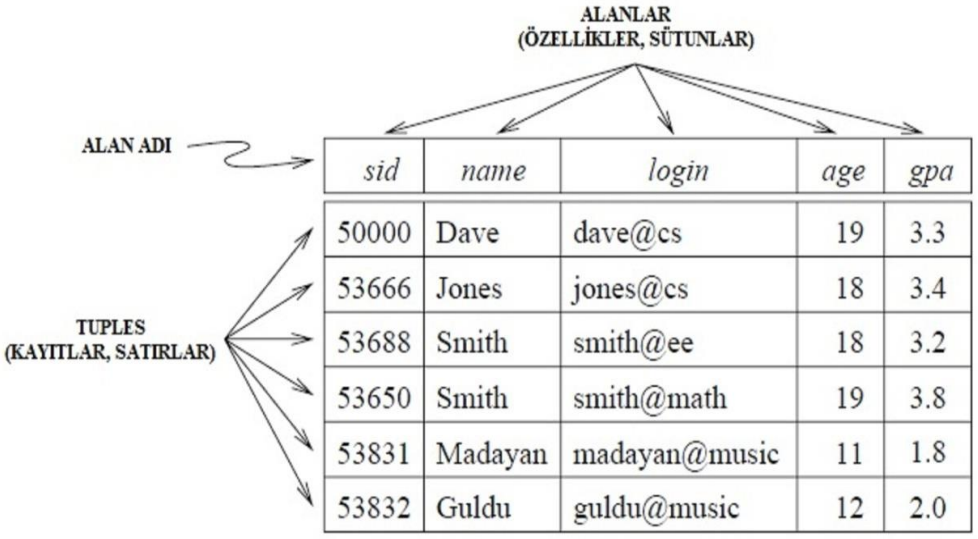
Veri tabanı konusunu anlayabilmek için veri modeli kavramından da bahsedilmesi gerekir. Veritabanları yapısal olarak aynı değildirler(Özkan, 2008, s.16). Veritabanında asıl önemli kavram, kayıt yığını ya da bilgi parçalarının tanımlanmasıdır. Bu tanıma Şema adı verilir. Şema veritabanında kullanılacak bilgi tanımlarının nasıl modelleneceğini gösterir (Silberschatz vd., 2010, s.8). Buna Veri Modeli (*Data Model*) yapılan işleme de Veri modelleme denir.Verilere erişim ve verileri depolama bakımından farklılık gösterirler. Bunları 3 ana başlık (Hiyerarşik veritabanları, ilişkisel veritabanları ve nesneye yönelik veritabanları) altında toplayabiliriz. Şuan ençok kullanılan veri modeli ilişkisel veri modellemesidir. İlişkiselveritabanı mantığı 1970 yılında Dr. EdgarF. Codd tarafından yazılan "A Relational Model of Data forLargeShared Data Banks" adlı makaledeortaya atılmıştır (Garcia vd., 2008, s.4).

İlişkisel veri modeli ikiboyutlu tablolardan oluşan bir yapıdır (Garcia vd., 2008, s.61). Veri tabanı oluşturulurken tablolara ve onların özelliklerinin olduğu şemaya ihtiyaç vardır. Örnek olarak;

Öğrenciler(sid: string, name: string, login: string, age: integer, gpa: real) gösterilebilir.

("string" (harf), "integer" (tam sayı) ve "real" (ondalıklı sayı) terimleri veri tabanına kayıt yapılabilmesi için tanımlanmış veri giriş alanlarının özellikleridir.)

Tablolar satırlardan (kayıt, "tuples") ve sütunlardan (alan, özellik) oluşur. Ayrıca veri tabanından bahsederken "domain"(etki alanı) kavramını da açıklamak gerekir (Şekil 3). Etki alanı, bir tablodaki verilerin özelliklerine göre alabileceği değerleri ifade eder. Örneğin: Şekil3'de görülen "name" sütunun özelliği eğer metinsel (string) değer olarak atanmışsa rakam girilememesi; "age" sütunu "integer" ifade olarak atanmış ise "float" bir rakamın o sütuna yazılamaması gibi açıklanabilir. Buradan yola çıkarak, domain şemalarla ve tablolarla direk ilişkilidir, verilerin alabileceği değerlerin sınırlarını belirlemesi açısında da "atomik"tir, denilebilir (Elmasri ve Navathe, 2003, s.127).



sid	*name*	*login*	*age*	*gpa*
50000	Dave	dave@cs	19	3.3
53666	Jones	jones@cs	18	3.4
53688	Smith	smith@ee	18	3.2
53650	Smith	smith@math	19	3.8
53831	Madayan	madayan@music	11	1.8
53832	Guldu	guldu@music	12	2.0

Şekil 3. İlişkisel Veri Tabanı Örneği

(Kaynak: Ramakrishnan, R. ve Gehrke, J. (2000). *Database Management Systems (Second Edution).* California, USA, McGraw-Hill.)

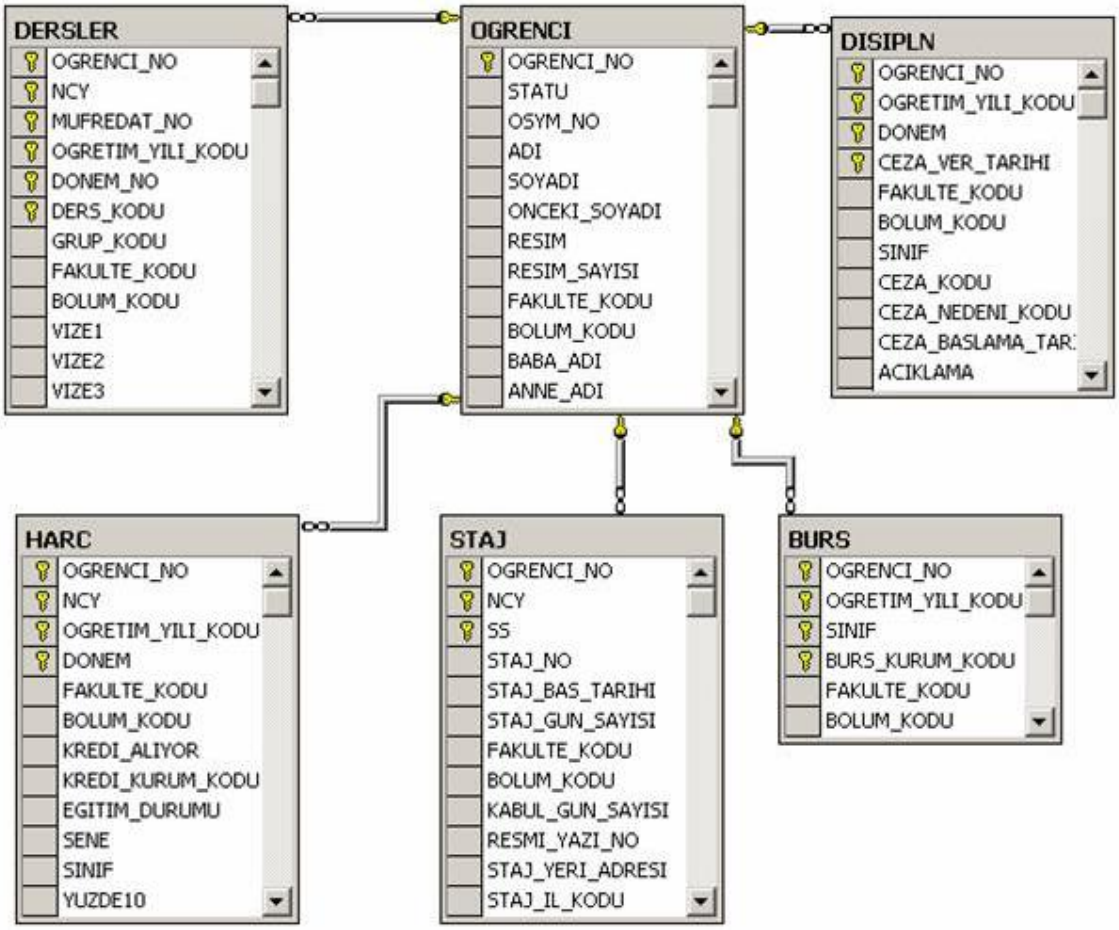
Kısaca ilişkisel veritabanlarındaki tabloların özelliklerini şu şekilde sıralayabiliriz (Silberschatz vd., 2010, s.12):

1. İki boyutludur; sütunlardan ve satırlardan oluşur.
2. Sütunların ayrı adları olmalıdır.
3. Satırlar birbirinden faklıdır.
4. Sütunlar, aynı "domain"deki (etki alanının) değeri alabilir.
5. Satırların sırası önemsizdir.
6. Sütunların sırası önemsizdir.

Satırların ya da sütunların kendi içlerinde yerdeğiştirmesi veri tabanını etkilemez.

İlişkisel veri tabanlarında, veriler tablolarda saklanır ve bu tablolar anahtar alan (primerykey) ile birbirlerine bağlanabilir. Böylelikle, verilerindoğru ve etkin bir biçimde saklanılması, veri bütünlüğü ve herhangi bir problemde verilerin kurtarılması sağlanır; ilişkisel veritabanı modelinin amacı olan yüksek verimlilik sağlanmış olur. Aşağıdaki Şekil 4'te ilişkisel veri modellemesiyle hazırlanmış tabloların birbirlerine

bağlanması görülebilmektedir. Ayrıca anahtar şeklindeki simge ilgili veri tabanındaki anahtar alanı (primary key) temsil etmektedir.



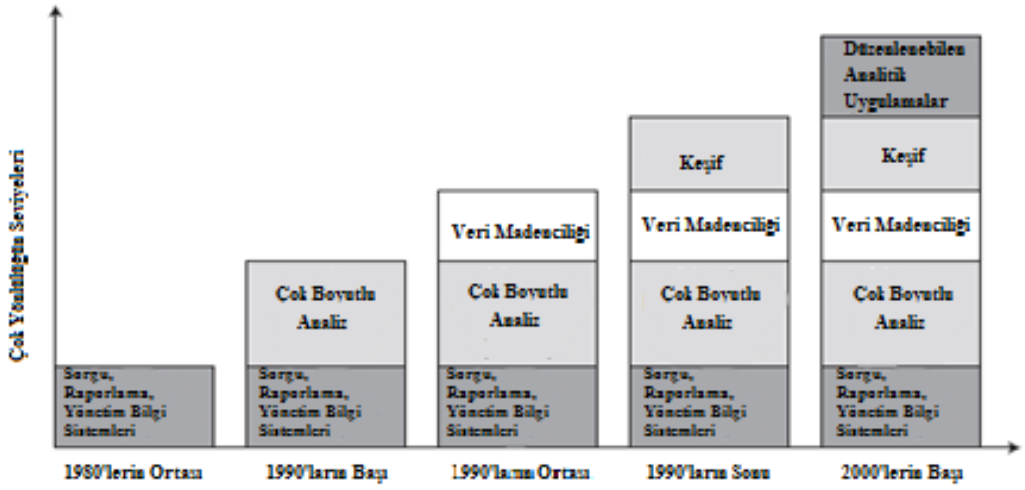
Şekil 4. İlişkisel Veri Tabanına Tablolar Bazında Örnek

(Kaynak: http://www.yazgelistir.com/Makaleler/Resimler/1000000753_image002.jpg, 01.02.2013)

2.3. VERİ AMBARI

Veri ambarı,sadece organizasyonlarda bulunanişlemsel düzeydeki personele değil, karar verici (yönetici) kişiler için de tasarlanmış ortamdır, denilebilir. Esasen veri ambarı kuruluş çapında raporlama ve analiz için tutarlı bir kaynak sağlayan veri tabanıdır. Günümüz koşullarında veri tabanı sistemlerinin karar destek uygulamalarında etkinliği tartışılmaktadır (Özkan, 2008, s.20). Böylelikle veri tabanın karşılaşılan ilgilenilen verinin boyut sorunuveri ambarı ile çözülebilmektedir.

Veri ambarının işlemleri de güncel olarak "tedarik zinciri"ne benzetilebilir. Veri üretimi, veri tedarikçilerinden sağlanan verilerin (operasyonel sistemler ve dış kaynaklar) geçici olarak bir veri ambarında depolanması şeklindedir(Moody ve Kortink, 2000). Görüldüğü gibi veri ambarı, giderek karmaşıklaşan ortamlardan düzgün bilginin elde edilmesi sürecinde karar vericilere destek olmaktadır. Bu sebeple veri ambarı, iş zekası (Business Intelligence) kavramın ayrılmaz bir parçasıdır. Basit raporlama ve üst yönetim bilgi sistemleri için, istatistik, veri madenciliği, özelleştirilebilir analitik uygulamlar ve çok boyutlu analiz günümüz şartlarında verinin bilgiye dönüştürülmesi sürecinde kullanılan yöntemlerdir (Imhoff vd., 2003, s.10). Veri madenciliğinin gelişimini Şekil 5'te görmek mümkündür.



Şekil 5.İşletme Zekasının Gelişimi

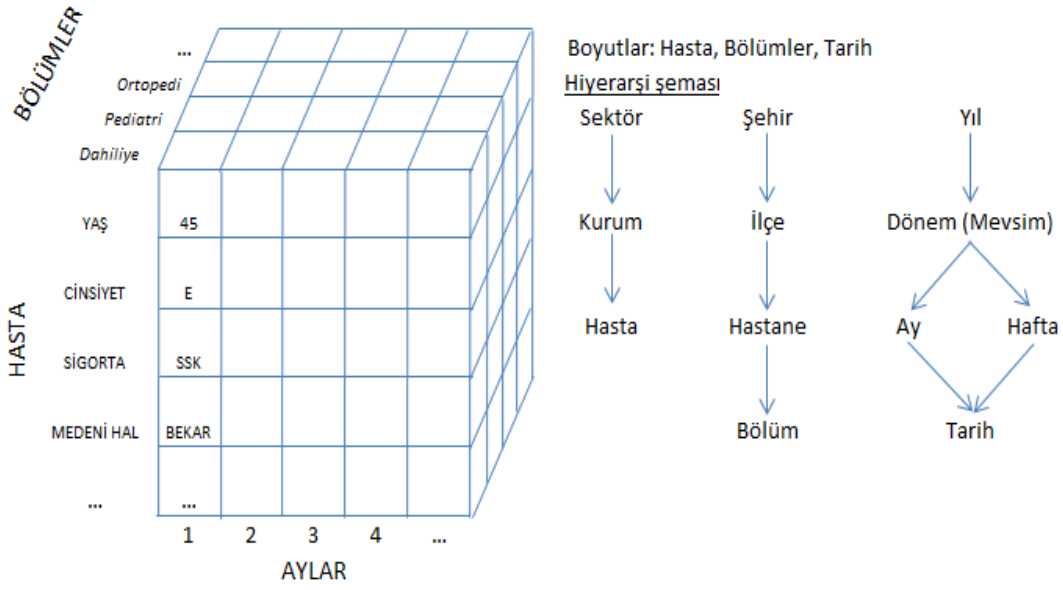
(Kaynak: Imhoff, C., Galemmo, N. ve Geiger, J. G. (2003). *Mastering Data Warehouse Design Relation and Dimensional Techniques.* Indianapolis, USA, Wiley Publishing.)

Veri ambarının daha iyi anlaşılabilmesi için OLTP, OLAP, alt küme (data mart), üst veri (meta data) kavramlarına değinilmesi gerekmektedir.

OLTP ⟶ Bir kuruluştaki güncel verilerin işlendiği veri tabanına, OLTP (OnLineTransactionProcessing) veri tabanı denir (Özkan, 2008, s.21). OLTP sistemi veri tabanındaki kayıtlar güncellenebilir, okunabilir, silinebilir ve yeni kayıt eklenebilir, ancak veri ambarında oluşturulan veriler sadece okunabilir niteliktedir. Bu kavramı sağlık sektöründe göstermek istenirse; hastanelerin kullandığı otomasyon yazılımlarının veri tabanları OLTP sistemlerine uygun bir örnek olacaktır. Gelen hasta hakkında kayıt ekleme yapılabilir, kişi evlendiyse ya da adres bilgisi değişmişse güncelleme yapılabilir, kişi hakkındaki yanlış bir kayıt silinebilir ve hasta hakkındaki bilgiler doktor tarafından veri tabanından okunabilir.

OLAP ⟶ OLAP (OnLineAnalyticalProcessing) veri tabanlarından çok boyutlu sorgulama yapılmasını kolaylaştıran bir yöntemdir, bir veri tabanı değildir

(Thomsen, 2002, s.3). OLAP işlemi, toplama (rollup) ve inme (drill-down); tek veya çok boyutlu hiyerarşik yapı ve çok boyutlu gösterim (pivot) içermektedir (Şekil 6). Verilerin bazı özelliklerini birleştirerek ya da göz ardı ederek tablo sayısını azaltabilir böylelikle "rollup" işlemi yapmış olunur, ancak verilerin özelliklerine göre yeni tablolar ya da yeni ayrıntılar oluşturursak "drill-down" işlemi gerçekleştirilmiş olunur.Ayrıca zaman boyutu da bu yöntemin önemli bir elemanıdır. Şekil 6'da veri tabanının boyutları ve boyutlardaki hiyerarşik adımları görmek mümkündür.



Şekil 6. Çok Boyutlu Veri Gösterimi

DATA MART ⟶Veri ambarlarının alt kümeleridir (alt veri tabanlarıdır). Şekil 6'ya bakacak olunursa, bir bölümün kendi hakkında yapacağı bir çalışma için tüm veri tabanını içeren sorgulama yapmasına gerek yoktur. Kendi bölümünün verisini içeren veri tabanıyla yapılan bir çalışmayla, daha basit sorgular ve daha hızlı sonuçlar çalıştırılabilir. Ancak şu durum da ortadadır ki, alt kümeler veri ambarındaki kadar ayrıntılı veri içermezler.

META DATA ⟶ Meta Data (üst veri), veri hakkındaki bilgilerdir. Şöyle ki, veri tabanı hazırlanırken belirlenen alanlar, veri tipleri, alanın boyu gibi verilerle ve tablolarla hatta veri tabanıyla ilgili bilgilerin bulunduğu alanlardır. Örneğin bir hastanın

ağırlığının kilogram ya da gram olarak yazılması, boyunun santimetre yada metre olarak yazılması; kaydedilecek bilginin sayı yada metinsel ifade olması; veri tabanın satır sayısı, sütun sayısı ya da versiyonu gibi bilgiler üst verinin kapsamındadır.

Kurumlar veya şirketler için, veri ambarı olusturulurken dikkate alınması gereken hususlar aşağıda sıralanmıştır (Imhoff vd., 2003, s.20):

1) İş zekasının kapsadığı bir sorunun ortaya konması,
2) Gereksinimlerin belirlenmesi,
3) Çözüme ulaşmak adına, nihai kullanıcılar için uygun olan teknolojilerin belirlenmesi,
4) Bir prototip oluşturmak, oluşturulan veri ambarının kurumsal kullanıcılarla işlevselliğini sınamak ve yeniden tasarlanması,
5) İhtiyaçlara ve iş modellerine göre veri ambarındaki veri modellerinin oluşturulması,
6) Veri ambarından, veri modellerinden ve geri operasyonel sistemlerden "data mart" haritasını belirlenmesi,
7) Veri hazırlama süreçleri ve veri toplama-taşıma-dönüştürme kodları oluşturulmalıdır. Hata bulma ve düzeltme ile denetim prodeürleri unutulmamalıdır.
8) Veri ambarı ve data mart (alt küme) oluşturma süreçleri test edilmelidir. Veri parametrelerinin kalitesi ve işletme için uygun üst veriler (meta data) kontrol edilmelidir.
9) Üstteki maddelerin gerçekleştirilmesiyle, veri ambarının ve alt kümelerin sürece dahil edilmesi,

gerekmektedir.

Veri ambarındaki verilerle analiz yapılması, zaman boytu da eklenerek verilerin isteğe uygun olarak analitik işlemlere tabi tutlması demektir. Ancak veri ambarının yapısı gereği diğer veri tabanlarından ayrılan bazı özellikleri mevcuttur. Bu özellikler ve açıklamaları aşağıda verilmiştir (Imhoff vd, 2003, s.21).

Konuya Yöneliktir ⟶ Veri ambarı oluştururken dikkat edilmesi gereken hususlarda bahsedildiği gibi, veri ambarı organizasyonun asıl amacına hizmeteder nitelikte olmalıdır. Bir hastanedeki temel birim hastadır. Dolayısıyla veri ambarının tasarımı da hastalar baz alınarak hazırlanmalıdır. Tasarım; doktor, idari personel, ilaç-malzeme stoğu gibi alanlara yönelik olarak hazırlandığı varsayılırsa sonuçta hastane içi işleyiş düzenlense bile eksik olan nokta hastaların memnuniyet dereceleri hakkında veri toplanmaması olacaktır.

Bütünleşiktir ⟶ Veri ambarının diğer bir özelliği de bütünleşik olmasıdır. Veriler kaydedilirken aynı kodlamaya ya da ölçü birimine sahip olmalıdırlar. Diyelim ki hasta hakkında boy, kilo verileri alınsın; ancak bazı kayıtlarda boy santimetre, bazılarında metre cinsinden kaydedilmiş olsun. Bu durumda işlemsel veri tabanından veri ambarına aktarım yapılırken ölçüm birimlerinin ortak bir birime dönüştürülmesi gerekecektir. Bu sayede veri ambasrındaki verilerde tutarlılık sağlanmış olur.

Zaman Boyutu ⟶ Veri ambarındaki kayıtlar belirli bir döneme aittir. Dolayısıyla işlemsel veri tabanlarından farklı olarak anlık kayıtlar değil; ay, hafta, yıl gibi dönemsel veriler tutulmaktadır. Buradan yola çıkarak veri ambarındaki kayıtlarda değişiklik yapılamaz sonucuna varılmaktadır.

Sadece Okunabilir ⟶ Zaman boyutu özelliğindeki "veri ambarında değişiklik yapılamaz" özelliğinden faydalanarak; veri ambarındaki kayıtlar silinemez ve güncelleştirilemez, sınucuna varılabilir. Verile sadece analiz yapacak kişiler için okunabilir düzeydedir. İşlemsel veri tabanlarındaki kayıt ekleme, silme, güncelleştirme gibi işlemler yapıyı düzenleştirdiği için kullanıcıların işlem hızlarını düşürmektedir. Dolayısıyla veri ambarının yapısı gereği, yapı bozulmaz ve işlem hızı düşmez.

3. VERİ MADENCİLİĞİNE GENEL BAKIŞ

3.1.VERİ MADENCİLİĞİNİN GELİŞİMİ

Veri madenciliği esasen yüzyıllardır farklı şekillerde ve uygulama alanında kullanılan ancak bilgisayarın keşfi ve bu keşifle gelen teknolojilerin geliştirilmesiyle daha da belirginleşen bir kavramdır. Bu noktada veri madenciliğinin bir kavram olduğunun özellikle belirtilmesi gerekmektedir. Veri madenciliği bir analiz yöntemi ya da bir veri tabanı değildir. Çeşitli istatistiksel yaklaşımlarla veri tabanındaki kayıtlardan; anlamlı, ilk bakışta görülemeyen örüntülerin ortaya çıkartılması, bunların kullanılabilirliğini belirleyerek ve görsel olarak hazırlarak sunulması, fayda maksimizasyonunun sağlanması için kullanılan tüm işlemlerin toplamı veri madenciliğidir.

Veri madenciliğinin başlangıcı, Bayes teoremi olarak varsayılabilir. İlk defa Thomas Bayes tarafından bahsedilen bu teorem 1812 yılında Pierre SimonLaplace'ın "Olasılıkların Analitik Teorisi (Théorie Analytique Desprobabilités)" kitabıyla yayımlanmıştır. Kısaca Bayes teoremi bir olayın gerçekleşip gerçekleşmemesinin birden fazla tekrarlamasıyla oluşan olasılık frekansı yöntemidir. Tekrarlarla oluşan örüntü bir sürecin tamamlanması, bir malın alımı gibi olayların tahmin edilmesinde yardımcı olabilmektedir. Bayes denklemi:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

şeklinde hesaplanmaktadır.

$P(A|B)$: B olayının bilinmesi koşulunda, Aolayının olması olasılığıdır. Bayes Regresyon analizi, örüntü bulmadaki yöntemlerden bir tanesidir. Bu analiz ilk olarak "en küçük kareler" yöntemi olarak 1805 yılında Adrien Marie Legendre tarafında ortaya atılmıştır. Ancak "regresyon" terimi ile ilk defa Francis Galton'un aile bireyleri arasındaki kalıtım araştırmasında karşılaşılmıştır (Mogull,2004, s.59). İlerleyen yıllarda bilim insanlarının bu teoremi geliştirmesiyle, sadece kalıtım değil daha geniş genel

istatistiksel problemler için bu yöntem kuıllanılabilir olmuştur (Aldrich,2005). Ancak günümüzde regresyon kavramı, verileri arasında doğrusal model bulmak ya da veriler için bir belirli modellere (üssel, logaritmik, parabolik, kübik, tersinir, büyüme, lojistik, v.b.) ait uygun eğrileri belirlemek şeklindedir. Doğrusal regresyon modeli:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

şeklinde formülize edilmektedir. Denklemdeki parametrelerin tahmini ise şu şekilde hesaplanmaktadır:

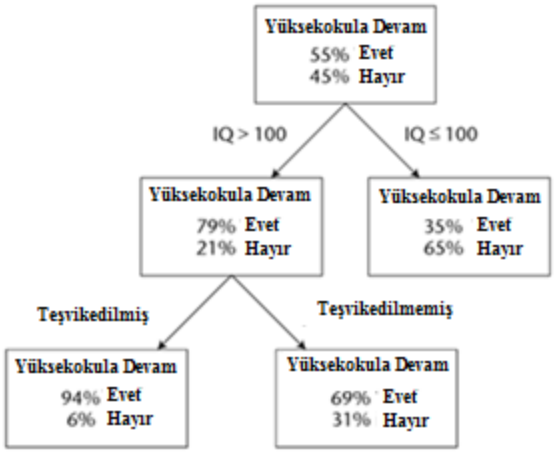
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Bilgisayarın keşfi ve gelişmesiyle daha geniş veri setlerinin analizi de uygulanabilir hale gelmiştir. Ayrıca günümüzde piyasa şartları ve küresel ekonominin de etkisiyle sadece bölgesel olarak değil aynı zamanda ülkeler arası ticaret gelişmiş ve daha fazla veriyle karşılaşılmıştır. Verilerin karmaşık yapısı gereği analiz yöntemleri de bu bağlamda yetersiz kalmaktadır. Bayes teoreminin ikili olasılıksal yönü ve regresyon analizinin olmazsa olmaz varsayımları nedeniyle çoklu parametre analizleri yapılamamaktadır. 1950'lerde sinir ağları, kümeleme analizi, genetik algoritmalar geliştirilmesiyle temel varsayımsal sıkıntılar aşılabilmiştir. Bu analizler bilgisayar sistemlerine bağımlıdır. Şöyleki; binlerce satırlık veri tabanlarındaki verilerin elle işlenmesi ve hatasız olarak hesaplanması kaynak kaybına sebep olmaktadır. Her üç yöntem de kendi içinde algoritmalar barındırmaktadır. Temelde, genetik algoritma ve sinir ağları olasılıksal olarak örüntüleri bulurken; kümeleme analizi, veri setindeki elemanların birbirlerine uzaklığına göre örüntüleri ortaya çıkartmaktadır.

Günümüze en yakın geliştirilen iki yöntem olarak: karar ağaçları ve destek vektör makineleri yöntemleri ile karmaşık örüntüler ve karar verme işlemleri çözümlenmiş olmaktadır (Kandartzic, 2011, s.171). Karar ağaçları düğümlerden ve

dallardan oluşan anlaşılması kolay bir sistemdir. İlgili düğüme karşılık gelen değerlerin olasılıksal hesaplamalarıyla nihai sonuç, karar vericiler tarafından kullanılır.Bu noktada tekrardan hatırlatılmalıdır ki veri madenciliğinde ortaya çıkan örüntülerin yada sonuçların doğru olup olmaması söz konusu değildir. Çıktılar hangi işlem için kullanılacaksa yada hangi amaca hizmet ediyorsa, karar verici olarak bulunan kişiler tarafından değerlendirilmesi gerekmektedir.



Şekil 7. Karar Şeması Örneği

(Kaynak: http://www.yazgelistir.com/Makaleler/Resimler/1000001858_DataMining1.png, 05.06.2013)

Şekil 7'de öğrencilerin matematiksel zeka (IQ) ve teşvik edilme bilgileri değerlendirilmiş ve olasılıksal olarak bir öğrencinin yüksek okul eğitimine devam etme durumu analiz edilmiştir. Sisteme göre, IQ seviyesi 100 puanın üzerindeolan ve teşvik edilmiş bir öğrencinin yükseköğrenime devam etme olasılığı %74,26 olarak hesaplanmaktadır.

3 .2 VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

Veri madenciliği yapısı gereği raporlamanın ve veri tabanının olduğu her türlü süreçte kullanılabilinmektedir. Sektörel ve işlem odaklın düşünülürse veri madenciliği şu alanlarda kullanılabilir (http://en.wikipedia.org/wiki/Data_mining, 06.06.2013):

- Perakende satış
- Finansal hizmetler
- Risk ve kredi analizi
- Pazarlama
- Müşteri davranış analizi
- Fiyatlandırma
- Telekomünikasyon
- Tedarik zinciri analizi
- Ulaştırma
- Tarım (şarap üretimi, pestisit kullanımının optimumlaştırılması)
- Hava durumu tahmini
- Öğretimi desteklemek için geri bildirim
- Öğrenci performans tahminlemesi
- Öğrenci davranışlarının tespiti
- Sosyal ağ analizi
- Kavram haritalarının oluşturulması
- Eğitim planalaması
- Bilgi güvenliği
- Silahlı kuvvetler güvenlik uyarıları
- Şehircilik ve planlama
- Coğrafi bilgi sistemleri
- Genom açıklama
- Moleküler etkileşim

- Doğal ürünlerin kimyasal çeşitliliği
- İlaç araştırmaları
- Klinik karar destek sistemleri
- Sahtekarlık algılama
- Müşteri tutundurma
- Çapraz satış kuralları
- Web içeri araştırma

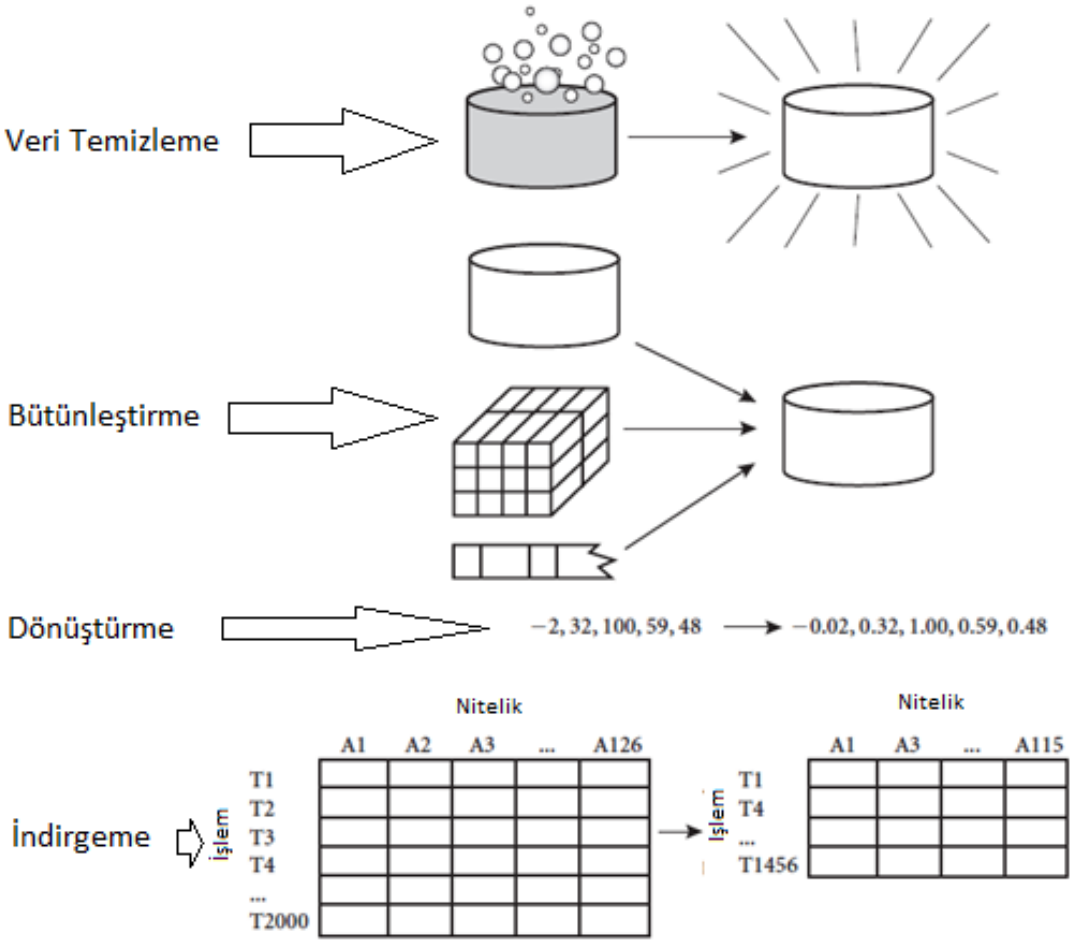
Belirtilen kullanım alanlarında; birçok veri yığın halinde bulunmaktadır, dolayısıyla klasik hesaplama yöntemleriyle analiz yapılması oldukça zordur.Buradaki sorun karar verici kişilerin uzmanlaşmasıyla bu alanlarda bir sonuca ulaşmasıdır. Ancak zaman ve maliyet kaybının oluşması sebebiyle kararların daha basit ve anlaşılır, aynı zamanda daha kesinleşmiş bir düzeye çekilmesi gerekmektedir.

3.3 VERİNİN BİLGİYE DÖNÜŞTÜRÜLMESİ

Günümüz bilgisayar teknolojilerinin gelişmesiyle veri madenciliği uygulamasının işletmelerde kullanımı giderek yaygınlaşmaktadır. Ancak elde edilen verilerin doğruluğu ve bu ham verilerden elde edilen bilgilerin kesinliği tartışılır bir boyuttadır. Dolayısıyla ham veriler belirli işlemlere tabi tutularak analize uygun hale getirilmesi gerekmektedir.

Hastane veri tabanlarından yola çıkılacak olunursa; bir hastanın laboratuvar sonuclarında tarih girilmemiş olabilir yada hastanın yaşı girilmemeiş olabilir. Böyle kayıtlara “missingvalue” denir. Başka örnek olarak hastanın kimlik numarası 11 haneden fazla olabilir, doğum tarihiyle ilgili olarak; 12 aydan fazla rakam yazılabilir yada doğum yılı 1946 olan hasta için kayıt 1646 olarak kaydedilebilir. Bu gibi durumlarda verinin veri setindeki aykırı değerlere “gürültülü veri” denir (Han vd., 2012,s.89).

Veri setinin hatalardan kurtarılması için verilerin ön işleme tabi tutuluarak analize uygun hale getirilmesi gerekmektedir. Veriden bilgiye ulaşırken kullanılan veri işleme prosedürleri Şekil 8'de gösterilmiştir.



Şekil 8. Ham Verinin Dönüşümü

(Kaynak: Chakrabarti, S.,Cox. E., Frank, Eibe., Güting, R. H., Han J., Kamber, M., Lighstone, S. S., Nadeau, T. P., Neapolitan R. E., Pyle, D., Refaat, M., Schneider, M., Teorey, T. J. ve Witten I. H. (2009). Data Mining Know It All. Massachusetts, USA, Elsevier.)

3.3.1 Temizleme

Verilerdeki eksiklikleri giderebilmek için verisetindeki eksik verileri ya da kayıtları, uç verileri, gürültülü verileri belirlemek gerekmektedir (Chakrabarti vd., 2009, s.72) . Mevcut eksikliklerin giderilememesi durumunda ilgili verinin bulunduğu kayıt veri setinden kaldırılabilir(Young ve Johnson,2010).

3.3.1.1. *Kayıp Veri*

Veri setindeki kayıp veri sık karşılaşılabilinir bir durumdur. Özellikle veri sirkülasyonunun fazla olduğu veri tabanlarında kişilerin (hastalar, müşteriler) tüm bilgileri o an kayıt altında alınamayabilir. Bu durumda oluşan eksikliği aşmak için bazı yöntemler geliştirilmiştir (Han vd., 2012, s.91).

1. **Kaydın Çıkartılması**: Eğer kişi hakkındaki kayıtlarda eksiklik varsa o kişinin kaydı veri tabanından çıkartılabilir. Çıkartılacak kayıt sayısı toplam kayıtlardaki bilgileri değiştirmeyecek şekilde olmalıdır. Şöyleki hasta kaydında kimlik numarası bilgisi boş olan hastayı veri tabanından çıkartırsak hastaların ortalama yaş bilgisi, hizmet sayısı v.b. bilgilerin de bozulması analmına gelir.Sonuç olarak analizden elde edilecek sonuçlar gerçeği yansıtmaz, yanlı bir tahminde bulunulur ve karar vericilerin yanlış yönlenmesine sebep olur.
2. **Kayıp verinin elle doldurulması:** Veri setindeki kayıpların elle doldurulması da mümkündür. Ancak bu uygulama objektiftir ve hataya oldukça açıktır. Kayıp verinin çok olduğu büyük veri tabanlarına uygulanamaz.
3. **Genel bir sabit değer atanaması:** Veri tabanında boş olan hücreler için genel bir terim atanabilir (Enders, 2010, s.22). Örneğin kimlik numarası olmayan bir hasta için ilgili alana “boş”, “0”,”-“,”yok” v.b. ayırt edici karakter ve kelimeler yazılarak, bu kişiler analize dahil edilebilir. Bu yöntem yanlıdır. Şöyleki; boş olan hücrelere bu yöntem uygulanırsa, ilgili parametre için yeni bir değişken atanmış olunur. Böylece hastanedeki bölümleri düşünecek olursak (dahiliye, K.B.B., göğüs Hastalıkları v.b.), sisteme yeni bir bölüm eklemiş olunur.

4. **Parametre ortalamasının atanması:** Veri girişi yapılmamış hücreler için bulundukları parametreye ait olan ortalama değer atanabilir. Örneğin hastaneye gelen hastalrın ortalama yaşı "54,3" olsun. Hastanın hangi bölüme geldiği yada hangi işlemi yaptırdığı önemli olmaksızın yaş bilgisi bulunmuyorsa bu hücreye "54,3" yazılır. Böylelikle kayıp verinin genel ortalamayı etkilemesi engellenir.
5. **İlgili Sınıfın ortalamasının atanması:** Eksik verilere ortalama atama işlemi sadece o parametrenin genel ortalamasına göre değil, sınıflandırarak da yapıla bilir. Dördüncü maddedeki örnek geliştirirlirse; genel ortalama 54,3 ancak dahiliye polikliniğine gelen hastalar için, yaş ortalaması 62,8 varsayılsın. Sonuç olarak dahiliyede muayene olan ve yaş bilgisi bulunmayan hastaların yaşları 62,8 olarak atanacaktır.
6. **Gerçekleşme olasılığı en yüksek olan değerin atanması:** veri setindeki verilerin frekans tablosu hazırlanarak olasılıksal olarak ilgili parametrede bulunan enyüksek olasılıklı değer boş hücrelere atanır(Ader ve Mellenbergh, 2008, s.316). Diyelimki bir hastanın hangi polikliniğe geldiği bilinmiyor, bu durumda tüm hastalar içinden polikliniğe başvuru oranları bulunarak olasılıksal olarak en yüksek değere sahip olan bölüm boş hücrelere atanır.

Yukarıda bahsedilen yöntemler çok kullanılan popüler yöntemlerdir. Bunlardan başka veri tipine ve özelliğine göre regresyon analizi, karar ağaçları, zaman serisi analizi, Bayesyen sınıflandırma yöntemleri kullanarak da veri tabanındaki eksik veri problemi giderilebilir (Silahtaroğlu, 2008, s.21). Bahsi geçen altı yöntemden 3. ve 6. yöntemler yanlılık arz etmektedir(Chakrabarti vd., 2009,s.73). Dikkat edilmesi gereken bir diğer nokta da; veriler bazı durumlarda isteyerek de boş bırakılabilir. 18 yaş altı bir hastanın kaydında sigorta kısmının boş bırakılması buna örnek gösterilebilir.

Aşağıdaki Tablo 1'de verilen örnek veri tabanında üçüncü sıradaki hastanın yaş ve bölüm bilgilerinin eksik olduğu görülmektedir. Yaş verisi doldurulmak istendiğinde, genel bir sabit değer, parametre ortalaması ya da ilgili sınıfın ortalaması

atanabilmektedir. Ayrıca bölüm bilgisi içinde, iligili sınıfın modu ya da gerçekleşme değeri en fazla olan değer atanabilmektedir.

Tablo 1. Örnek Veri Tabanı

ID	Hasta Adı	Yaş	Cinsiyet	Bölüm	…
1	Osman Katı	35	E	Dahiliye	
2	FatihSandıkkaya	30	E	GenelCerrahi	
3	SelimÇam		E		
4	BirselAkay	20	K	K.B.B	
5	ÜmitÇavdar	25	E	Dahiliye	
…					

3.3.1.2 *Gürültülü Veri*

Gürülütülü veri, veri setindeki uç değerler, kişisel ölçüm yanlışlarından doğan veya hatalı veri girişi yapıldığında karşılaşılan veriler olarak tanımlanabilir (Chakrabarti vd., 2009, s.75). Veri setinin düzgün analiz edilebilmesi için gürültülü verilerin "düzgünleştirilmesi" (smoothing) gerekmektedir (Van den Broeck vd., 2005).Verilerin dügünlerştirilmesinde en çok kullanılan ve kolay sonuca ulaşılan yöntemler kutulama (binning), doğrusal regresyon (linearregression) ve kümeleme (clustering) olarak görülmektedir (Enders, 2010, s.86).

Kutulama yöntemi, veri setindeki gürültülü veriyi de barındıracak şekilde bir grup veri alınarak uygulanır. Bu yöntemde dikkat edilmesi gereken nokta alınacak gruptaki eleman sayısını çalışmacı, konunun içeriğine göre kendisi belirler (Silahtaroğlu, 2008, s.21). Kutulama metodu, seçilen grubun; ortalamasına, medyanına ya da sınır değerlerine göre uygulanır. Şöyleki:

A:{1, 5, 9, 11, 17, 18, 27, 29, 33, 35, 36, 100} elemanlarını içeren bir veri seti olsun. Bu veri seti düzgünleştirmek istenirse;kutulama metodundaki, ortalama, medyan ve sınır değerlerine göre şöyle sonuçlarla karşılaşılacaktır;

1. **Ortalamaya Göre:**

Öncelikle veri setini eşit hacimlerde alt kümelere ayrılsın.

A_1:{1, 5, 9, 11}

A_2:{17, 18, 27, 29}

A_3:{33, 35, 36, 100}

Alt kümeler belirlendikten sonra kümelerin kendi ortalamaları hesaplanarak her bir elemanın yerine kümenin ortalaması yazılır. Aritmetik ortalama:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

şeklinde formüle edilmektedir.

Oluşan yeni veri seti şu şekilde olacaktır:

A:{6,5; 6,5; 6,5; 6,5; 22,75; 22,75; 22,75; 22,75; 51; 51; 51; 51}

2. **Sınır Değerlerine Göre:**

Tekrardan veri seti eş hacimdeki alt kümelere ayrılsın. Hesaplama medyana göre yapılsın:

$$m = \frac{a - b}{N}$$

a= alt kümedeki üst sınır değeri

b= alt kümedeki alt sınır değeri

N= alt kümedeki eleman sayısı

A kümesinin alt kümelerine yukarı daki işlemi uyguladığımızda oluşan yeni küme şu şekilde olacaktır:

$$m_{A_1} = \frac{11-1}{4} \Rightarrow m_{A_1} = 2,5 \text{ ve } m_{A_2} = 3,\ m_{A_3} = 16,75$$

A:{2,5; 2,5; 2,5; 2,5; 3; 3; 3; 3; 16,75; 16,75; 16,75; 16,75}

Sınır değerlerine göre yapılan bir diğer düzgünleştirme işlemi de; alt ve üst sınır değerlerine en yakın elemanlara alt ya da üst sınır değerlerinin atanmasıdır. Şöyleki;

A_1:{1, 5, 9, 11} $\Rightarrow$ A_1:{1, 1, 11, 11}

A_2:{17, 18, 27, 29} $\Rightarrow$ A_2:{17, 17, 29, 29}

A_3:{33, 35, 36, 100} $\Rightarrow$ A_3:{33, 33, 33, 100}

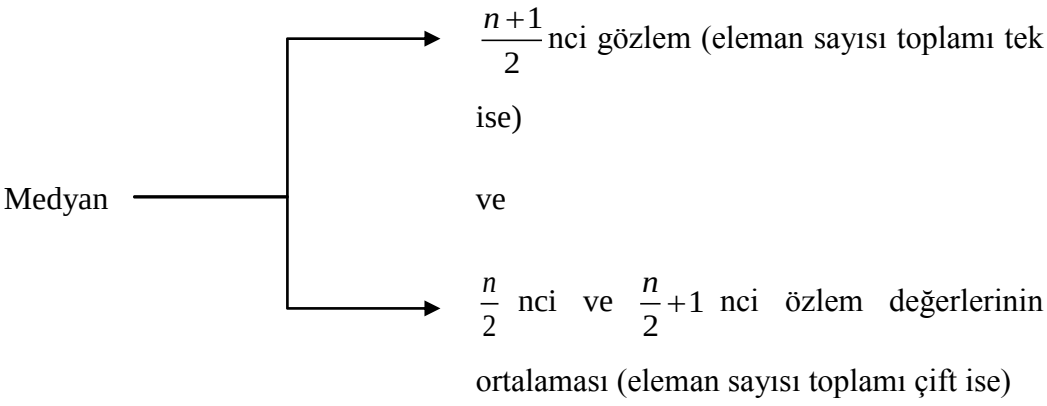
Oluşturulan yeni alt kümeler yukarıdaki gibi olacaktır.

3. Medyana Göre:

Veri düzgünleştirme işlemi alt kümeler oluşturulduktan sonra bu kümelere ait medyan değeri atanarakda yapılabilir. Ancak verilerin dağılımı medyan değerinin hesaplanmasını değiştireceğinden, alt kümelerin dağılımlarının da bilinmesi gerekmektedir.

- Normal dağılımda medyan ve ortalama değerleri eşittir
- Uniform dağılımda medyan ve ortalama eşittir ve basitçe şu şekilde hesaplanır: (a + b) / 2.
- Parametre değeri λ olan bir üssel dağılımda medyan değeri $\lambda^{-1}\ln 2$ olacaktır.
- $X \sim Couchy\ (x_0, \lambda)$için medyan x_0 olacaktır.
- $X \sim Weibull(k, \lambda)$için medyan $\lambda \ln(2)^{1/k}$ olacaktır.

Basit seriler için medyan bulunmak istenirse:

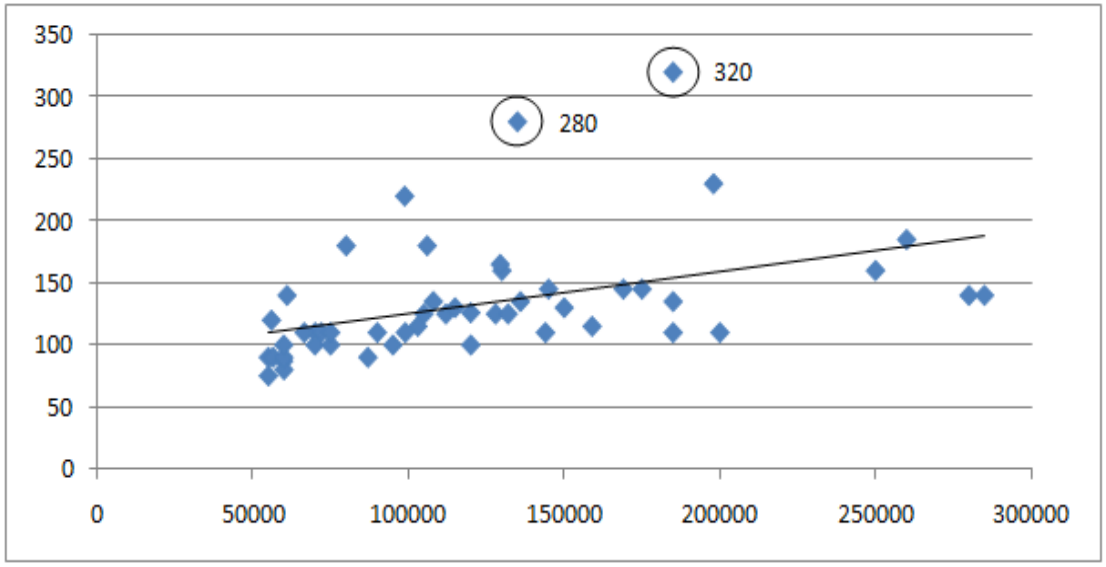


İşlemleri medyanın bulunmasında kullanılabilir.

Buradan yola çıkarak A kümesinde düzgünleştirme işleminin sonucu olarak oluşan yeni A kümesi şöyle olacaktır.

A:{7; 7; 7; 7; 22,5; 22,5; 22,5; 22,5; 35,5; 35,5; 35,5; 35,5}

Veri düzeltme işlemi doğrusal regresyon modeli kurarak da yapılabilmektedir(Enders, 2010, s.92). Veri setindeki uç değerler nokta diyagramı yardımıyla görsel olarak belirlenebilir ve regresyon modeli yardımıyla uygun olan değer atanabilir. Gürültülü verinin yerine yenisini atayabilmek için hazırlanan iki değişkenli doğrusal regresyon modeline birden fazla bağımsız değişken koyulabilir. Böylelikle iki değişkenli bir modelden ziyade çok değişken ile analize için daha uygun bir düzletme yapılmış olunur (Han vd., 2012,s.106).



Grafik 1. Uç Değer Örneği

Grafik 1.'de görüldüğü gibi regresyon eğrisinden uzakta bulunan veriler seriyi bozmaktadır. Bu değerleri modelden çıkartarak tekrardan regresyon eğrisi oluşturmak ve ilgili veriyi modele göre tahminlemek gerekmektedir.

Bir diğer metod olarak kümeleme yöntemi uç değerlerin belirlenmesinde kullanılabilir (Chakrabarti vd., 2009, s.76). Kümeleme yönteminde eldeki veriler analiz sonucu kümelere ayrılarak, küme dışında kalmış veriler belirlenir. Uç değer kabul edilen bu verilere yakın oldukları kümenin elemanın ya da küme ortalamasının değeri atanır (Silahtaroğlu, 2008, s.23).

Veri düzenleme işleminin ilk adımı uyumsuz verilerin tespit edilmesi için veri temizleme işlemi uygulanmaktadır (Aderve Mellenbergh, 2008, s.336). Veri setindeki uyumsuzluk kişilerin yanlış veri girmesi, kasıtlı hatalar, eski veriler, kişilerin anlamakta zorlanacağı formların hazırlanması, ölçüm araçlarının ölçümleme eksikliği, veri tabanının veriler için uygun hazırlanmaması sebepleriyle açıklanabilir. Ayrıca verilerin amaçları dışında kullanılması ve veri tabanları birleştirilirken birim farklılıklarının olması da veri setindeki tutarsızlığı açıklayabilmektedir (Han vd.,2012, s.93).

3.3.2. Bütünleştirme

Veri madenciliği analizleri sırasında analizin gerçekleşitirilebilmesi için veri tabanlarını birleştirmek gerekebilir. Farklı veri tabanlarından alınan verilerin birim farklılığı birleştirme sırasında analizi yapan kişilere zorluk çıkarmaktadır (Han vd., 2012, s.93). Örneğin hastalar ile ilgili poliklinik ve laboratuvar olarak iki veri tabanı olduğu varsayılsın. Poliklinik veri tabanında hastaların cinsiyet bilgisi "E", "K" ve laboratuvar veri tabanında "Bay", "Bayan" olarak kaydedilmiştir. Bu iki veri tabanını birleştirirken karşımıza dört farklı cinsiyet çıkmaktadır. Bu gibi birim farklılıklarından kaynaklanan hatalar olduğu gibi meta data (veri tabanındaki verilerin özellikleri) hatalarıyla da karşılaşılmaktadır. Örneğin bir veri tabanında kilo bilgisinin girildiği hücre "integer", diğer bir veri tabanında aynı bilgi "float" olacak şekilde ayarlansın. Veri tabanları birleştirildiğinde sayısal olarak bir hata gözükmeyecektir. Ancak analizde araştırmacı, verilerinondalıklıkısmlarına göre daha hassasbir sonuca ulaşmak isterse, veri kaybıyla karşılaşılacağı kaçınılmazdır.

Veri tabanlarının birleştirilmesinde tekrarlama (redundancy) sorunu görülmektedir. Buradaki problem farklı veri tabanlarının aynı özelliğe ve bilgiye sahip kayıtlarınyada türetilmiş verilerin bütünleştirme sonrasıda birden fazla şekilde tekrarlamasıdır (Doorn ve Rivero, 2002, s.209). Tekrarlama olup olmadığını anlayabilmek için parametrik verilerde korelasyon analizi, kategorik verilerde Ki-Kare Uyum analizi uygulanır (Chakrabarti vd.,2009, s78.). Pearson korelasyon katsayısı:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\overline{A}\overline{B}}{N\sigma_A\sigma_B}$$

formülü ile hesaplanmaktadır.

$r =$ Korelasyon katsayısı

$A =$ A parametresi değerleri kümesi

$B =$ B parametresi değerleri kümesi

$a_i =$ A kümesindeki i'nci değer

$b_i =$ B kümesindeki i'nci değer

$N =$ Toplam eleman sayısı

$\sigma_A =$ A kümesinin standart dağılımı

$\sigma_B =$ B kümesinin standart dağılımı

Korelasyona örnek verilmek istenirse; hastanenin doluluk oranı, yatan hastalar ve yatak sayısıyla hesaplanır ve birlerştirilmek istenen iki veri tabanında doluluk oranı, yatak sayısı ve yatan hasta sayısı olduğu varsayılsın. Veri tabanlarının birleştirilmesinden sonra doluluk oranı bilgisi türetilmiş veri olarak veri tabanının fazla hacim kazanmasına sebep olacaktır. Aynı şekilde parametrik değilde kategorik veriler varsa ne yapılacaktır?. Bu durum Ki-Kare Uyum testiyle çözülebilmektedir. Bu test şu şekilde gösterilmektedir:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$c =$ A kümesinin eleman sayısı

$r =$ B kümesinin eleman sayısı

$o =$ Gözlenen değer

$e =$ Beklenen değer

Bu analiz ile A ve B kümeleri birbirlerinin gözlenen ve beklenen değerleri olarak uyumlu ya da uyumsuz olarak elealınır.

Korelasyon işlemi sonucu, katsayı +1'e yaklaşırsa A ve B arasında pozitif ilişki, -1'e yakınsa negatif ilişki var, sonucuna ulaşılmaktadır. Ki-Kare testinde A ve B kümlerinin bir birlerinde uyumlu olup olmaması test edilir.Ki-Kare testinin sonucu olarak %95 güven düzeyinde $P_{hesap}<0,05$ şartı sağlandığında A ve B arasında istatistiksel olarak anlamlı bir farklılığın olduğu, dolayısıyla A ve B veri kümelerinin türetilmiş ya da aynı olmadığı sonucuna varılır. Böylelikle bütünleştirme sonrası A ve B arasında tekrarlama vardıryada yoktur denilebilir. Ancak bir ilişki ya da benzerlik versa bu iki grubun aynı olup olmadığı yada A'nın B'den mi yoksa B'nin A'dan mı türetildiği bilinememektedir(Han vd., 2012, s.96).

Veri tabanlarında bütünleştirmede ortaya çıkan sıkıntıları aşabilmek adına veri tabanı dizaynının veri ve veri özelliklerinin iyi düzenlenmesi gerekmektedir. Bütünleştirme işlemi sonrası tutarsızlıkların, fazlalıkların ve türetilmiş verilerin bulunması çıkarımların doğrulunu sağlamada ve analiz hızını arttırmada yardımcı olacaktır (Chakrabarti vd., 2009, s.81).

3.3.3 Dönüştürme

Birçok istatistiksel işlem analizin yapılabilmesi için veri setinin bazı varsayımlara sahip olması ister. Bu varsayımlar genellikle veri setinin normal dağılıma uygun olup olmaması yadadoğrusallık ile alakalıdır (Nisbet vd., 2009, s.54). Veri setindeki veri aralığının büyük olması, ilişkilendirilen verilerin matematiksel farkının fazla olması analizi etkilemektedir (Silahtaroğlu, 2008, s.24). Verilerin oluşturduğu desenler görselleştirilerek eğimin yapısı görülebilir ve doğrusallaştırmak adına da veri dönüştürme yoluna gidilebilir.Normalizayon işlemi sinir ağları, en yakın komşu sınıflandırması ve kümeleme yöntemlerindeki mesafe ölçümleri içeren analizlerde özellikle yararlıdır (Han vd., 2012, s.113).

3.3.3.1 *Min-Max Normalleştirme*

Veri setini 0-1aralığına sıkıştırmak için kullanılan yönteme min-max normalleştirmesi denir. Bu yöntemde basitçe veri setindeki en büyük değer ve en küçük değere göre hesaplama yapılır.

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

$X_i^* =$ Dönüşüm sonrası yeni X değeri

Örnek veri seti, X={28,5; 65; 116; 145; 250; 490} olarak belirlenmiş olarak kabul edilsin. 0-1 ve 1-5 aralığına dönüştürülmüş olan veri setleri Tablo 2'de gösterildiği gibi olacaktır.

Tablo 2. Normalleştirme İçin Örnek Veri Tabanı

X	**X*** **(0-1)**	**X*** **(1-5)**
28.500	0	1
65.000	0,079090	1,316360
116.000	0,189599	1,758397
145.000	0,252438	2,009751
250.000	0,479957	2,919827
490.000	1	5

Analizin gerekliliğine ve araştırmacının isteğine göre min-max normalleştirme işleminin aralığı genişletilebilir. Şöyleki:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} * (X_{yeni\_\max} - X_{yeni\_\min}) + X_{yeni\_\min}$$

formülü ile veriler istenilen sınır değerleri arasına dönüştürülebilir.

3.3.3.2 *Z-Score Standartlaştırması*

Bu yöntem istatistiksel varsayımların tamamının sağlandığı normal dağılımda kullanılan Z değeri hesaplamasıyla aynıdır. Veriler veri setinin ortalaması ve standart sapmasına bağlı olarak yeni z değerine dönüştürülür.

$$X^* = \frac{X - (\frac{1}{n}\sum_{i=1}^{n} X_i)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - (\frac{1}{n}\sum_{i=1}^{n} X_i))}{n-1}}}$$

İşlemiyle z score yapılabilir. Formülü daha basit bir şekilde yazmak istenirse:

$$X^* = \frac{X - \overline{X}}{\sigma_X}$$

olarak da belirtilebilir.

Örnek olarak verilen X={28,5; 65; 116; 145; 250; 490} veri setine z score dönüşümü uygulandığında Tablo 3'teki X^*={-0,91184; -0,69561; -0,39347; -0,22167; 0,400381; 1,822202} veri seti elde edilmektedir.

Tablo 3 Z-Score Normalleşmesi İçin Örnek Veri Tabanı

X	**X^* (z score)**
28.500	-0,91184
65.000	-0,69561
116.000	-0,39347
145.000	-0,22167
250.000	0,400381
490.000	1,822202

Z score standartlaştırma işleminde min-max normalleştirmede olduğu gibi sınır değerleri belirtilmemiştir.

3.3.3.3 *Ondalıklı Normalleştirme*

Bu yöntemde veri setindeki verileri ondalıklı hale getirecek şekilde normalleştirme işlemi uygulanır. Şöyleki:

$$X^* = \frac{X}{10^j}$$

Buradaki j değerine, veri setindeki en büyük değeri $Max\left(\left|X^*\right|\right) < 1$ koşulunu sağlayacak şekilde en küçük değer atanmalıdır(Han vd.,2012, s.115).

Min-Max normalleştirme ve Z Score normalleştirme işlemlerinde kullanılan X veri seti (ondalık değeri 6 olacak şekilde) ondalıklı normalleştirmeyle dönüştürüldüğünde Tablo 4'teki X^* veri setine dönüşmüş olacaktır.

Tablo 4. Ondalıklı Normalleşme İçin Örnek Veri Tabanı

X	**X^* (j=6)**
28.500	0,0285
65.000	0,065
116.000	0,116
145.000	0,145
250.000	0,25
490.000	0,49

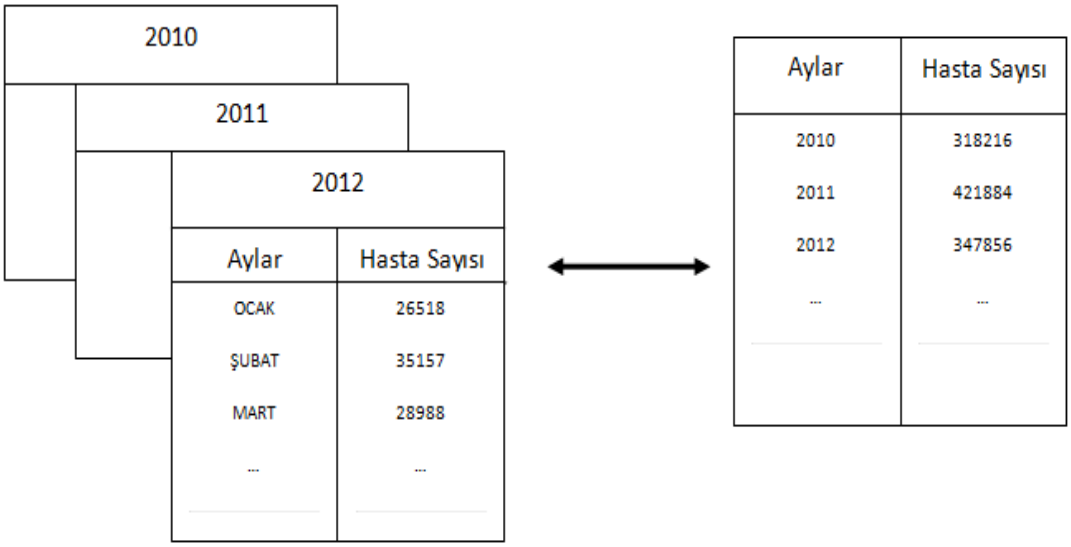
3.3.4 İndirgeme

Birçok yöntem içeren veri indirgemenin temel amacı, çok boyutlu yapıdaki veri tabanının sahip olduğu bazı parametreleri birleştirerek toplam parametre sayısını azaltmaktır (Terlemez, 2008, s.22).Saha çalışmalarında alınan değerlerde bazı problemlerle karşılaşılabilinir. Şöyleki, veriler toplanırken yöntem hatası ya da kişisel hatalar yapılabilir. Ayrıca tüm kitleye ulaşılamadığından, alınan örnekler, metod ve kişisel hatalar içerebilmesi yanında, örneklemin tüm kitleyi temsil edememesinden gibi doğal bir problemle de çalışmacıları karşı karşıya bırakmaktadır (Jackson, 2009,s.63-64). Bir diğer durum olarak da benzer özelliklerin veri tabanın da farklı farklı tanımlanması gürültüye yol açar ve işlem hızını düşürür (Karahoca, 2012, s.43).

Veri tabanını analizlerinde gürültülü verilerin ya da herhangi bir tutarsızlığın belirlenebilmesi için sınıflandırma veya gruplama uygulanabilir. Bu işlemler yapılmadan önce de veriler üzrinde indirgeme yapılması gerekmektedir (Dua ve Du, 2011, s.87). İndirgeme işlemleri veri indirgeme ya da boyut indirgeme şeklinde gerçekleştirilebilir.

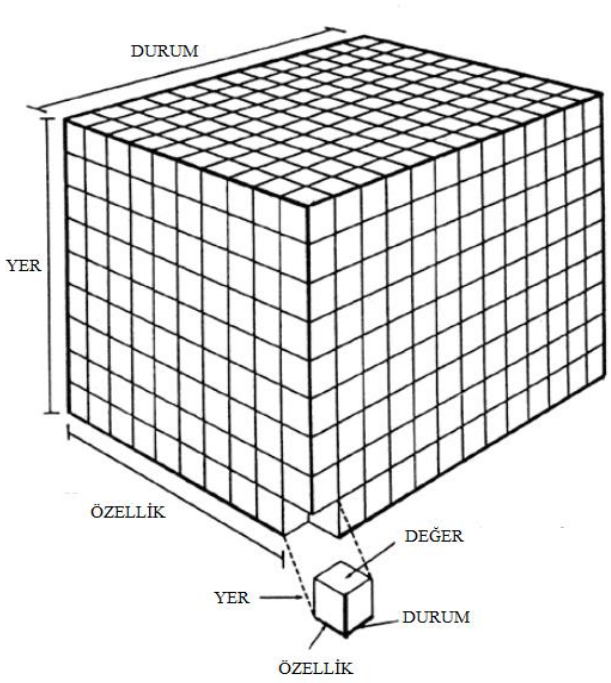
3.3.4.1 *Veri Birleştirme Veya Veri Küpü*

Veri birleştirme işlemi, aynı özelliklere sahip, farklı tablolarda bulunan parametrelerin tek bir tabloda birleştirilmesidir. Birleştirme işlemi sonrası veri sayısının ve değişkeninin azaltılması ile bazı durumlarda da veri ölçeğinin değiştirilmesi durumlarıyla karşılaşılabilir (Vatansever, 2008, s.24). Örneğin; bir hastaneye gelen aylık hasta sayısını içeren ve yıllara göre ayrılmış tablolar olduğunu varsayalım. Yıllık hasta sayısı üzerinden yapılacak işlemler analiz sırasında fazla işlem yapılması demektir. Ancak birleştirme işlemi yapıldığında, yıllık tek bir tabloya indirgenmiş tabloya ulaşılacaktır (Şekil 10).



Şekil 9. Tablo Birleştirme Örnek Gösterimi

Veri küpü, veri birleştirmeden farklı olarak çok boyutludur. Veri setindeki istenilen parameteler seçilerek gereksiz veriler veri tabanından çıkartılır. Veri küpleriyle özet bilgiye daha hızlı ulaşılır (Vatansever, 2008, s.24). Veri küpü için örnek vermek gerekirse; hastalara yapılan işlem (özellik), işlemin yapıldığı birim (yer) ve işlem sonucu (durum) bilgilerinin olduğunu varsayılsın. Bu üç parametreyi işlem numarası (değer) olarak birleştirebiliriz. Böylelikle her üç parametreyi barındıran ve işlem hızını arttırıcı veri küpünü elde edilmiş olunur (Şekil 11).



Şekil 10. Veri Küpünün Örnek Gösterimi

3.3.4.2. *Boyut İndirgeme*

Veri tabanlarında birçok parametre bulunmaktadır. Örneğin hastane veri tabanlarında hastalar hakkında ana başlık olarak demografik, işlem, tedavi, ilaç, tıbbi malzeme, laboratuvar verileri vardır. Bunların alt başlıklarının da olduğu düşünülürse onlarca değişkenin veri tabanında yer aldığı söylenebilir. Değişkenlerin bazıları yapılacak araştırma için ya da kullanılacak yöntem için uygun olmayabilir ya da birbirinden türemiş veriler veri tabanında gürültüye sebep olabilir. Bu gibi durumları aşabilmek için ileri yönlü sezgisel seçim ya da geri yönlü sezgisel seçim kullanılır.

İleri yönlü sezgisel seçimde veri tabanındaki ana değişken belirlenerek, diğer değişkenlerden en uygun olanları sezgisel olarak seçilerek bu kümeye dahil edilir. Bu durum tersi olarak geri yönlü sezgisel seçimde veri tabanındaki tüm değişkenler kümeye baştan dahil olur, ana değişkene göre uygun olmayan değişkenler kümeden çıkartılır. Bu aşamalarda katılan ya da çıkartılan değişkenler araştırmacının kararına bırakılır.

3.3.5 Modelin Belirlenmesi

Veri madenciliği çalışmalarında, ölçüm yada gözlem ile eldeedilen ham verinin yukarıda belirtilen temizleme, bütünleştirme, indirgeme, dönüştürme işlemleri yapılarak işlenebilir hale getirilmesi gerekmektedir (Özkan, 2008, s.37). Kullanıcılar eldeki veriye göre, keşfedilmemiş örüntüleri bulabilmek amacıyla, veri madenciliği teknikler bütününden en uygun yöntemi alarak veri setine uygular (Koyuncugil ve Özgülbaş,2009).

Veri madencilİği yöntemlerini genel anlamda denetimli olan ve olmayan diye ikiye ayırmak mümkündür. Denetimli veri madenciliği tekniklerinde, eğitim verisi kullanarak gizli bir fonksiyon tahminetmeyi amaçlar. Eğitim verisi, tüm veri setinden alınan ve genel veri setini iyi olarak temsil edebileceği düşünülen bir bölümdür (Dua, 2011, s.6). Denetimsiz veri madenciliği yöntemlerinde ise eğitim verisi kullanılmaz, tüm veri seti işleme tabi tutulur.

Model kurma aşaması çalışmanın temel bir noktası olduğundan, çalışmanın amacı iyi bir şekilde belirlenmelidir. Ayrıca uygun algoritmayı seçmek herzaman bulunamayabilir, bu durumda veri madenciliği yöntemlerinin denenip en uygun yönteme karar vererek çalışma sonlandırılabilir (Tekerek, 2011).

Bu çalışmada veri madenciliğindeki klasik olan yöntemlerden olan kümele yöntemiyle analiz gerçekleştirilecektir.

4. KÜMELEME ANALİZİ

Kümeleme analizi, günümüz teknolojilerinin gelişmesi ve giderek karmaşık bir yapıya sahip olan sosyal yaşamdan (bankacılık, e-devlet, alışveriş vb.)elde edilen verilerin anlaşılmasında ve analiz edilmesinde kullanılan, veri madenciliği yöntemlerinden birisidir. Kümeleme analizinde amaç, çalışılan veri tabanındaki elemanların kümelere ayrılmasıdır (Doğan, 2007, s.18). Küme elemanları nitelikleri bakımından birbirlerine benzemelidir. Denetimsiz bir öğrenim yöntemi olması sebebiyle başlangıç aşamasında kaç kümenin oluşturulacağı ve hangi değişken özelliklerine göre küme özelliklerinin belirleneceği bilinmemektedir (Ghahramani, 2004).Dolayısıyla kümeler yeniden oluşturulmuş kümelere dayalı çalışmazlar (Altıntaş, 2006, s.19).

Denetimli öğrenmenin temsilcisi olan sınıflandırma analizinde, örnek olarak; hastaneye geliş ve yaş parametrelerine göre hastaya verilecek tedavi tipi çalışmasında sınıflar daha önceden "Yatarak" ve "Ayakta" tedaviler olarak ayrılmış olabilir.. Ancak kümeleme analizinde bu sınıflar önceden bilinmemektedir (Göral, 2007, s.24). Kümeleme algoritmalarına göre alınan aynı veriler, farklı benzerliklere göre farklı sonuçlar verebilirler. Kümeleme analizi birçok uygulamada yaygın olarak kullanılmaktadır. Örneğin Pazar araştırmaları, imaj örüntülerinin tanılanması, internet aramaları, biyoloji ve güvenlik araştırmaları konuları söylenebilir (Han vd.,2012, s.327). Pazar araştırmalarında kullanılan kümeleme yöntemiyle, hangi ürünün veya ürün grubunun hangi yaş grubuna, hangi eğitim düzeyine ya da hangi varlık seviyesindekilere hitap edeceği bulunarak; ürünün piyasaya ne şekilde sunulacağı planlanabilmektedir. Benzer şekilde; internette yapılan aramalar için kullanılan kümeleme analizi aranan kelimelerin gruplanmasında ve bu kelimelerle ilişkilendirilmiş diğer kelime ya da kelime gruplarının arayıcılara getirilmesi şeklinde olduğu söylenebilir. Bu açılardan bakıldığında kümeleme analizi ileriye dönük tahminleme yaparken kullanılan temel bir basamak, bir veri tabanı tanımlama işlemi olarak görülebilir.

Daha geniş kullanım alanlarını örneklemek gerekirse (http://en.wikipedia.org/wiki/Cluster_analysis, 10.06.2013);

- Hayvan ve bitki ekolojilerinin araştırılması,
- Genler üzerinde; gen aileleri, değişimleri, evrimi gibi çalışmalarda,
- Görüntüleme tekniklerinden de faydalanılarak teşhis tedavi süreçlerinde,
- Sağkalım çalışmalarında,
- Pazar araştırmalarında,
- Ürün konumlandırma,
- Pazar bölümlendirmesinde,
- Müşteri davranışı analizlerinde,
- Web ortamındaki pazarlama işlemleri, metin analizlerinde,
- Yazılım geliştirmede,
- Suçların ve sahtekârlıkların ortaya konmasında,
- Robotik biliminde, durumsal hareketlerin analizinde,
- Matematiksel kimyada,
- Atmosfer olaylarının tahmininde,
- Jeolojik araştırmalarda

kümeleme analizi kullanılır.

Kümeleme analizine, büyük veri setlerini ayrıştırması açısından veri segmantasyonu da denebilmektedir (Han vd., 2012, s.445). Herhangi bir verinin herhangi bir kümeye ait olmaması ya da tek veriye ait bir kümenin varlığıyla da uç değer analizi yapılabilmektedir. Bu yöntem dolandırıcılık araştırmalarında yardımcı olmaktadır. Örneğin internet üzerinden yapılan çok yüksek meblağlarda bir ürün alınması dolandırıcılık olasılığına dikkat çekmektedir. Kümeleme analizinde istatistiksel yöntemler genellikle uzaklık temelli kümeleme analizi yöntemlerinde kullanılmaktadır. Buradan yola çıkarak temel istatistik yöntemlerinin kümeleme analizi için vazgeçilmez olduğu sonucu çıkartılabilir.

Her yöntemde olduğu gibi kümeleme analizinde de uygulamayı zorlaştıran, nesneler arasındaki uygun mesafenin belirlenmesi, kümeleme algoritmasının seçimi, sonuçları değerlendirilmesi gibi yönleri vardır (Bruno ve Fiori, 2011). Kümeleme analizinin kullanılabilmesi için bazı özelliklerin kümeleme algoritmalarında bulunması gerekmektedir. Çalışmanın başlangıç aşamasında, ilerlerken ve sonuçlandırırken analizin uygun sonuçlar vermesi ve etkin olabilmesi için araştırmada şu maddelere dikkat edilmesi gerekmektedir(Han vd., 2012, s.452):

Ölçeklenebilirlik: Kullanılan bir kümeleme tekniği sadece yüzlerle ifade edilebilecek bir veri tabanıyla değil milyonlarca veriye sahip bir veri tabanında da çalışabilmelidir.

Farklı ölçek türleri ile ölçülmüş veri setlerinde uygulanabilme: Günümüzde kişiler, nesneler hakkında kayda geçirilen verilere ve bu verilerin özelliklerine bakılacak olunursa, bu verilerin birçok ölçek tipine uygun olduğu görülmektedir. Örneğin hastaneye gelen bir kişinin cinsiyeti ikili veridir yani nominal ölçeğe uygundur. Aynı hastanın eğitim durumu ordinal olacaktır. Bu ve bu gibi örneklerin gösterdiği gibi kümeleme analizi sırasında kullanılan algoritma birçok farklı ölçek tipini de aynı anda çalıştırabilmelidir.

Gürültülü verilerle çalışabilme: Gerçek hayatta toplanan verilerle oluşturulan veri tabanlarında eksik, aşırı, hatalı veriler olabilir. Şöyle ki; bir hastanın cinsiyeti veri tabanında işlenirken “Erkek ”, “erkek”, “E” veya “e” olarak yazılabilir. Aynı şekilde hastanın yaşı yazılamayabilir ya da hastanın 16 olan yaşı 61 olarak yazılabilir. Kümeleme algoritmalarının gürültülü verilere olan hassasiyeti kümelerin kalitesinin azalmasına sebep olmaktadır. Bu sebeple kümeleme algoritmasının gürültülü verilerle en uygun şekilde çalışabilmesi gerekmektedir.

Yüksek boyutluluk: Veri tabanları birçok boyuttaki ve özellikteki verileri içermektedir. Genelde ise kümeleme algoritmaları iki-üç boyutlu olarak çalışmaya

uygundur. Analizin kalitesinin arttırılması, daha uygun ve kritik çözümlerin ortaya konması için kümeleme algoritmalarının daha çok boyutla çalışabilmesi gerekmektedir.

Giriş parametresi belirlemek için alan bilgisi gerekliliği: Veri analizinin yapılabilmesi için kümeleme analizi bazı önemli değişkenlerin analize konmasını gerektirmektedir. Çok boyutlu çalışmalarda bu özelliğin ne olacağını belirlemek analizi yapan kişiyi zorlamaktadır. Seçimin uygun yapılamaması uygulanan testlerin ve algoritmaların güvenirliliğini azaltmaktadır.

Artan kümeleme ve veri girişine duyarsızlık: Günümüzde kullanılan veri tabanlarının yapısı gereği anlık olarak sürekli kayıt alınmaktadır. Bu durumda çoğu kümeleme analizi uygulamaları yeni veri girişlerinden etkilenmektedir. Uygulanan kümeleme algoritması artan veri kümelerine ve yeni veri girişlerine, verikümesi yapılarında oldukça farklı yeni sonuçlar oluşturabilmektedir. Dolandırıcılık örneğini söylemek gerekirse; kredi kartıyla alınan yüksek fiyatlı bir ürünün genel ortalamayı yükseltmesindense aykırı değer olarak ele alınması gerekmektedir.

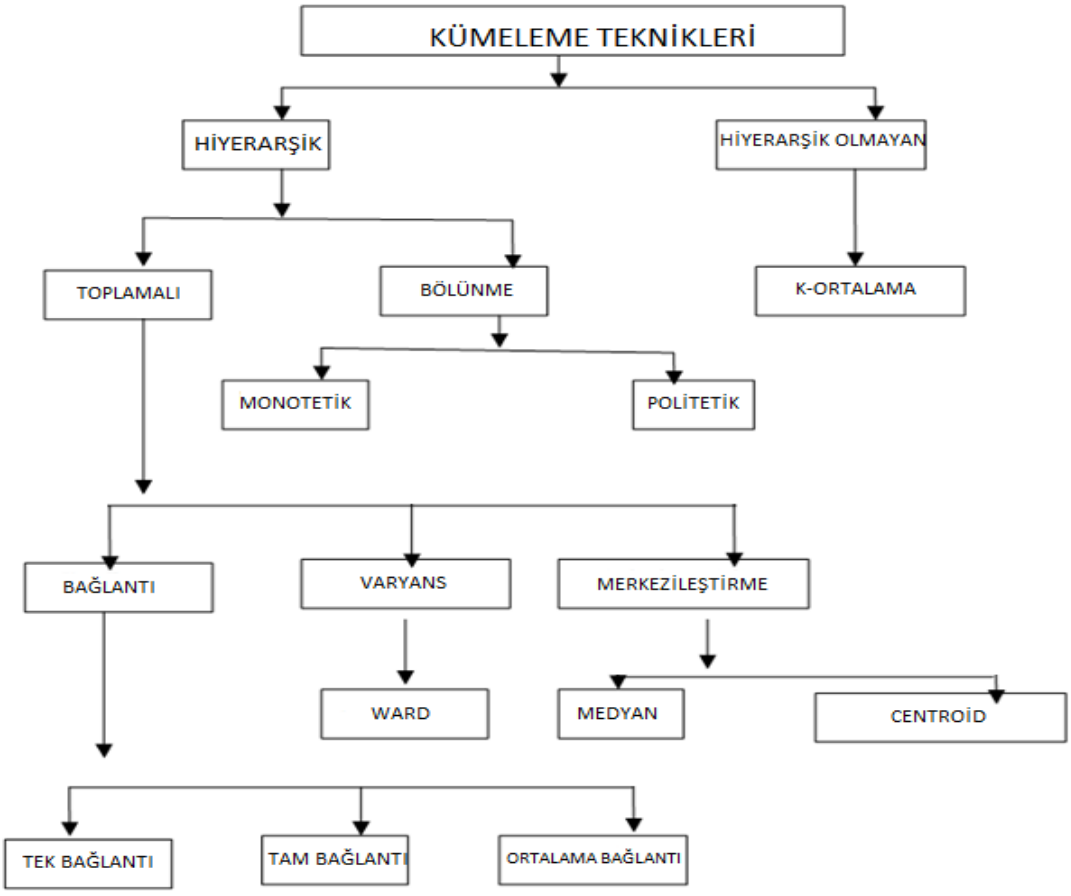
Kümelerin keyfi olarak oluşması: Kümeleme algoritmaları genellikle temel uzaklık ölçüleri kullanılarak çalışma yaptığından, oluşturulan kümeler benzer yapılarda olmaktadır. Ancak çalışmanın daha iyi sonuçlanması açısından kümeler matematiksel işlemlerin getirdiği sonuçlardan ziyade kendi doğal yapılarında olmalıdır. Bu noktada tekrardan analizi yapan kişinin bu doğal yanlışlıklara karşı bilgisi olması ve bu hataları ayıklaması gerektiği yorumuna ulaşılmaktadır.

Kısıtlama temelli kümeleme: Çalışmalar yapılırken bazı durumlarda verilere kota koymak, sınırlandırmak gerekebilir. Örneğin tüm hastanenin hasta verilerinin olduğu varsayılsın, bu durumda çalışma 12-18 yaş arasıyla sınırlandırılmak istenebilir. Benzer bir şekilde hastaneye başvuran hastaların ikametlerine de sınırlandırma uygulanabilir.

Yorumlanabilme ve kullanılabilme: Kümeleme algoritması uygulandıktan sonra ortaya çıkan sonuçların yorumlanabilme ve başka yerlerde de kullanılabilme

kabiliyetleri olmalıdır. Sonuçlardan bazı özellikli anlamlar ortaya konula bilinmelidir. Kümeleme yönteminin seçimi ve küme özelliklerinin seçimi sonuçların istenen sonuçla ne kadar örtüştüğünü ortaya koymaktadır.

Kümeleme analizi temel anlamda, hiyerarşik ve hiyerarşik olmayan, iki yöntem olarak ele alınmıştır. Veri setinin analizi yapılırken bu iki yöntemi de ayrı ayrı kullanmak, analizin başarısını arttıracaktır. Böylelikle hangi tekniğin veri setiyle daha uyumluğu olduğu görülebilecektir (Akın, 2008, s.8). Kümeleme analizindeki tekniklerin klasik anlamda ayrımı Şekil 13'de gösterilmiştir.



Şekil 11. Kümeleme Teknikleri

(Kaynak: Akın, Y. K. (2008). Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi. (yayınlanmamış Doktora Tezi). Marmara Üniversitesi/Sosyal Bilimler Enstitüsü, İstanbul.)

4.1. ÖLÇEK TİPİNE GÖRE UZAKLIKLARIN BELİRLENMESİ

Veri madenciliği doğası gereği veri tabanından alınan birçok veriyi ve bu verilerin özelliklerine göre birçok veri tipi içermektedir. Örneğin bu çalışmada kullanılacak olan parametrelerden biri hastanın tedavi tipi olarak Ayaktan Tedavi ya da Yatarak Tedavi gibi ikili bir değişken iken hastanın yaş bilgisi de oransal bir değişken olarak veri tabanında yer almaktadır. Bu durum hangi veri madenciliği yönteminin kullanılması ya da veri üzerinde dönüşüm yapılıp yapılmaması gerektiği gibi sorunları akla getirmektedir. Aralıklı ve oransal ölçekli veriler için:

- Öklid uzaklığı
- Ölçekli Öklid uzaklığı
- Minkowski uzaklığı
- Mahalanobis uzaklığı
- Hotteling T^2 uzaklığı
- Canberra uzaklığı
- Manhattan City-Block uzaklığı
- Açısal benzerlik ölçüsü
- Korelasyon benzerlik ölçüsü

hesaplamaları kullanılabilir.

Oransal ya da aralıklı ölçeğe sahip verilerde kullanılan uzaklık ölçümleri nominal ve aralıklı ölçeklerde işe yaramayacaktır. Nominal ölçekli verileri uzaklık değerleriyle değil benzeşme değerleri ile kümelemek daha uygun olacaktır. İkili değişkene sahip nominal değişkenler için kullanılan benzeşme yöntemleri:

- Jaccard benzerlik katsayısı
- Ochiai benzerlik katsayısı
- Rao benzerlik katsayısı
- Basit eşleşme benzerlik katsayısı
- Binary Öklid uzaklığı

- Binarykaresel Öklid uzaklığı

olarak söylenebilir

4.1.1.Aralıklı Ve Oransal Ölçeklere Göre Uzaklik Ölçümü

Aralıklı ve oransal ölçekler sayısal olarak gösterilen matematiksel işlemlerin yapılabildiği ölçeklerdir. Bu ölçeklerdeki verilerde ölçümler arası düzen ve sabit bir uzaklık mevcuttur. Bu iki tip ölçek matematiksel işlemlere uygun olduğu için aynı zamanda herhangi bir dönüşüm gerektirmeden ortalama, standart sapma vb. Tüm istatistiksel işlemlerin yapılabilmesine olanak tanımaktadır (Nakip, 2013, s.192). Aralıklı ölçekle oransal ölçeğin anlamsal olarak farklılıkları bulunmaktadır. Oransal ölçekte sıfır bir yokluk ifade ederken aralıklı ölçekte etmemektedir. Ayrıca oransal ölçeğe sahip veriler birbirlerinin eş katlarıyken bu durum aralıklı ölçeğe sahip verilerde gözükmemektedir. Örnek vermek gerekirse; Ocak ayının 1. ve 2. Günleri arasında iki katlık biroran yoktur. Aynı şekilde bir takvimde sıfırıncı gün sadece o an bulunulan gündür ve herhangi bir yokluğu göstermemektedir. Ancak 10 kg olan bir nesne 20 kg olan bir nesneden iki kat daha ağırdır ve eğer nesne 0kg ise öyle bir nesnenin olmadığı söylenebilmektedir. Bu ölçek tiplerindeki uzaklık ve benzerlikler aşağıdaki yöntemlerle hesaplanmaktadır.

4.1.1.1. *Öklidyen Uzalik Ölçümü*

En çok kullanılan uzaklık ölçümü yöntemidir. Öklid uzaklığı hesaplanırken veriler hakkında bir ön bilginin bilinmesi gerekli olmadığı için daha çok tercih edilmektedir (Akın, 2008, s.19).

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$ Öklidyen Uzaklık formülü

Eğer veriler arasında belirgin bir önemlilik varsa yukarıdaki formül şu şekilde düzenlenmektedir:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} w_k \left(x_{ik} - x_{jk}\right)^2}$$

d: noktalar arasındaki uzaklığı gösterir.

i: p boyutlu nesnedeki i'nci terimi niteler.

j: p boyutlu nesnedeki j'nci terimi niteler.

k: k boyutlu nesnedeki k'ncı terimi niteler.

p: olası tüm durumlar yani nesnelerin tamamıdır.

4.1.1.2. *Canberra Uzaklik Ölçümü*

Canberra uzaklığı 1966 yılındaLance ve Williams tarafından "Computer Programs ForHierarchicalPolytheticClassification" adlı makalede bahsedilmiştir. (Jurman vd., 2009). Bu uzaklık ölçümü sıfıra yakın olan ve negatif olmayan değerler için uygundur. Canberra uzaklığının formülasyonu şu şekildedir:

$$d(x_i, x_j) = \frac{\sum_{k=1}^{p} |x_{ik} - x_{jk}|}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}$$

d: noktalar arasındaki uzaklığı gösterir.

i: p boyutlu nesnedeki i'nci terimi niteler.

j: p boyutlu nesnedeki j'nci terimi niteler.

k: k boyutlu nesnedeki k'ncı terimi niteler.

p: olası tüm durumlar yani nesnelerin tamamıdır.

4.1.1.3. *Hotteling T^2*

İki grubun ya da iki kümenin uzaklığını hesaplamada kullanılır. Hotteling T^2 uzaklığının formülasyonu şu şekildedir:

$$T^2 = \frac{n_1 n_2}{n}\left(\bar{x}_i - \bar{x}_j\right)^{-1} s^{-1}\left(\bar{x}_i - \bar{x}_j\right)$$

n_1: 1. kümenin ya da grubun eleman sayısı
n_2:2. kümenin ya da grubun eleman sayısı
s: verilerin tamamının standart sapması
$\bar{x}_i$: i.nci nesnenin vektörü
$\bar{x}_j$: j.nci nesnenin vektörü

4.1.1.4. *Minkowski Uzakliği*

Minkowski uzaklığı, Öklidyen uzaklığın ve Manhattan uzaklığının öklidyen uzaydaki genelleştirilmiş halidir. Minkowski uzaklığı şu şekilde formüle edilmiştir:

$$d(x_i, x_j) = \left\{\sum_k^p \left|x_{ik} - x_{jk}\right|^{\lambda}\right\}^{\frac{1}{\lambda}}$$

Değişkenler önem derecelerine göre ayrılmışsa, ağırlık etkisinin formüle eklenmesi gerekmektedir. Bu değişiklik:

$$d(x_i, x_j) = \left\{\sum_k^p w_k \left|x_{ik} - x_{jk}\right|^{\lambda}\right\}^{\frac{1}{\lambda}}$$

şeklinde gösterilmektedir.

d: noktalar arasındaki uzaklığı gösterir.

i: p boyutlu nesnedeki i'nci terimi niteler.

j: p boyutlu nesnedeki j'nci terimi niteler.

k: k boyutlu nesnedeki k'ncı terimi niteler.

p: olası tüm durumlar yani nesnelerin tamamıdır.

$\boldsymbol{\lambda}$: kuvvet ortalaması

$\boldsymbol{\lambda}$değeri ne kadar büyürse denklem Chebyshev uzaklığına o derece yaklaşacaktır.

4.1.1.5. *Manhattan City Block Uzakliği*

Bu uzaklık ölçümü Minkowski uzaklık ölçümündeki $\boldsymbol{\lambda}$ değerinin 1 olduğu durumdaki halidir. Ayrıca HermannMinkowski'nin ortaya attığı öklidyen uzaklık hesaplaması yerine ortaya attığı bir hesaplama yöntemidir(Gardner, 1997).Bu uzaklık ölçümü şu şekilde formüle edilmektedir.

$$d(x_i, x_j) = \sum_k^p |x_{ik} - x_{jk}|^1$$

d: noktalar arasındaki uzaklığı gösterir.

i: p boyutlu nesnedeki i'nci terimi niteler.

j: p boyutlu nesnedeki j'nci terimi niteler.

k: k boyutlu nesnedeki k'ncı terimi niteler.

4.1.1.6. *Mahalanobis D^2 Uzaklığı*

Öklidyen ve Minkowski uzaklıklarının ölçümlerinde karşılaşılan sıkıntılardan kurtulmak amacıyla değişkenler arasındaki korelasyonunda hesaplamaya eklenilmesiyle oluşan uzaklık ölçüm birimidir. Formülden de anlaşılabileceği gibi korelasyonun sıfır olduğu bir durumda sonuç Öklid değerinin karesine eşit olacaktır. Mahalanobis D^2 değeri şu şekilde formüle edilmektedir:

$$d_{ij} = (x_i - x_j) \sum^{-1} (x_i - x_j)$$

d: noktalar arasındaki uzaklığı gösterir.

i: p boyutlu nesnedeki i'nci terimi niteler.

j: p boyutlu nesnedeki j'nci terimi niteler.

k: k boyutlu nesnedeki k'ncı terimi niteler.

Σ : i ve j değerlerinden oluşan kovaryans matrisi değeridir.

4.1.1.7. *Korelasyon Benzerlik Ölçümü*

Korelasyon değeri ile hesaplanan benzerlik ölçümü değeri, Öklidyen ve Mahalanobis uzaklık ölçümleri gibi sıklıkla kullanılmaktadır. Bu değerin elle hesaplanması zor olacağından günümüz teknolojilerinin gelişimiyle bu ölçüm imkanının kazanıldığı söylenebilmektedir (Ma ve Chow., 2004).Korelasyon benzerlik ölçümünün formülasyonu şu şekildedir:

$$s_{ij} = \frac{(x_i - \bar{x})^T (x_j - \bar{x})}{\sqrt{(x_i - \bar{x})^T (x_i - \bar{x}) (x_j - \bar{x})^T (x_j - \bar{x})}}$$

4.1.1.8. *Açısal Benzerlik Ölçüsü*

Açısal benzerlik ölçümü, iki değişkene sıfır noktasından çekilen vektörlerin aralarındaki açının kosinüsü olarak hesaplanmaktadır. Bu değerin sıfır ile bir aralığında olması benzerlik olduğunu göstermektedir. Ölçümün denkleminden de anlaşılabileceği gibi hesaplama işlemi pozitif değerler ile yapıldığında doğru sonucu vermektedir. Ayrıca açısal benzerlik ölçümü çok boyutlu çözümlemeler için kullanılabilmektedir.Açısal benzerlik ölçümü:

$$s = \frac{u^T v}{\sqrt{u^T u v^t v}}$$

şeklinde hesaplanmaktadır.

u: değişken vektörü

v: değişken vektörü

s: benzerlik değeri

4.1.2. Nominal Değişkenler Için Benzerlik Ölçümü

Nominal ölçeğe sahip verilerde benzeşme değerini ölçmek için kontenjans tablosu kullanılmalıdır. Benzeşme değeri bilinmek istenen iki verinin olasılık uzayda birlikte olup olmamalarına göre oluşturulan tablo şu şekildedir (Tablo 5):

Tablo 5.Kontenjans Tablosu

	j.Gözlem			
i.Gözlem		Var	Yok	Toplam
	Var	a	b	a+b
	Yok	c	d	c+d
	Toplam	a+c	b+d	a+b+c+d

İkili değişkenler için kullanılan bu kontenjans tablosundaki verilerden yola çıkarak benzerlik ölçümleri hesaplanabilmektedir.

Tablo 6.Benzerlik Ölçüleri

BENZERLİK ÖLÇME YÖNTEMLERİ	**DENKLEMLER**
İkili Öklid	$\sqrt{b+c}$
İkili Karesel Öklid	$b+c$
Jaccard Benzerlik	$\frac{a}{a+b+c}$
Ochiai Benzerlik	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Rao	$\frac{a}{a+b+c+d}$
Basit Eşleşme	$\frac{a+d}{a+b+c+d}$

(Kaynak: Vatansever, M. (2008). *Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması.* (Yayınlanmamış Yüksek Lisans Tezi). Yıldız Teknik Üniversitesi/Fen Bilimleri Enstitüsü, İstanbul.)

Eğer nominal ölçekteki veriler ikili olarak değişmiyorlarsa (örneğin: ilçe, şehir gibi aralarında hiyerarşi olmayan değişkenler) aşağıdaki formül aracılığıyla değişkenler arasındaki uzaklık hesaplanmaktadır:

$$d(i,j) = \frac{p-m}{p}$$

p: toplam değişken sayısı

m: eşleşen değişken sayısı

4.1.3.Ordinal Değişkenler Için Benzerlik Ölçümü

Keyfi bir sıralama ölçütüne dayanan ve değerler arasındaki farkların anlamlı olmadığım ölçüm birimidir. Örneğin bir firmanın yaptığı ankette 5'li likert tipi olduğu varsayılsın. Cevap olarak "çok beğeniyorum","beğeniyorum","kararsızım","beğenmiyorum","hiç beğenmiyorum" şıklarını içeren bir ankette bu seçimlerin arasında matematiksel bir işlem yapılamamaktadır. Bir ürüne kararsızım diyen birisi hiç beğenmiyorum diyen birine göre üç kat beğenmiş denilemez. Bu sebeple verileri sıfır ve bir aralığına indirgeyerek Kısım 4.1.1. ve Kısım 4.1.2. başlıkları altında bahsedilen denklemler aracılığıyla uzaklık ya da benzeşme değerleri hesaplanabilmektedir.

$$z_{if} = \frac{r_{if}-1}{M_f-1}$$

r_{if}: herhangi bir x_i değerinin veri kümesindeki sıralamış durumdaki yerini belirtmektedir.

M_f: işlem yapılan veri kümesinin sıralanmasıyla elde edilen yeni verilerin en büyük değeri

4.2.KÜMELEME ANALİZİ TEKNİKLERİ

Önceki konusunda da bahsedildiği gibi kümeleme analizi gerçekleştirilirken verilerin ölçekleri ve analiz yönteminin büyük hacimli veri tabanlarında uygulanması gibi problemlerle karşılaşılmaktadır. Bu sebeple istenilen ölçümleri yapabilecek kümeleme algoritmaları geliştirilmiştir. Kümeleme analizi teknikleri genel olarak Tablo 7'de gruplandırılmıştır ve bu gruplamalara göre olan özellikler verilmiştir (Tablo 7).

Tablo 7. Kümeleme Analizi Teknikleri

Yöntem	**Genel Özellikler**
Bölmeli	Mutlak ayrık kümeleri bulma
	Uzaklık ölçümü temelli
	Küme merkezlerini bulmada ortalama ya da medoid kullanır
	Küçük ya da orta hacimli veri setlerine uygundur
Hiyerarşik	Hiyerarşik bir ayrışmayı varsayar
	Hatalı veri olan veri setlerin uygun birleştirme ya da ayrılma yapamaz
	Mikro kümeleme ya da bağlantı teknikleri gibi diğer teknikleri içerebilir
Yoğunluk Temelli	Rasgele şekillerdeki kümeleri oluşturabilir
	Veri setinde düşük yoğunluklara göre ayrım yapabilir
	Uç değerleri ayırabilir
Izgara Temelli	Çoklu sonuç üretebilir
	Hızlı işlem yapar

(Kaynak: Vatansever, M. (2008). *Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması.* (Yayınlanmamış Yüksek Lisans Tezi). Yıldız Teknik Üniversitesi/Fen Bilimleri Enstitüsü, İstanbul.)

4.2.1.Bölmeli Yöntemler

Bölmeli yöntemler en basit ve temel kümeleme analiz yöntemidir. Analiz için, n nesneden oluştuğuve her bir kümenin en az bir nesne içerdiği varsayılır (Kanungo vd., 2002). Böylelikle k adet kümeden oluşan analiz gerçekleştirilirken kümeleme işlemi her bir nesnenin ayrı ayrı hesaplanması yapılarak küme aidiyetlikleri belirlenir (Han vd., 2012). Bölmeli yöntem uygulanırken genellikle uzaklık ölçümleri hesaplanır. Bu ölçümler sonucu nesnelerin hangi kümeye ait oldukları bulunur. Aşamalı olarak yapılan bu işlemde yakın olan ya da benzer olan verilerin aynı kümeye dâhil edilebileceği gibi,

bu kümeleme yönteminde en benzeyen ya da aralarındaki mesafe en çok olan verilerin ayrılması şeklinde de uygulama yapılabilmektedir. Bu yöntemde de amaç süreç sonunda oluşturulan kümelerin kendi içlerinde sahip oldukları verilerle en çok benzeşmesi ve diğer küme üyeleriyle en az benzeşmesidir.

Bu başlık altında en yaygın olarak kullanılan k-Means, k-Medoids ve bu iki yöntem temel alınarak oluşturulan algoritmalardan bahsedilecektir. Bu yöntemlerden k-Means tamamen nümerik veriler ile çalışabilirken, k-Medoids kategorik veriler ile çalışabilmektedir (Elkan, 2003). Ayrıca k-medoids algoritmasından geliştirilen CLARA ve CLARANS algoritmaları tüm ölçek tipleriyle çalışabilmektedir. Her bir algoritma aykırı değerlere karşı duyarlıdır ve bu durum analizin sonucunu etkilemektedir. Ayrıca veri setinin örüntüsüne bakılacak olunursa bu algoritmalar konveks olmayan şekle sahip kümelerde daha uygun sonuçlar vermektedir.

4.2.1.1.*K-Means Algoritmasi*

K-Means algoritmasıyla n adet veriden oluşan bir veri setinden merkez noktası c ile temsil edilen k adet küme oluşturulmaktadır (Vattani, 2011). K-Means algoritmasına Lloyd's algoritması da denilmektedir(Frahling ve Sohler, 2005). Bu algoritma şu şekilde gösterilebilmektedir (Lin veWu, 2009):

1. k adet küme keyfi olarak atanır ve küme merkezleri belirlenir.
2. Herbir veri belirlenen küme merkezlerine göre yakınlıkları belirlenerek ilgili kümeye dahil edilir.
3. Her bir küme merkezi tekrardan hesaplanır.
4. 2.ve3. adımlar tekrarlanır.

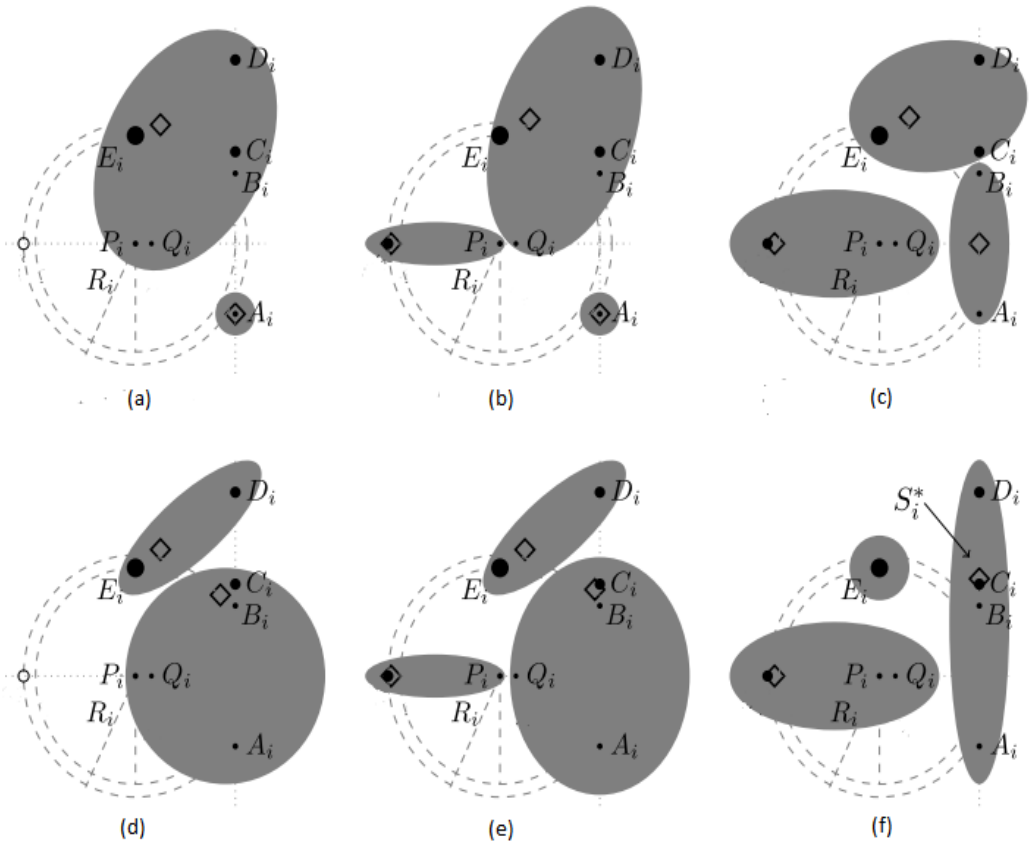
Küme merkezleri değişmeyecek hale gelene kadar algoritma kendisini tekrar eder. Merkezin değişmeyecek hale gelmesi, hata kareler toplamının minimum olmasıyla anlaşılır. Hata kareler toplamının minimum olması şu denklem ile bulunmaktadır:

$$J_k = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2$$

J_k: k. kümedeki nesnelerin hata kareleri toplamı

x_i: i.nesneye ait değer

m_k: k.kümenin merkez noktası



Şekil 12. K-Means Yönteminin Bir Saat Üzerindeki Kümeleme Çalışması

(Kaynak: Vattani, A. (2011). K-Means Requires Exponential MAny Iterations Even in the Plane. Discrete Computer Genom. 45, 596-616)

Şekil 12.'de de görüldüğü gibi küme merkezleri Şekil 12 (a)'da rastgele koyulmuştur. Devam eden altı adımda kümeler ve küme merkezleri değişerek en uygun

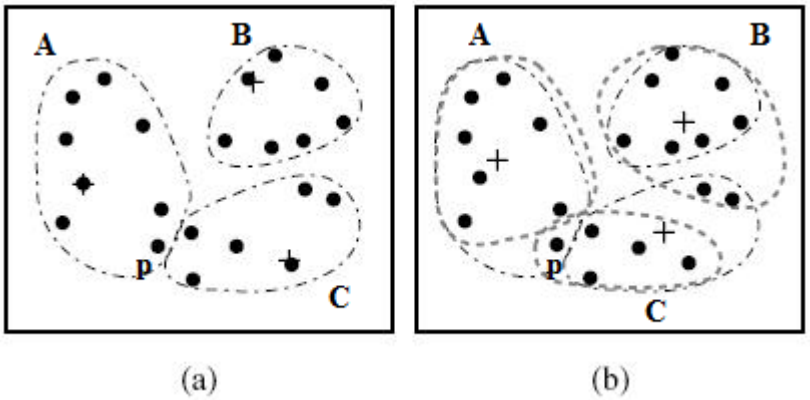
hale gelmiştir. Bu örnekte normal bir kümeleme analizinin yanı sıra şekildeki noktaları bazılarının büyük olması verilerin ağırlıklandırılmış olduğunu göstermektedir. İki küme olması hedeflenen örnekteki çalışmada en son durum üç kümeyle sonuçlandırılmıştır.

K-Means kümeleme algoritması öklidyen uzaklık hesaplamalarını kullandığı için bu yöntem nominal ve ordinal ölçekli verilerde uygun sonuç vermeyecektir (Han vd., 2012, s.454). Ayrıca elde edilen kümelerin konveks şekillere sahip olduğuna dikkat edilmelidir. Uygun olmayan ölçekli verilere sahip veri setlerinin çözümlenme problemini benzerlik katsayıları kullanarak ya da küme ortalaması yerine mod değeri hesaplanarak aşılabilmektedir. Ayrıca bu algoritma uç değerlere karşı hassas olmamaktadır. Aksine uç değerin varlığı ya da analiz yapan kişinin yanlış küme sayısı ile yapacağı bir bölümlendirme, uç değerlerin analiz sonuçlarını etkilemesine yol açacaktır.

4.2.1.2.*K-Medoids Algoritmasi*

K-Medoids algoritmasında küme merkezini temsil eden nokta medoid olarak adlandırılır. Bu algortimadaki amaç aynı k-Means algoritmasında da olduğu gibi veri setini k adet kümeye bölerek birbirine benzeyen ya da yakın olan verilerle oluşturulan kümeleri diğer kümelere göre farklı hale getirmektir (Işık ve Çamurcu, 2007). K-Medoids algoritması 1987 yılında Kaufman ve Rousseeuw tarafından ortaya atılmıştır.

K-Medoids algoritması da k-Means algoritmasındaki aynı işlem adımlarına sahiptir. Tek farkı keyfi olarak küme merkezlerinin seçilmesi ve bunun üzerine işlem yapılmasındansa küme sayısı kadar veri medoid olarak seçilerek tekrarlama işlemleriyle kümeler ve veri aidiyetlikleri belirlenir. K-Medoids algoritması uç değerlere karşı daha az hassastır. Ayrıca algoritmanın baştan çalıştırılmasıyla oluşacak analizler her zaman aynı sonucu vermeyebilmektedir. Bu olaya sebep olan durum medoidlerin keyfi olarak seçilmesidir.



Şekil 13. K-Medoids Yöntemi Ile Kümeleme Gösterimi

(Kaynak: http://www.iszekam.net/image.axd?picture=2009%2F5%2F3-11%28k-meds%29.jpg, 24.06.2013)

Şekil 13 (a)'da görüldüğü gibi küme medoidleri şekildeki gibi oluştuğunda p verisi A kümesine ait olmaktadır. Ancak küme medoidleri Şekil 13(b)'deki gibi olursa p verisi B kümesine ait olmaktadır.

K-Medoids yöntemini daha iyi açıklayabilmek gerekirse algoritmanın adımları şu şekildedir:

1) k adet nesne seçilir.
2) Verilere en yakın olan medoidlerle küme oluşturulur ve aidiyetlikler belirlenir.
3) Medoid olmayan rasgele bir veri seçilir.
4) 3.basamakta seçilen veri medoid gibi kabul edilir ve verilere olan yeni uzaklıklar hesaplanır.
5) Temsili medoid olarak belirlenen veri daha iyi sonuç veriyorsa yeni medoid olarak atanır.
6) Sistemin performansı değişmeyecek hale gelene kadar 2. ve 5. Adımlar arası tekrar edilir.

Sistemin performansı şu şekilde hesaplanmaktadır;

$$maliyet(x, c) = \sum_{i=1}^{k} |x_i - c_i|$$

x: herhangi bir veri

c: herhangi bir medoids

k: küme sayısı

Başta belirlenen medoid ile yapılan hesaplama sonucu, temsilen seçilen medoid ile yapılan hesaplama sonucundan çıkartılır. Sonuç sıfırdan büyükse bu ilk seçilen değerin daha iyi olduğunu ve medoidin değiştirilmemesi gerekir. PAM (PartitioningAroundMedoids) ve CLARA (Clustering LargeAplications) algoritmaları üretilen ilk k-medoids algoritmalarıdır. Büyük veri setleriyle çalıştırıldığında istenilen hassasiyeti gösteremeyen PAM algoritması yerine CLARA geliştirilmiştir.

4.2.1.3.*Clara ve Clarans*

CLARA algoritması 1990 yılında k-medoids yönteminin geliştiricileri Kaufman ve Rousseeuw tarafından ortaya atılmıştır. Ayrıca Raymond T. Ng ve Jaiwei Han tarafından 1994 yılında CLARANS (Clustering LArgeAplicationsBased on RandomizedSearch) yöntemi geliştirilmiştir (Pilevar ve Sukumar, 2005; Krishnapuram vd., 1999). Bir önceki kısımda da belirtildiği gibi büyük veri setlerinde çalışmak adına geliştirilen CLARAalgoritması veri setinden örneklemeler alarak analizi gerçekleştirir. Oluşturulan daha küçük boyutlardaki veri setlerinde medoid belirlemek için tekrardan PAM algoritması kullanılır. Veri setinden seçilen örneklemlerin rassal olarak seçilmesine dikkat edilmelidir. Böylelikle seçilen verilerle oluşturulan medoidlerin temsili de daha iyi olacaktır. CLARA üzerine yapılan çalışmalar sonucunda veri setinden 5 örneklem seçilmesi ve her örneklemde toplam veri sayısının 40+2k'sı kadar veri olması durumları sonuçların anlamalı olduğunu ispatlamıştır(Barioni vd., 2006).

CLARA algoritması şu şekildedir (Ng ve Han, 2002):

1. i= 1'den 5'e kadar aşağıdaki adımlar tekrar edilir.
2. Tüm veri setinden 40+2k'lık örneklem seçilir ve PAM algoritması ile medoid bulunur.
3. Veri setindeki tüm nesneler için en uygun k medoid bulunur.
4. 3.adımda bulunan kümelerin ortalama uzaklıkları hesaplanarak 2. adımda bulunan medoidler ile karşılaştırılır. Hangi medoid en uygun ise küme medoidi olarak atanır.
5. 1. adıma geri dönülür.

CLARA algoritmasında önce örneklem belirlenir ve PAM algoritmasıyla da medoidler belirlenirken CLARANS algoritmasında her ikisi de birlikte kullanılır. Bu bağlamda CLARANS algoritması PAM ve CLARA algoritmalarının birleşimidir (Chen vd., 1996). Bu analiz yöntemi kısaca, n adet veriden oluşan veri setinden k adet küme seçebilmek için veri setini k adet bölüme ayırır. Böylelikle her bir bölme bir kümeyi ifade eder (Zhang vd., 2004). CLARA algoritmasındaki eksiklik olan veri setinin büyüklüğüne göre analizin başarısının değişmesi CLARANS ile giderilmiştir (Koperski vd., 1998).

CLARANS algoritmasında oluşturulan her bir küme ya da bölüm bir düğüm olarak ele alınmaktadır. Buradan yola çıkarak her bir düğümün de bir çözüm olabileceği yorumu yapılmaktadır (Ng ve Han, 2002).Bu düğümler yani kümeler k medoidleri olarak adlandırılmaktadır.

CLARANS algoritması şu şekildedir (Akın, 2008, s.84):

1. Rassal olarak k adet aday medoid seçilir.
2. Rassalolarakseçilennoktalardanbirtanesininseçilmemişbirnoktailedeğişebilmesiir delenir.
3. Bu değişimsistemidahaiyibir hale getirecekse, yaniperformansdeğeridüşecekse 2.adım tekraredilir.

4. Tersiolarak değişimden sonra performans değeriartacaksabaştavarolan durum tercihedilir.
5. Her ikiçözümdekarşılaştırılır.
6. En uygunçözümeulaşılamamışsa ilk adımdanitibarenalgoritmatekrareder.

Algoritmada yapılacak olan tekrar sayısı en fazla $[250, k(n-k)]$ kadar olmalıdır (k: küme sayısı ve n: toplam veri sayısı).

Daha önceki yöntemlerde olduğu gibi algoritmanın sonuçlandığını performans değeri bulunarak karar verilir. Algoritmanın 3. ve 4. adımlarında bahsedilen hesaplama olanperformans değerişu şekilde hesaplanmaktadır:

$$\text{maliyet}(M, D) = \frac{\sum_{i=1}^{n} benzemezlik(O_i, temsil(M, O_j))}{n}$$

D: Kümelenecek veri seti

M: Medoid

O_i:Veri setindeki herhangi bir veri

O_j:Veri setindeki herhangi bir veri

n:Veri setindeki toplam veri sayısı

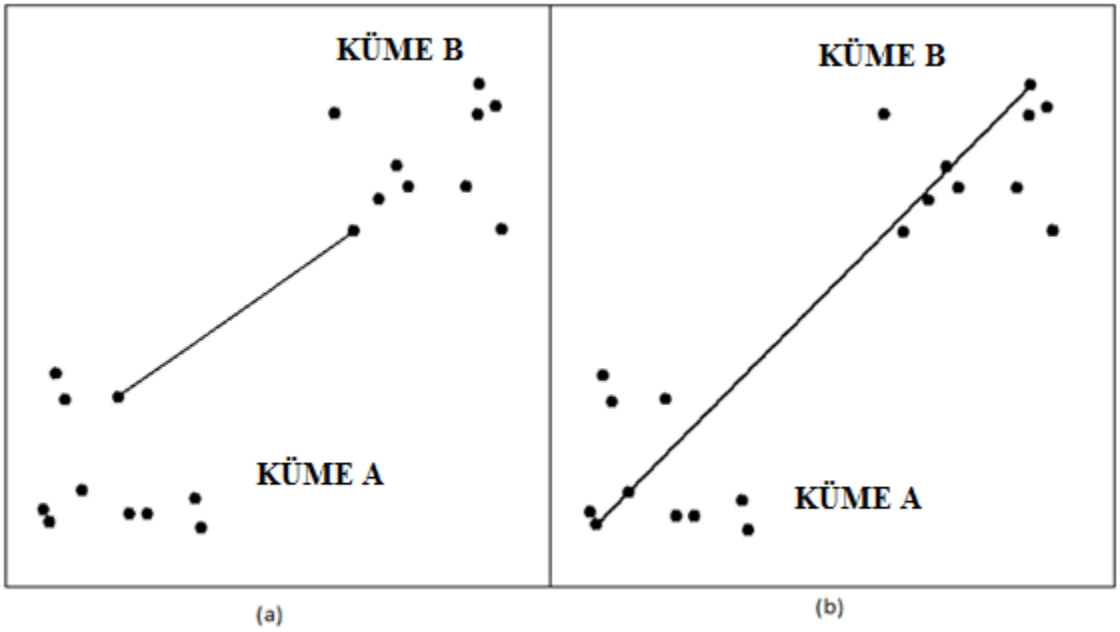
4.2.2.Hiyerarşik Kümeleme Analizi

Hiyerarşik kümeleme analizi, veri madenciliğinde kullanılan temel kümeleme analizlerinden birisidir. Bu analiz yöntemi de Gruplayıcı ve Ayırıcı olarak iki ana başlıkta anlatılabilir (Atbaş, 2008, s.15). Elbette bilişim teknolojilerinin gelişmesi analizler yapılırken ki karşılaşılan ölçek, veri tabanı hacmi gibi problemlerin aşılmasında yeni yöntemlerin geliştirilmesine olanak vermiştir. Temel olarak iki başlık altında toplanan hiyerarşik analiz yöntemlerinde daha iyi sonuçlar elde edilmesi amacıyla çeşitli algoritmalar da geliştirilmiştir. Bu bölümde Toplayıcı ve Ayırıcı yöntemlerden ve bunların bağlantı metotlarıyla nasıl gerçekleştirildiği, ayrıca zamanla geliştirilen algoritmalardan BIRCH, CURE, ROCK ve CHAMELEON algoritmalarından bahsedilecektir.

Hiyerarşik kümeleme analizi yöntemleri kullanılırken başlangıçta kaç küme olduğu bilinmemektedir (Altun Ada, 2011). Kullanılan yönteme göre başlangıçta n adet veriden n adet küme ya da n adet kümeden 1 adet küme oluşturulur. Buradan da anlaşılacağı gibi her adımda küme sayısı k-1 ya da k+1 olacak şekilde değişmektedir. Hiyerarşik kümeleme analizi hacimli veri tabanlarında uygun sonuçlar veremeyebilir (Öz vd., 2009). Bunun sebebi de; analiz adımsal olarak ilerlerken bir gruba üye olmuş değişken diğer adımlarda başka bir kümeye dahil olamayacağıdır (Sibson, 1972). Ancak hiyerarşik kümeleme algoritmalarının tüm ölçek tiplerini kullanılabilir olması, nesnelerin yoğunluğu konusunda esnek olması ve uzaklık ölçülerini kullanabilmesi, bu yöntemin tercih edilebilirliğini açıklamaktadır. Hiyerarşik kümeleme analizi sonuçlarını ve süreç adımlarını ağaç benzeri grafikte göstermek mümkündür. Bu yapıya "dendogram" denilmektedir.

4.2.2.1.*Tek Bağlanti Ve Tam Bağlanti Teknikleri*

Tek bağlantı tekniğindeki amaç, iki küme arasındaki uzaklığın en az olduğu durumu belirlemek ve bu iki kümeyi tek küme olarak birleştirmektir (Penrose, 1995; Allen vd., 1991). Böylece toplam küme sayısı azalmış olacaktır. Tek bağlantı tekniğinde kümeler arasındaki uzaklık durumunu en yakın küme elemanlarının yakınlığı ya da en uzak küme elemanlarının uzaklığı şeklinde ölçmek mümkündür.



Şekil 14. Bağlanti Yöntemleri

(Kaynak: Rao., A. R. ve Srinivas, V. V. (2005). Regionalization of Watersheds by Hybrid-Cluster Analysis. *Journal of Hydrology*, 318, 37-56)

Şekil 14 (a)'da da görüldüğü gibi A ve B kümelerinin üyeleri arasında en yakın iki noktanın yakınlığı alınarak hesaplama uzaklık hesaplaması yapılmaktadır. Başka bir şekilde, Şekil 14 (b)'de de A ve B kümelerinin en uzak noktanın arasındaki mesafe hesaplanmaktadır. Bahsi geçen yöntemlere ait hesaplamalar:

$$D(X,Y) = \min_{x\in X, y\in Y} d(x,y)$$

$$D(X,Y) = \max_{x\in X, y\in Y} d(x,y)$$

şeklinde hesaplanmaktadır.

X: X kümesi

Y: Y kümesi

x: X kümesindeki herhangi bir eleman

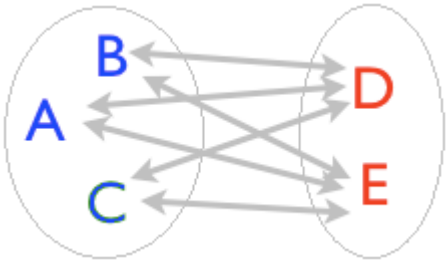
y: Y kümesindeki herhangi bir eleman

D,d: Uzaklık değeri

Uzaklık değerleri daha önce bahsedilen ölçüm yöntemleriyle yapılabilmektedir.

4.2.2.2.*Ortalama Grup Bağlanti Tekniği*

Ortalama bağlantı tekniği, bir kümedeki verilerin diğer bir kümedeki verilerle olan uzaklıklarının ortalaması olarak hesaplanır. Kümeler arasındaki ortalama uzaklık hangi iki küme arasında daha az ise bu iki küme birleştirilerek yeni bir küme oluşturulur. Bu yöntem, biyoloji alanında, canlıların köken araştırmalarında yani türlerin hangilerinin ortak kökene sahip olduğu ya da türlerin ne zaman ayrılarak farklı bir tür olduğunu anlamak amacıyla sıklıkla kullanılmaktadır.



Şekil 15. Ortalama Bağlanti Gösterimi

Şekil 15'te de gösterildiği gibi; iki kümedeki elemanların kombine olarak diğer kümedeki elemanlarla arasındaki uzaklık hesaplanır. Kullanılan yöntemin formülizasyonu şu şekildedir:

$$D(X,Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X} \sum_{y \in Y} d(x,y)$$

X: X kümesi

Y: Y kümesi

x: X kümesindeki herhangi bir eleman

y: Y kümesindeki herhangi bir eleman

D,d: Uzaklık değeri

Bu yöntem her adımda iki benzeşen ya da yakın kümenin birleştirilmesi şeklinde çalışmaktadır.

4.2.2.3.*Ward Tekniği*

Ward tekniğinde önceki tekniklerdeki gibi grup elemanları arasındaki uzaklık hesaplanmaz. Bu teknikte yöntem birleştirilecek kümelerin elemanlarının toplam hatayı en az olacak şekild kabul etmesidir. Şöyle ki; iki küme birleştirildiğinde grup içi hata diğer bir kümeyle birleştirilmesine kıyasla daha az olacaksa uygulama sonlanır, aksi bir durum söz konusuysa üçüncü olan diğer küme ile birleşim sağlanır. Tekniğin matematiksel olarak üstün oluşu ve rahatlığı günümüzde birçok alanda kullanılmasına olanak tanımaktadır(Cohen ve Shannon, 1981). Ward tekniğinde hata kareler şu şekilde hesaplanmaktadır:

$$W_k = \sum_{i=1}^{p} \sum_{j=1}^{n_k} \left(x_{ijk} - \bar{x}_{ik}\right)^2$$

W_k: k kümesinin Ward değeri (hata kareler toplamı)

p: p kümesi

n_k: k kümesindeki eleman sayısı

i: i. veri

j: j. veri

k: herhangi bir küme

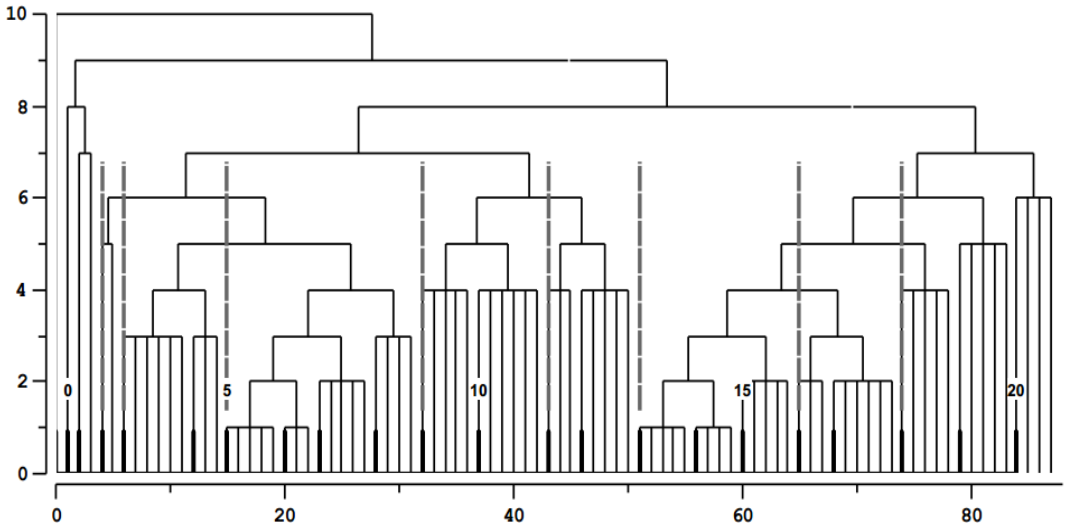
Bu formülde küme içi hata kareler hesaplanır ve her bir küme için aynı işlem uygulanır. Hesaplama işleminden sonra p ve q gibi iki küme olduğu varsayılsın. Bu iki kümenin birleşiminden t kümesi elde edilmiş olsun. Bu durumda Ward değerindeki artış:

$DW_{pq}=W_t-W_p-W_q$

şeklinde hesaplanmaktadır.

4.2.2.4.*Toplayıcı Küme Teknikleri*

Toplayıcı kümeleme analizi işlem alttan üste doğru olan ve her adımda kümelerin birleşmesini esas alan bir tekniktir (Taşkın ve Emel, 2010). Analiz her bir verinin kendi başına tek elemanlı bir küme oluşundan başlar ve en son noktada tüm elemanlar aynı kümede birleşmiş olurlar (Chidananda Gowda ve Ravi, 1995).



Şekil 16. Toplayıcı Kümeleme Için Dendogram

(Kaynak: Yaari, Y. (1999). *Segmentation of Expository Texts by Hierarchical Agglomerative Clustering.* 2nd International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 24 September.)

Dendograma örnek olarak Şekil 16 gösterilebilmektedir.

Toplayıcı kümeleme analizinin her bir adımında, başlangıçta n olan küme sayısı n,n-1,n-2,…..,1 olacaktır. Bu analizi yaparken süreç istenilen herhangi bir adımda durdurulabilir. Bu durdurma işlemine karar verecek olan çalışmayı yapan kişidir (Zhao ve Karypis, 2002). Toplayıcı kümeleme analizi her ne kadar hızlı olsa da bu durum analiz sonuçlarının uygunluğunu düşürmektedir. Ayrıca genellikle yatay bir düzlemde bağlantı kurmaya daha yatkın olan bu yöntem sistem hatasının artmasına sebep olmaktadır (Freanti vd.,2006).

4.2.2.5.*Ayırıcı Kümeleme Teknikleri*

Ayırıcı kümeleme tekniklerinde, toplayıcı kümeleme tekniklerinde olanın tersine bir işlem söz konusudur. Başlangıçta tüm veri setindeki elemanlardan olan 1 adet küme vardır ve en son adımda ise tüm verilerin kendi başlarına küme oluşturdukları n kümeli bir yapı mevcuttur. Algoritmanın amacı veri setini olası en ideal iki parçaya ayırmaktır. Bu tekniğin kategorik verilerden oluşan veri setlerini kümelemede k-Means ve Ward tekniklerine göre daha başarılı olduğu kanıtlanmıştır (Chavent vd., 2007). Analiz gerçekleştirilirken uygulanacak yöntemler uzaklık ölçümlerine ya da benzeşme değerlerine göre yapılacağından, toplayıcı tekniklerin tersine bu teknikte daha fazla işlem yoğunluğu vardır. Dolayısıyla hatalı sonuç ya da sonuçlar üretme olasılığı da aynı şekilde artmaktadır.

4.2.2.6.*Birch Algoritması*

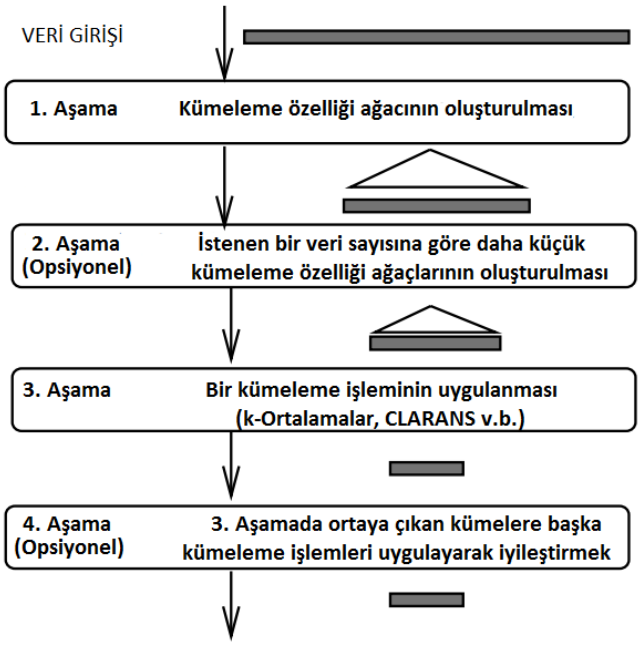
Veri madenciliğinde kümeleme analizlerinin denetimsiz yöntemler olmasından dolayı, sistemler dinamik çalışmalarda eksiklikler gösterebilmektedir. Ayrıca gerçek hayatta karşılaşılan analizlerde ya da yapılan çalışmanın içeriğine göre veriler aynı ağırlıkta olmayabilir. Önceki yöntemlerde de bahsedildiği gibi kümeleme analizinde büyük hacimli veri tabanlarında yapılan çalışmalarda uygun sonuçların üretilememesi problemiyle karşılaşılmaktadır. Bu problemi aşmak adına veri tabanları daha küçük

parçalara ayrılarak sonuç üretilmeye çalışılmaktadır. Bu durum da işlem yükünün veri analizinden ziyade işlem yapılacak en uygun parçalanmanın nasıl olacağı noktasında harcanmaktadır. Böylelikle işlem hacminin artması ek maliyet unsuru doğurmaktadır. BIRCH algoritması yapısı gereği bu tip bir işlem yoğunluğunu egale etmektedir. Ayrıca büyük hacimli veri tabanlarında ayrımın yatay bir düzlemde olması ya da dikey bir düzlemde olması başka bir problemdir(Fu vd., 1999). Eğer kümeleme yatay olacaksa az elemanlı fazla küme olacak, dikeyse çok elemanlı az küme olacaktır. Bu durumda karara varılamayacak kadar çok oluşmuş küme ya da kümeleme özelliği kaybetmiş, kesinliği azalmış kümelerle karşılaşılacaktır.

Veri tabanları yatay, dikey veya keyfi olarak ayrılmış olabilir. Bu durum ortak verilerin paylaşılması noktasında sonuçların güvenilir olmamasını sağlamaktadır ve hangi kümelerin uygun bir şekilde ayrılacağı problemiyle karşılaşılmasına sebep olmaktadır. Eğer veri tabanının yapısı bu şekildeyse BIRCH algoritması kümeleme analizinin gerçekleştirilmesi için uygun olacaktır(Prasad ve Pandurangan, 2005).

BIRCH algoritmasında kümeleme özelliği (CF) ve kümeleme özelliği ağacı (CF-tree) kavramlarıyla karşılaşılmaktadır. Kümeleme özelliği, veriler hakkındaki üç özelliğin (N,LS,SS) birleştirilmesidir (Zhang, Ramarkrishnan ve Livny, 1997). N, toplam veri sayısını; LS, verilerin doğrusal toplamını ve SS verilerin hata kareler toplamını ifade etmektedir. Kümeleme özelliği ağacı, alt kümelerde toplanan bilgilerin birleştirildiği, denge ağacıdır. Kümeleme özelliği ağacında dal faktörü (B: Branch) ve eşik (T: Threshold) olarak iki istatistik kullanılır. B, oluşabilecek alt düğüm sayılarını (child node) ve T, alt kümelerin barındırabileceği veri sayısını belirtir. Buradan yola çıkarak, T değeri büyük ise kümeleme özelliği ağacı küçük olacaktır, yorumunda bulunulabilir(Zhang, Ramarkrishnan ve Livny, 1997).

BIRCH algoritması dört aşamadan oluşmaktadır. Bu aşamalar Şekil 17'de gösterilmektedir.



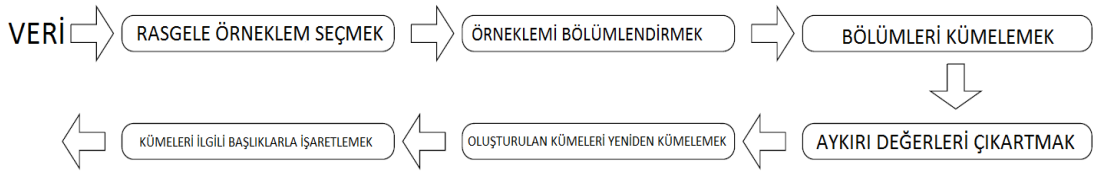
Şekil 17. BIRCH Algoritması

(Kaynak: Zhang, T., Ramarkrishnan, R. ve Livny, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery, 1,* 141-182.)

BIRCH algoritmasının faydaları olmasının yanı sıra, veri ekleme, çıkartma, merkez noktası bulma gibi işlemler yapması sebebiyle nominal ve ordinal ölçeğe sahip verilerde düzgün çalışmamaktadır (Ghanti vd., 1999).

4.2.2.7.*Cure ve Rock Algoritmaları*

CURE (Clustering Using Representative) algoritması küresel olmayan şekilli, aykırı ya da uç değerlere sahip ve büyük hacimli veri tabanlarında kümeleme yapmak için uygun bir algoritmadır (Guha vd.,1998). CURE algoritmasında kümeler için tek bir nokta değil birden fazla temsil notası belirlenir. Diğer kümeler arasındaki benzerlikler bu temsil noktalarına göre belirlenir. Benzerlik açısından kabul gören ve birleştirilen kümelerin yeni merkezleri, oluşturulan kümenin elemanları göre ve daraltma faktörü kullanılarak bulunur. Daraltma faktörü α ile gösterilir. CURE algoritması şu şekildedir:



Şekil 18. CURE Algortiması

(Kaynak: Guha, S.,Rastogi, Rastogi, R. ve Shim, K., (1998). CURE: An Efficent Clustering Algorithm for Large Databases, ACM SIGMOD Record, 27(2), 73-84.)

Küme merkezlerini belirlerken kullanılan daraltma faktörü aynı zamanda aykırı değerlerin tespit edilmesi ve hesaplamadan çıkartılmasına da yardımcı olmaktadır. Veri tabanının olası kümelere ayrılarak hesaplama yapılması işlem hızının arttırılmasını sağlamaktadır. En uygun ayrımın yapılması tek seferde olmayabileceği için şekil 18'deki işlem sırası sürekli kendini tekrar etmektedir.

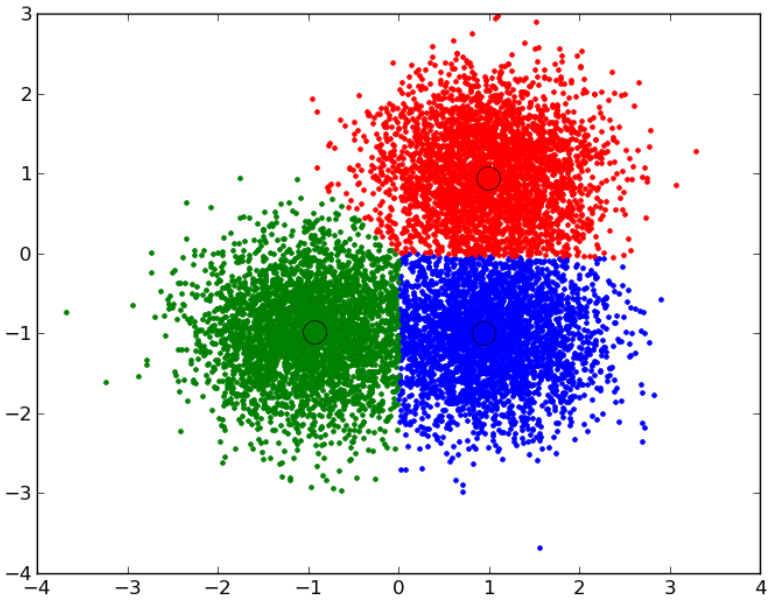
ROCK algoritmasında da süreç veri tabanından rasgele örneklem seçmekle başlar. Hiyerarşik kümeleme analizi yaparken kategorik verilerle analiz yapan diğer algoritmalardan olan STIRR, BUBBLE ve CACTUS yanı sıra ROCK algoritması en çok kullanılan algoritmadır (Dutta vd., 2005). Bu algoritmada kümelerin elemanlarının nisbi yakınlıkları hesaplanır ve birleştirilecek kümeler bulunur. Veri tabanından kaç küme oluşturulacağı (k) ve bu kümelerin benzerlikleri (Ɵ) algoritmadaki kriter değerlerdir ve sonuçların uygunluğuna bu parametrelere göre karar verilir. ROCK algoritması şu şekilde gösterilebilir (Şekil 19):

Şekil 19. ROCK Algoritması

(Dutta, M., Kakoti Mahanta, A. ve Pujari, A. K. (2005). QROCK: A Quick Version of The ROCK Algorithm for Clustering of Categorical Data. *Pattern Recognition Letters, 26,* 2364-2373.)

ROCK ve CURE algoritmaları hiyerarşik kümeleme teknikleri olduğundan doğası gereği modellemeler de statik olmaktadır. Dolayısıyla veri tabanına gelen yeni bir değişken özelliğine göre küme yapısını değiştirebilir, hatta küme aidiyetliklerini de değiştirebilir. Analizi gerçekleştiren kimsenin bunu dikkate alarak sadece sayısal ya da tablo üzerindeki durumları değil grafikler aracılığıyla da verileri takip etmesi gerekebilir.

CURE algoritmasında kümeler arasındaki temsili noktaların minimum uzaklıkları kullanılırken iki küme arasındaki verilerin bağlılığı hesaplanmaz. Benzer şekilde ROCK algoritmasında da kümelerdeki verilerin birbirlerine göre uzaklıkları hesaplanırken kümelerin sınır değerlerinin yakınlığı hesaplanmamaktadır. Bu durum elbette analizlerin uygunluğu noktasında çalışmacıları düşündürmektedir. Veri yapısına ve çalışmanın amacına göre en uygun model belirlenmesi çalışmacıya ait olacaktır.



Şekil 20. ROCK ve CURE Algoritmalarına Göre Birleşecek Küme Seçimleri

(Kaynak: http://scikit-learn.org/stable/_images/plot_mean_shift_11.png, 12.08.2013)

Şekil 20'de de görüldüğü gibi kümeler CURE algoritmasına göre birleşebilecekken ROCK algoritmasına göre birleşemezler. Çünkü küme merkezleri CURE algoritması için olması gerekenden uzaktadır.

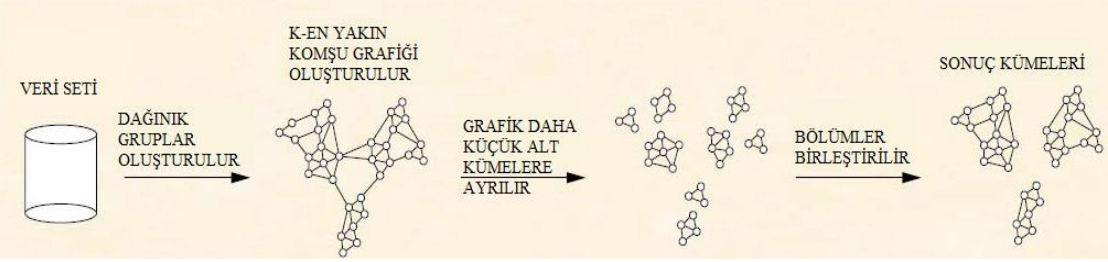
4.2.2.8.*CHAMELEON Algoritması*

CHAMELEON algoritması Karypis tarafından geliştirilmiştir. ROCK ve CURE algoritmalarındaki eksiklikleri gidermek için geliştirilmiş bu yöntem iki bölüme ayrılmıştır (Kogan vd.,2006, s.33; Rajagopal ve Selvi, 2006). İlk bölümde tüm veri tabanındaki veriler alınarak küçük kümelere ayrılır. Bu kümlerdeki veriler tartışmasız bir şekilde en çok benzeşen doğal kümlerdir (Abubaker, 2011, s.56). İkinci bölümde ise birleştirici kümeleme yöntemleri uygulanarak benzeşme ve yakınlık ölçüm değerlerine göre ilk adımda oluşturulan kümeler birleştirilir (Gupta vd.,2013). Herhangi bir veri setinden C kümesi oluşturulduğu varsayılsın ve bu kümelerin daha küçük kümelere ayrıldığı düşünülsün. Bu yeni kümeler C_i ve C_j gibi iki alt küme olacaktır. Buradan yola

çıkarak bağlanabilme (relativ interconnectivity) $RI(C_i,C_j)$ olacaktır. Aynı şekilde de bu iki alt kümenin yakınlığı (relative closeness) $RC(C_i,C_j)$ olacaktır.

CHAMELEON algoritmasında küçük kümelere bölme ve k-means yöntemi ile birleştirmenin yapılması sürecin dinamik ve oluşan kümelerin daha doğal olması sağlanmaktadır (Abubaker, 2011). Bu algoritmanın iki aşamalı olması ayrıca kümelerin rasgele şekiller içermesine ve rasgele yoğunluklarda olmasına karşı daha iyi sonuçlar üretmektedir. Ancak CHAMELEON algoritması çok boyutlu veri tabanlarında çözümleme için uygun değildir (Rafsanjani vd.,2012).

CHAMELEON algoritması Şekil 21'deki gibi şematize edilebilir:



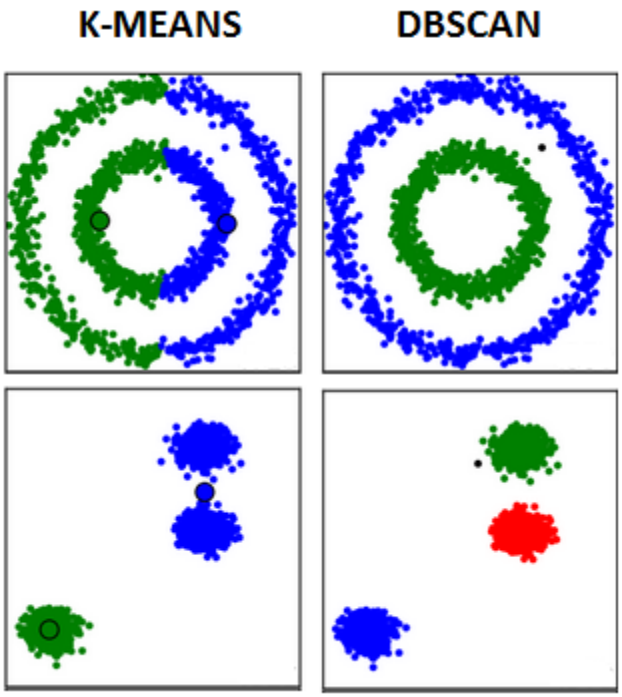
Şekil 21. CHAMELEON Algoritması

(Kaynak: Rafsanjani, M. K., Varzaneh, Z. A. ve Chukanlo, N. E. (2012). A Survey Hierarchical Clustering Algorithms. *The Journal of Mathematics and Computer Science, 5(3),* 229-240.)

4.2.3.Yoğunluğa Dayali Kümeleme Analizi

Günümüzde yüksek hacimli veri tabanlarıyla karşılaşılmaktadır ve uzaklık ölçümleriyle yapılan kümeleme analizleri bu tip veri tabanlarında her zaman istenilen sonucu vermemektedir. Verilerin dağılımlarının karmaşık şekillerde olması matematiksel modellemeler için zorluk teşkil etmektedir. Ayrıca büyük veri tabanlarında daha çok gürültülü verilerle karşılaşılması muhtemeldir. Gürültülü verileri sistemden çıkartarak sonuç bulmaya çalışmak, veri kaybına yol açabileceğinden her zaman uygun olmayacaktır.

Yoğunluğa dayalı kümeleme teknikleri daha önce bahsedilen kümeleme teknikleri gibi çalışmamaktadır. Bu tekniklerde uzaklığa dayalı küme seçiminden ziyade verilerin yoğunluğuna göre bir kümeleme işlemi yapılmaktadır. Yoğunluğa dayalı kümeleme tekniklerinde, kümeler veri tabanındaki daha yüksek yoğunluklu alanlar olarak tanımlanmaktadır (Ester vd., 1996). Küme yoğunluklarının seyrek olduğu alanlarda ise ya gürültülü veriler ya da küme sınırını oluşturan veriler bulunmaktadır (Jiang vd., 2003).



Şekil 22. K-Means ve DBSCAN Yöntemleri Arasindaki Kümeleme Farki

(Kaynak: http://scikit-learn.org/stable/_images/plot_cluster_comparison_11.png, 14.08.2013)

Şekil 22'de görüldüğü gibi k-ortalamalar yöntemiyle yapılan kümeleme yönteminde veri seti iki parçaya ayrılmışken yoğunluğa dayalı kümeleme algoritmalarından olan DBSCAN yöntemiyle veriler yoğunluklarına göre ayrılmıştır. Uzaklığa dayalı kümeleme yöntemlerinde Şekil 22'deki gibi bir ayrım yapılması mümkün değildir. Bu bölümde en çok kullanılan yoğunluğa dayalı kümeleme

algoritmalarından olan DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ve OPTICS (Ordering Points to Identify the Clustering Structure) algoritmalarından bahsedilecektir.

4.2.3.1.*Dbscan Algoritması*

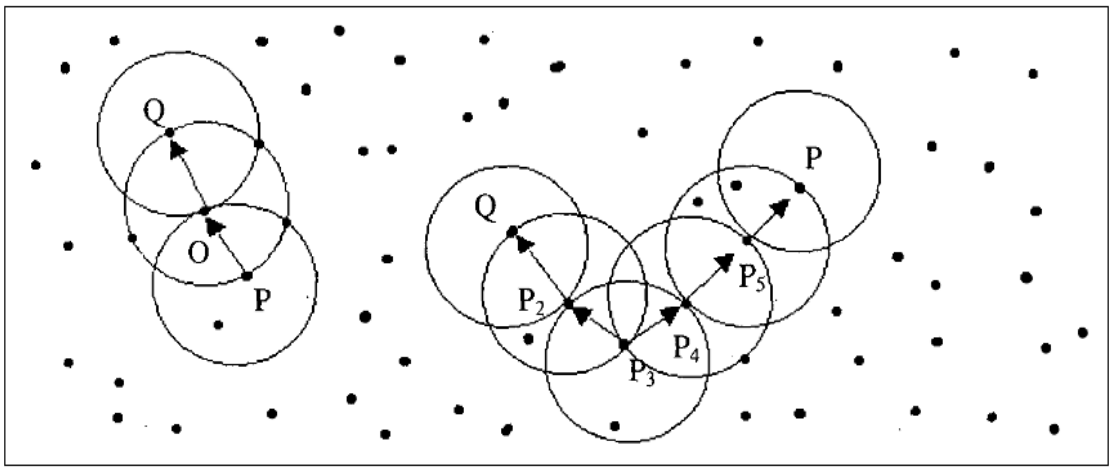
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algoritması kümeler oluşturulurken veri yoğunluklarını dikkate almaktadır (Duan vd., 2007). Bu algoritmayı büyük veri tabanları ve gürültülü verisi çok olan yapılarda kullanmak oldukça uygundur(Ester vd., 1996; Kriegel vd., 2011). Farklı büyük ve şekillerdeki kümelerin oluşturulmasında da bu algoritma kullanılabilecek bir yöntemdir.

DBSCAN algoritması iki istatistik doğrultusunda hareket etmektedir. Bunlar, kümede yer alan her bir nesnenin yarıçap komşuluğu olan *Eps*ve küme çevresindeki en az sayıdaki elemanı gösteren *MinPts* istatistikleridir. Algoritmanın çalışmasını daha iyi anlamak amacıyla algoritmanın bazı özelliklerinin belirtilmesi gerekmektedir (Ye vd., 2003).

- Herhangi bir p noktasının Eps değeri MinPts değerinden daha fazla veri içeriyorsa kümenin yeni merkezi (core object) bu p değeri olacaktır.
- Eğer bir veri merkez nokta değilse sınır değer (border object) olacaktır. Bu sınır değer başka bir merkez noktası için yoğunluğa katılabilir (density reachable) niteliktedir.
- Bir p noktası q noktasının yoğunluğuna katılabilirse; bu durum p_1=q ve p_n=p koşullarında ve p_1,..,p_n nesnelerinde zincirleme ise p_{i+1} nesnesi de yoğunluğa katılmış olacaktır (Borah ve Bhattacharyya, 2004).
- Eğer yukarıdaki maddelere uygun herhangi bir veri yoksa; yani kendi başına kalmış bir veri söz konusuysa, bu veri gürültü olarak nitelendirilmektedir

DBSCAN algoritmasının adımları şu şekildedir(Borah ve Bhattacharyya, 2004):

1. Veri setindeki her bir verinin Eps değerini bul.
2. 1.adımda MinPts değerinden daha büyük bir Eps değerine sahip veri varsa bunu küme merkezi olarak kabul et.
3. Kümeyi, küme merkezlerinden doğrudan yoğunluğa katılacak verilerle genişlet.
4. Herhangi bir kümeye atanacak veri kalmadığında algoritmayı sonlandır.
5. Algoritma sonlanmasına rağmen atanmamış veri kalmışsa bunları gürültü olarak nitele.



Şekil 23. DBSCAN Algoritmasinda Yoğunluğa Katılma (Reachable) Ve Bağlanma (Connective)

(Kaynak: Borah, B. ve Bhattacharyya, D. K. (2004). *An Improved Sampling-Based DBSCAN for Large Spatial Databases.* 8th International Conference on Spoken Language Processing, Jeju Island, South Korea, 4-8 October.)

DBSCAN algoritmasındaki katılma ve bağlanma kavramlarının gösterimi Şekil 23'te gösterilmektedir.

Yöntemin bir eksikliği olarak; DBSCAN algoritmasında Eps ve MinPts değerleri çok boyutlu veri tabanlarında bulunması kolay olmayan istatistik değerlerdir, söylenebilmektedir. Dolayısıyla bu durum sonuçların başarısını da etkilemektedir,

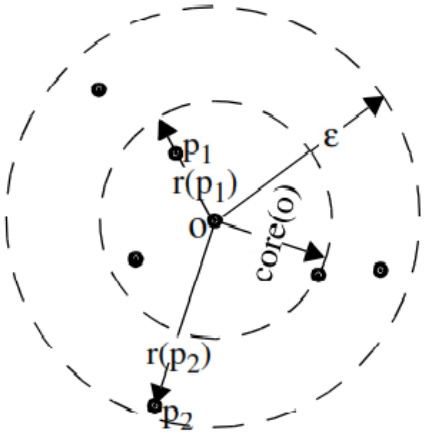
çünkü veriler arasındaki uzaklıklar belirlenirken öklidyen uzaklık hesaplaması yapılmaktadır (Sander, 1998, s.150).

4.2.3.2.*Optics Algoritması*

Optics (Ordering Points to Identfy the Clustering Structure) algoritması, DBSCAN algoritmasındaki girdi parametresi güçlünü ortadan kaldırmak amacıyla Ankerst, Breunig, Kriegel ve Sander tarafından geliştirilmiştir ve 1999 yılındaki uluslararası SIGMOD konferansında *"OPTICS: Ordering Points To Identify the Clustering Structure"* makalesiyle sunulmuştur.

DBSCAN algoritmasının kümeye ve gürültüye dayalı yapısının aksine OPTICS algoritmasında direkt olarak yoğunluk temellidir (Omrani vd., 2011). Bu algoritmada başlangıç olarak her bir nokta merkez noktası olarak kabul edilir (Ankerst vd., 1999). Bu sebeple yeni iki kavram ortaya çıkmıştır. Bu kavramlardan ilki merkez uzaklığı (core distance) ve diğeri ulaşılabilir uzaklık (reachability distance) kavramıdır(Berkhin, 2006).

- D olarak tanımlanmış bir veri tabanında p nesnesi bir merkez noktası ise bu merkez noktasının uzaklığı Eps olacaktır. Eğer p, merkez nokta değilse merkez uzaklık değeri hesaplanmamaktadır.
- D olarak tanımlanmış bir veri tabanında p ve q değerlerinin birer veriyi temsil ettiği varsayılsın. Bu veriler arasındaki uzaklık Öklid uzaklığı olarak hesaplanır ve eğer q verisi bir merkez nokta değilse p ve q verileri arasındaki ulaşılabilir uzaklık hesaplanmamaktadır.



Şekil 24. OPTICS Algoritmasında Merkez Uzaklığı ve Ulaşılabilir Uzaklık

(Kaynak: Ankerst, M., Breunig, M. M., Kriegel H. P. ve Sander, J. (1999). *OPTICS: Ordering Points to Identify the Clustering Structure.* ACM SIGMOD'99 International Conference on Management of Data, New York, USA, 1-3 June.)

Şekil 24'te de görüldüğü gibi merkez kabul edilen bir o noktasının Eps değeri core(o) olarak gösterilmiştir. Aynı şekilde merkez nokta ile p_1 noktası arasındaki uzaklık da core(o) değerine eşittir, çünkü o noktasının Eps değerinden daha yakın bir durumdadır. Merkez nokta ile p_2 noktası arasındaki uzaklık da öklidyen uzaklık ile ölçülür ki, bunun da sebebi bu iki nokta arasındaki uzaklığın o noktasının Eps değerinden büyük olmasıdır. Ayrıca yukarıdaki maddelerde bahsi geçtiği gibi merkez olmayan p_1 ve p_2 noktalarından biri merkez nokta olmadığıı için aralarındaki uzaklıklar hesaplanmaz.

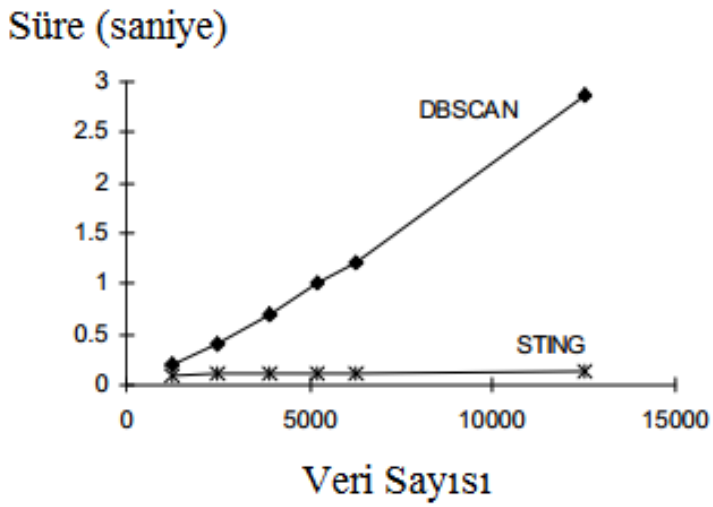
4.2.4.Izgara Tabanlı Kümeleme Teknikleri

Izgara tabanlı kümeleme tekniklerinde, veri tabanı ızgara şeklindeki kümelere bölümlenir. Bu bölümlere ait olan veriler incelenerek bölümlerin ayrıştırılması veya birleştirilmesi yapılarak analiz sonucundaki kümelere ulaşılır (Hinneburg ve Keim, 1999).ızgara tabanlı yöntemler kümeleme analizleri için çok boyutlu veri tabanlarının

analizinde kullanılabilmektedir. Ayrıca ızgara yapısı sayesinde analizi yorumlayacak kişiler için çok çözümlü bir yapı sunmaktadır.

Daha önceki konularda da bahsedildiği gibi kümeleme analizi yaparken problem sadece veri sayısının fazlalığı değil aynı zamanda veri tabanındaki boyut da önemlidir. Izgara tabanlı yöntemler çok boyutlu çözümlemelerde kullanılmak için uygundur (Pilevar ve Sukumar, 2004; Lu vd., 2005). Analizde kullanılan ızgaraların büyüklüklerinin ve sayısının önceden belirlenmesi bu tekniklerdeki zayıflık olarak öne çıkmaktadır (Akın, 2008, s.116).

Izgara tabanlı algoritmalar büyük veri sayısına sahip veri tabanlarında kullanıldığından işlem hızı da önemli bir etkendir. Klasik kümeleme analizlerine oranla dahi iyi bir çözümleme hızı olan yoğunluk temellik kümeleme algoritmaları henüz ızgara tabanlı sistemlere yetişememektedir (Wang vd.,1997). Ancak ızgara tabanlı algoritmaların hızı da hücre sayısının fazlalığına göre artmakta ya da azalmaktadır (Mave Chow, 2004). Analizin başında bölümlendirmenin fazla olması da katmanlar arttıkça hücre sayılarının üssel olarak artmasına, dolayısıyla sürecin yavaşlamasına neden olmaktadır (Wang vd., 1997). Bu durum şekil 25'teki yoğunluk temelli algoritmalardan olan DBSCAN ve ızgara tabanlı algoritmalardan olan STING arasındaki karşılaştırma sonucunda gösterilmektedir.



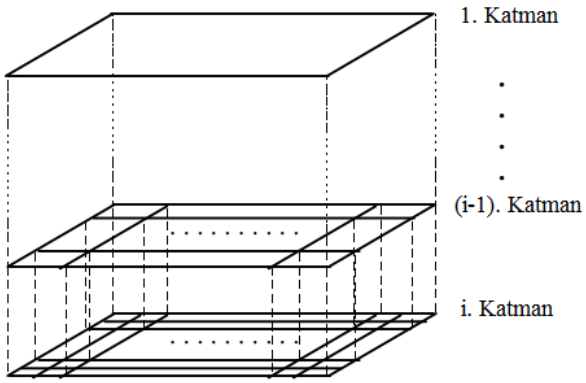
Şekil 25. DBSCAN ve STING Algoritmaları Analiz Hızı Sonuçları

(Kaynak: Wang, W., Yang, J. ve Muntz, R. (1997). *STING: A Statistical Information Grid Approach to Spatial Data Mining*. 23th Very Large Database Conference, Athens, Greece, 20 February.)

Bu başlık altında ızgara tabanlı tekniklerden olan STING algoritması ve WaveCluster algoritmaları anlatılacaktır.

4.2.4.1.*Sting Algoritmasi*

STING algoritması veri tabanını, boyutsal olarak ele alarak, dikdörtgenlere böler ve analizi hiyerarşik bir yapıyla ele alır. (Wang vd.,1997). Hiyerarşik yapı ızgara sistemi katmanlarının her birinde bölümlendirmenin olmasını sağlar. Yapı en yukarıda tüm verileri içeren tek bir hücre şeklindeyken verinin yapısına göre aşağı katmanlara inildikçe bölümlendirilir.



Şekil 26. STING Algoritmasında Hiyerarşik Yapı

(Kaynak: http://lh6.ggpht.com/_e7UyIXjsjN8/SbU6lyv5LII/AAAAAAAADUs/UQdsMwEIr6c/s800/sting.JPG, 20.09.2013)

Şekil 26, İki boyutlu STING algoritmasındaki hiyerarşik yapıyı göstermektedir. Şekil 26'dan da anlaşılabileceği gibi katmanlar arttıkça bölümlendirme de artmaktadır. Hücreler sahip oldukları verilere göre farklı dağılımlar gösterebilirler. Bu dağılımlara göre diğer hücrelerle birleşmek ya da hücrelerin bölümlendirilmesi mümkün olmaktadır. STING algoritmasındaki hücrelerin istatistikleri şu şekilde olmaktadır (Wang vd.,1997):

- n: Hücredeki veri sayısı
- m: Hücredeki verilerin ortalama değeri
- s: Hücredeki verilerin standart sapması
- min: Hücredeki verilerin en küçük değeri
- max: Hücredeki verilerin en büyük değeri
- dağılım: Hücredeki verilerin dağılım tipi

Verilerin dağılımının bilinmesi alt katmanlardaki hesaplamaların yapılma hızını etkilemektedir. STING algoritmasının adımları şu şekildedir(Wang vd., 1997):

1. Katman sayısı belirlenir.

2. Her hücre için güven aralığı ya da tahmini aralık olasılıkları belirlenir.
3. 2.Adımda hesaplanan aralıklarla ilgili hücrenin geçerli olması ya da geçerli olmaması bilgisi alınır.
4. Eğer işlem yapılan katman, alt katman olarak kabul edilirse 6.adıma geçilir, aksi halde 5.Adımdan devam edilir.
5. Katman bir basamak arttırılır ve hücreler daha yüksek bir katmanla ilişkilendirilebilirse 2.adıma dönülür.
6. Hücre bilgisiyle araştırılan konu örtüşüyorsa 8.adıma gidilir, aksi halde 7.adıma devam edilir.
7. Araştırılan konu hakkında daha fazla işlem gerekiyorsa bölümlendirme devam ettirilir, aksi halde 9.adıma gidilir.
8. Hücreler tamamen tekil hale gelene kadar 2.adımdan itibaren algoritma tekrarlanır.
9. Herhangi bir gereksinim kalmamışsa algoritma sonlandırılır.

Hücre yapılarının dikdörtgen şeklinde olması dairesel yapıdaki verilerde bazı verilerin hücre dışında kalmasına sebep olabilir. Bu durum analizin kalitesini etkilemektedir.

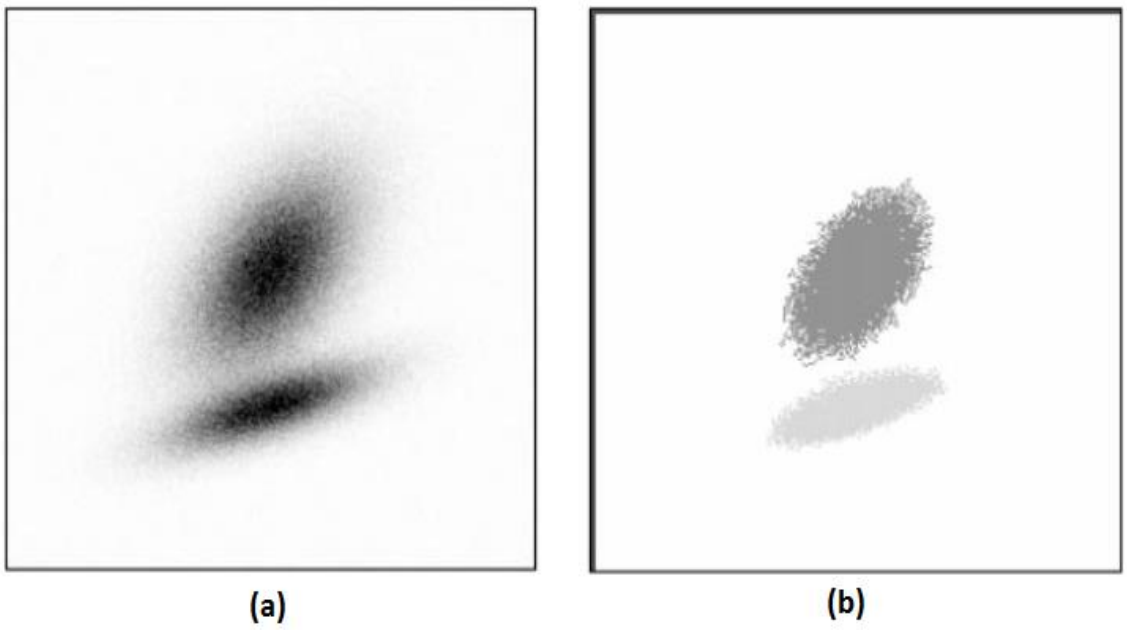
4.2.4.2.*Wavecluster Algoritması*

WaveCluster algoritması G. Sheikholeslami, S. Chatterjee ve A. Zhang tarafından geliştirilmiş olup, 1998 yılında VLDB konferansında "WaveCluster: a Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases" başlıklı çalışmayla yayınlanmıştır. Bu algoritmanın amacı, nitelik uzayına, yani veri tabanına, dalga dönüşümleri uygulamak ve yoğunluğa göre oluşan bölgeleri, yani kümeleri, ortaya çıkartmaktır (Li ve Luo, 2009). Dalga dönüşümü, veri tabanını farklı seviyelerde bölümlemek ve ayırt ediciliği yüksek olan kümeler bulmaktır. Algoritmanın bu yapısı

sebebiyle, kümeleme çalışması olarak algoritmanın hem yoğunluğa dayalı hem de ızgara tabanlı olduğu söylenebilmektedir.

WaveCLuster algoritması şu adımlardan oluşmaktadır (Sheikholeslami vd.,2000):

1. Veri tabanını nicelendir ve nesneleri temsili olarak hücrelere ata
2. Nicelendirilmiş veri tabanına dalga dönüşümü uygula
3. Farklı düzeyler için dönüştürülmüş veri tabanındaki bağlanabilecek kümeleri bul
4. Hücre adlarını belirle
5. Seçim kriteri oluştur
6. Verileri nihai kümelere ata



Şekil 27. Orjinal ve Dalga Dönüşümü Uygulanmış Veri Tabanlarının Grafiksel Gösterimi

(Kaynak: Sheikholeslami, G., Chatterjee, S. ve Zhang, A. (2000). WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases. *The Very Large Data Bases Journal*, *8*, 289-304.)

Şekil 27 (a)'da orijinal verinin grafiksel gösterimi ve Şekil 27 (b)'de bu veri setine dalga dönüşümü uygulanmış olan küme ayrımının grafiksel gösterimi verilmiştir.

Algoritmanın yapısı gereği şu faydalar ortaya çıkmaktadır (Li ve Luo, 2009; Yıldırım ve Özdoğan, 2011):

1. Veriler üzerinde dalga dönüşümü uygulanmaktadır. Bu sebeple sisteme giren verinin sırası önemli olmayacaktır. Ayrıca oluşan hücrelerin veri büyüklükleri de bu algoritma için önemsiz olmaktadır.
2. Algoritma veri tabanındaki verilerin tamamını en az bir defa okumaktadır. Dolayısıyla hesaplamalar diğer algoritmalara göre daha hızlı sonuçlanmaktadır. Özellikle az boyutlu veri tabanlarında bu algoritma oldukça etkilidir.
3. Aynı anda çoklu ölçekli veri tabanlarında kümeleme yapılabilir. Dolayısıyla algoritma çoklu çözüm imkanı sunar ve çalışmacının isteğine göre küme seçimi yapılabilir.
4. Küme birleştirilirken esas olarak küme özellikleri dikkate alınır. Bu durumda birleştirilecek kümeler arasındaki farklar önemsizdir. Ayrıca kümelerin şekillerinin de herhangi bir etkisi olmamaktadır.
5. Dalga dönüşümü yoğunluk farklarını ortaya çıkartacağından düşük yoğunluğa sahip olan gürültülü ve aykırı veriler sistemden çıkartılacaktır. Buradan da anlaşılacağı üzere gürültülü ve aykırı verilerin analiz üstüne bir etkisi olmayacaktır.
6. Küme sayısının analizin başında verilmesi gerekmemektedir. Algoritmanın sonucunda olası doğal kümeler kendiliğinden oluşacaktır (Akın, 2008, s.120)

5.UYGULAMA

Bu çalışmanın amacı, üçüncü basamak hizmet veren bir üniversite hastanesine sağlık hizmeti alan hastaların hastaneye başvurma ve bulunma durumlarına göre, hasta davranışlarını belirlemektir. Sağlık uygulamalarında hastalar, en zararsız kabul edilecek hastalıktan, iyileşme süresi en fazla olan hastalıklara kadar birçok başvuru ile hastanelere gitmekte ve aynı şekilde birçok tedavi ile karşılaşabilmektedir. Bu sebepten hasta, hastalık ve tedavi türleri oldukça fazla olmaktadır. Veri sayısının ve tipinin fazla olduğu analizlerde veri madenciliği yöntemlerinin kullanmak diğer klasik istatistik yöntemlerine göre daha uygun olacağı düşünülmektedir. Bu amaçla çalışmada veri madenciliği yöntemleri içerisinde kullanılmakta olan kümeleme analizi uygulanmıştır. Tabii ki kümeleme analizi içerdiği yöntemler açısından farklılıklar göstermektedir. Klasik kümeleme analizlerinden kabul edilen K-Means kümeleme ve (bilişim teknolojisinin gelişmesiyle birlikte ortaya çıkartılan) Yoğunluk Tabanlı Kümeleme Analizi teknikleri çalışmanın gerçekleştirilmesi amacıyla seçilmiş ve uygulanmıştır.

Çalışmada kullanılan veri seti 2011 yılında hastaneye başvurmuş olan hastaların hastaneye her bir başvurularının verisini içermektedir. Veri setinde kullanılan parametreler, hastaların ayaktan tedavi olması göz önüne alınarak poliklinik sayısı veya yatarak tedavi gören hastaların kaç defa yatarak tedavi aldığı ve bu yatışlarındaki toplam hastanede kalma süreleri alınmıştır. Hastaların yatarak tedavi olması gerekiyorsa bunun sebebi, hastayı gözetim altına almak mı? Cerrahi müdahalede bulunmak mı? soruları akla gelmektedir. Bu sebeple hastaların oldukları ameliyatların sayıları da çalışmaya dahil edilmiştir. Ayrıca hastalıklarda tedavinin süreci ve şiddetinde önemli faktörler olarak düşünülen yaş ve cinsiyet parametreleri de alınmıştır. Genel olarak hastaneye yapılan başvuruların dar bir aralıkta olup olmadığını anlamak amacıyla hekimler tarafından hastalara verilen teşhisler ve hastaların kaç farklı servis başvurduğu, çalışmada hasta davranışlarını ve üçüncü basamak bir hastaneyi incelemek için faydalı olacağı düşünülmüştür.

Hastaneler yapısı gereği birçok kamu kurumundan ve özel sektördeki şirketlerden ayrılmaktadır. Hastanelerin bulunduğu çıkmaz, hastalara hizmet vermek ve aynı zamanda işletmecilik adına devamlılığını sağlayabilmektir. Buradan yola çıkarak hastaneye başvuran hastaların sağladıkları maddi meblağ da veri setine katılmıştır. Çalışma Cumhuriyet Üniversitesi Hastanesi'ne başvuran hastaları içerdiğinden hastaların Sivas ilinden gelip gelmemesi de önemli bir değişkendir. Ayrıca hasta yapılarını belirlemek amacıyla hangi ilden gelirse gelsin hastaların il merkezinden mi? merkez dışındaki bölgelerden mi? geldikleri de bu çalışmada incelenmiştir.

Bahsi geçen parametreler toplanırken yaş unsurunun öne çıkarttığı bazı sorunlar olmuştur. Öncelikle, hastanedeki tedavi birimleri 18 yaş altı ve üstü olarak ayrılmaktadır. Çocuk hastalıklarına başvuran hastalara uygulanan işlemler, dolayısıyla tedavi süresi ve tipi farklı olacaktır. Ayrıca ileri yaşlarında bulunan hastalarındaki hastaneye başvuru miktarları, ameliyat sayıları ve yatarak tedavi edilen gün sayıları, bu parametrelerdeki ortalamayı yukarı çekmektedir. Bu sebeple veri seti yasalar tarafında yetişkin sayılan 18 yaş ile emeklilik sınırı olan 65 yaş arasında bulunan hastalara indirgenmiştir. Böylelikle veri seti 78.239 hastanın veri tabanından alınan verileri ile oluşturulmuştur. Buradan da 18 yaşındaki bir hastanın önceki yıllarda yaptığı hastane başvurularının nasıl kullanılacağı sorunu ortaya çıkmaktadır. Veri tabanından veriler alınırken bu sorunla karşılaşılan hastaların 18 yaş altındaki başvuruları veri setine dahil edilmemiştir.

Çalışmada kullanılan 7 parametre ve oluşacak kümelerin farklılıklarını analiz edebilmek amacıyla hastaların bilgileri Tablo 8'de verilmiştir.

Tablo 8. Çalışmada Kullanilan Değişkenler

	Değişken Adı	**Değişken Tipi**	**Değişken Aralığı**	**Değişken Kodu**
1	Yaş	Sayısal Değer	18-65	-
2	Cinsiyet	Metinsel Değer	-	E-K
3	Ayaktan Tedavi Sayısı	Sayısal Değer	1-230	-
4	Yatarak Tedavi Sayısı	Sayısal Değer	1-96	-
5	Yatılan Gün Sayısı	Sayısal Değer	0-1161	-
6	Ameliyat Sayısı	Sayısal Değer	0-86	-
7	Verilen Farklı Teşhis Sayısı	Sayısal Değer	1-72	-
8	Başvurulan Farklı Servis Sayısı	Sayısal Değer	1-29	-
9	Ücret	Sayısal Değer	1-428.682 Lira	-
10	İl Merkezi-Merkez Dışı Bölgeler	Metinsel Değer	-	1-2
11	Sivas İlinden Gelen-Sivas İli Dışından Gelen	Metinsel Değer	-	1-2

Çalışmada yer alan hastaların geçmiş yıllardaki tedavilerinden doğan ücretler Lira olarak baştan düzenlenerek veri setine dahil edilmiştir.

5.1.UYGULAMADA KULLANILAN YÖNTEMLER

Veri tabanından süzülen veriler uygulamaya hazır hale getirildikten sonra oluşan 78.239 hasta ve 11 parametre kullanılarak öncelikle veri madenciliği kümeleme analizinde sıkça kullanılan yöntemlerden olan K-Means kümeleme yöntemi kullanılmıştır. Kümeleme tekniklerinin veri tipine göre yeterlilikleri göz önüne alınarak bir diğer kümeleme tekniği olan Yoğunluk Tabanlı Kümeleme tekniği uygulanmıştır.

Veri seti farklı iki teknikle kümelere ayrıldıktan sonra kümeler arasındaki farklar hastaların demografik özelliklerine göre sınanmak istenmiştir. Ancak öncelikle karşılaştırma analizinin belirlenebilmesi için verilerin normallik sınamasını geçmesi gerekmektedir. Bu amaçla öncelikle verileri Kolmogorov-Smirnov Z testi uygulanmıştır. Verilerin normal dağılıma uygun olmaması sebebiyle kümelerin dağılımlarının farklılığını ölçmek amacıyla Kruskal-Wallis H testi ve küme ortalamalarının farklılığını test edebilmek amacıyla da Mann-Whitney U testi kullanılmıştır.

Kategorik olan değişkenlerin kümelere göre farklılığını ölçmek amacıyla Ki-Kare testi kullanılmıştır. Ki-Kare testi sonucunda farklı olan kümelerde test yapılan parametreler ile arasındaki ilişkinin gücü Spearman’s rho değeri ile test edilmiş olup sonuçlar ortaya konulmuştur.

Çalışmada kullanılan analizler SPSS 14.0 ve WEKA (http://www.cs.waikato.ac.nz/ml/weka/downloading.html) paket programları aracılığıyla yapılmış olup sonuçlar aşağıda belirtilmiştir.

5.2.K-Means Kümeleme

K-Means yöntemi, veri madenciliği yöntemleri arasında temel olabilecek ve en sık kullanılan yöntemlerden birisidir. Konveks şekillere sahip olmayan verilerde bu analizin uygun sonuçlar vermeyebileceği görülmüştür. Ayrıca sayısal olmayan veri setlerinde de uygulama öklidyen uzaklıkların hesaplanmasından çok parametrelerin mod değerinin hesaplanması yada benzerlik katsayıların hesaplanmasıyla gerçekleştirilebilmektedir. Ancak yapılan çalışmada küme sayısındaki değişime göre küme içi hata kareler toplamı ve benzemezlik katsayısı hesaplanarak Tablo 9'da verilmiştir.

Tablo 9.Küme Sayısına Göre Benzemezlik ve Hata Değerleri

Küme Sayısı (k)	Benzemezlik Değeri	Küme İçi Hata Kareler Toplamı
2	31,32	3505,35
3	31,14	2621,34
4	29,62	1965,51
5	29,64	1668,81
6	29,01	1392,43
7	29,03	1267,83
8	28,60	1150,24
9	28,38	1017,89
10	28,38	960,21

Tablo 9'dan da anlaşılabileceği gibi küme sayısı arttıkça benzemezlik değeri düşmekte ve benzer şekilde küme içi hata kareler değeri de düşmektedir. Ancak küme sayısının artması birçok parametre için ayırt ediciliğin azalarak yok olmasına neden

olmaktadır. Bu sebepten çalışmada küme sayısı üç olarak belirlenmiştir. Küme sayısının üç olduğu çözümde benzemezlik değeri 31,14 ve küme içi hata kareler toplamı da 2621,34 olarak hesaplanmıştır. Analiz sonucunda elde edilen parametrelere göre küme merkezleri ve standart sapma değerleri tablo 10'da gösterilmiştir.

Tablo 10.Küme Merkezleri ve Standart Sapmalari

Parametre	1.Küme		2.Küme		3.Küme	
	Merkez	Standart Sapma	Merkez	Standart Sapma	Merkez	Standart Sapma
Yaş	39,8697	±4,6706	**56,3847**	±5,0369	24,773	±4,0958
Poliklinik Tedavi	10,0018	±13,3344	**12,8086**	±16,6641	6,2562	±7,8753
Yatarak Tedavi	0,9148	±1,9614	**1,4784**	±2,6237	0,5431	±1,4597
Farklı Servis	4,1764	±3,3564	**4,8325**	±3,7842	3,2019	±2,4706
Ameliyat	0,857	±2,3521	**1,2546**	±3,4348	0,5377	±1,7343
Farklı Teşhis	5,7268	±6,1014	**6,9354**	±6,9826	4,0999	±3,9816
Yatılan Gün Sayısı	6,3585	±19,817	**12,5638**	±30,5707	3,0537	±12,404
Tedavi Ücreti	1939,1885	±5339,3704	**3680,5836**	±8435,9653	927,9309	±2757,257

Tablo 10'da görüldüğü gibi ikinci kümedeki parametre merkezleri ve standart sapma değerleri diğer küme istatistiklerinden daha yüksektir. Buradan da anlaşılabileceği gibi ikinci kümenin toplam değişkenliği daha fazladır. Bu durum yanıltıcı olarak görülebilir. Çünkü değişkenliğin daha büyük olması daha fazla veri

içerebildiği anlamına gelmemektedir. Kümelerin veri sayısı ve toplam veri setindeki oranları Tablo 11'de gösterilmiştir.

Tablo 11.Kümelere Göre Gözlem Sayısı Ve Oranları

Kümeler	Veri Sayısı	Oran
1	23044	%29
2	20685	%26
3	34510	%44

2. kümeye atanmış olan hastaların yaşlarının diğer kümedekilere göre daha büyük olması ve diğer parametrelerde de (yatış süresi, tedavi ücreti gibi) bu artışın olması, 2. kümenin sağlık hizmetini daha fazla alan grup olarak sınıflandırılabileceğini göstermektedir. Belirlenmiş küme aidiyetliklerine göre küme ortalaması ve standart sapma değerleri Tablo 12'de verilmiştir.

Tablo 12.Parametrelere Göre Küme Ortalamaları Ve Standart Sapmaları

Parametre	1.Küme		2.Küme		3.Küme	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma	Ortalama	Standart Sapma
Yaş	41,51	±6,112	**56,47**	±6,317	25,98	5,035±
Poliklinik Tedavi	10,23	±11,681	**15,32**	±19,356	5,10	±4,992
Yatarak Tedavi	0,84	±1,365	**1,92**	±3,364	0,39	±0,763
Farklı Servis	4,35	±3,339	**5,30**	±4,115	2,93	±2,022
Ameliyat	0,79	±1,693	**1,68**	±4,303	0,38	±0,964
Farklı Teşhis	5,97	±5,725	**7,94**	±7,956	3,55	±2,897
Yatılan Gün Sayısı	5,42	±11,712	**17,11**	±37,924	1,63	±4,430
Tedavi Ücreti	1.676,69	±2.439,622	**4.843,23**	±10.414,058	601,69	±888,951

Tablo 12'de de görüldüğü gibi parametrelerin ortalama değerleri, kümelerin isimlendirilmesi dolayısıyla hasta yapısı hakkında fikir vermesi bağlamında araştırmacıya yardımcı olmaktadır. Şöyle ki; sağlık uygulamalarından faydalanma ve sonucunda yapılan işlemlerin tutarlarında bütün olarak bir değişim gözükmektedir. 3.küme daha az sağlık hizmetinden faydalanmış dolayısıyla daha az sağlık ödemesi olmuştur. Aynı şekilde 2.küme daha fazla sağlık hizmetinden faydalanmış ve daha yüksek miktarlarda sağlık ödemesi olduğu sonucu ortaya çıkmıştır.

Kümeleme analizi sonucunda oluşan kümelerin tutarlılığını test etmek amacıyla kümeler arasındaki farklılıklar araştırılmıştır. Bu farklılık araştırmasını yapmadan önce verilerin normallik sınamasının yapılması gerekmektedir. Bu amaçla yapılan Kolmogorov-Smirnov Z uyum iyiliği testi sonucu Tablo 13'te verilmiştir.

Tablo 13.Normallik Sınaması Sonuçları

Parametre	1.Küme		2.Küme		3.Küme	
	K-S Değeri	Önemlilik Seviyesi	K-S Değeri	Önemlilik Seviyesi	K-S Değeri	Önemlilik Seviyesi
Yaş	10,867	0,000	14,115	0,000	20,075	0,000
Poliklinik Tedavi	32,588	0,000	33,035	0,000	38,186	0,000
Yatarak Tedavi	44,392	0,000	40,819	0,000	78,920	0,000
Farklı Servis	27,183	0,000	23,040	0,000	39,318	0,000
Ameliyat	54,327	0,000	50,108	0,000	82,470	0,000
Farklı Teşhis	29,262	0,000	27,526	0,000	37,607	0,000
Yatılan Gün Sayısı	48,863	0,000	46,878	0,000	70,304	0,000
Tedavi Ücreti	37,467	0,000	46,184	0,000	46,923	0,000

Tablo 13'e göre kümelere ayrılmış olana parametrelerin 0,05 ve 0,01 anlamlılık düzeylerin de istatistiksel olarak normal dağılıma uymadığı olmadığı belirlenmiştir. Kümelerin tutarlılığını ölçmek amacıyla parametrik olmayan testlerin uygulanması gerekmektedir. Bu amaçla parametrelerin dağılımlarının farklılığı için Kruskal-Wallis H testi ve ortalamasın farkı içinde Mann-Whitney U testi uygulanması gerekmektedir.

Tablo 14.Kruskal-Wallis Testi Istatistikleri Ve Önemlilik Düzeyleri

Parametreler	K-W İstatistiği	P-Değeri
Yaş	62948,699	0,000
Poliklinik Tedavi	6375,304	0,000
Yatarak Tedavi	7806,922	0,000
Farklı Servis	5541,348	0,000
Ameliyat	2882,964	0,000
Farklı Teşhis	6212,472	0,000
Yatılan Gün Sayısı	8590,327	0,000
Tedavi Ücreti	12466,137	0,000

Tablo 14'te Kruskal-Wallis H testi değeri ve p değerleri verilmiştir. Bu analizin sonucuna göre veri seti üç kümeye ayrıldıktan sonra bölümlendirmede parametrelerin dağılımları birbirleri ile benzeşmemektedir. Yani parametrelerin aynı oldukları hipotezi 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak reddedilmektedir. Kümeler ilgili tüm başlıklarda farklıdır denilebilmektedir. Dağılımların farklılığı gibi ortalamalarında farklılığı test edilmelidir. Bu amaçla yapılan Mann-Whitney U testi sonuçları aşağıdaki tabloda yer almaktadır.

Tablo 15.Mann-Whitney U Testi Istatistikleri Ve Önemlilik Düzeyleri

Parametre	Küme		M-W İstatistiği	P-Değeri
Yaş	1.Küme	2.Küme	2E+007	0,000
		3.Küme	3E+007	0,000
	2.Küme	3.Küme	3.358.927	0,000
Poliklinik Tedavi	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Yatarak Tedavi	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Farklı Servis	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Ameliyat	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Farklı Teşhis	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Yatılan Gün Sayısı	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000
Tedavi Ücreti	1.Küme	2.Küme	2E+008	0,000
		3.Küme	3E+008	0,000
	2.Küme	3.Küme	2E+008	0,000

Tablo 15'da görülen Mann-Whitney U testi sonuçları, kümelere ayrılmış olan parametrelerin ortalamasının birbirlerinden önemli ölçüde farklı olduğunu göstermektedir. Test sonuçlarına göre 0,05 ve 0,01 anlamlılık düzeylerinde parametrelerin farklı olduğu söylenebilmektedir.

Çalışmanın Cumhuriyet Üniversitesi Hastanesine gelen hastaları içermesi sebebiyle hastaların cinsiyeti, il merkezinde veya il merkezi dışında olması, ayrıca Sivas ilinden gelip gelmemesi durumlarının da çalışma için önemli olduğu düşünülmektedir. Hastaların belirtilen özelliklerine göre kümelerde nasıl bir yapıda olduğu Ki-Kare bağımsızlık testi ile incelenmiştir.

Tablo 16.Kümelere Göre Cinsiyet Sayıları ve Oranları

Parametreler			Kümeler			Toplam
			1	2	3	
Cinsiyet	Kadın	n	12.386	11.112	19.050	42.548
		Cinsiyete göre %	%29,1	%26,1	%44,8	%100,0
		Kümeye göre %	%53,7	%53,7	%55,2	%54,4
		Toplam %	%15,8	%14,2	%24,3	%54,4
	Erkek	n	10.658	9.573	15.460	35.691
		Cinsiyete göre %	%29,9	%26,8	%43,3	%100,0
		Kümeye göre %	%46,3	%46,3	%44,8	%45,6
		Toplam %	%13,6	%12,2	%19,8	%45,6
Toplam		n	23.044	20.685	34.510	78.239
		Cinsiyete göre %	%29,5	%26,4	%44,1	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0	%100,0
		Toplam %	%29,5	%26,4	%44,1	%100,0

Tablo 17.Cinsiyete Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	16,711	2	0,000
Benzerlik Oranı	16,715	2	0,000
Geçerli Gözlem	78.239		

Çalışılan veri setinde (Tablo 16), kümeler ve cinsiyet arasında 0,05 ve 0,01 anlamlılık düzeyinde istatistiksel olarak farklılık olduğu tespit edilmiştir. Hastaların cinsiyetlerine göre hangi kümelerde olduğunu görmek amacıyla oluşturulan çapraz tabloda (Tablo 17.), oransal olarak birinci ve ikinci kümedeki cinsiyet dağılımının eşit ve üçüncü kümede diğer iki kümeye oranla daha fazla kadın olduğu sonucu görülmektedir. Ayrıca ikinci kümedeki erkek sayısı birinci kümeye göre daha fazla iken kadınlarda bu durum tam tersi olduğu analiz sonucu ortaya çıkmıştır.

Tablo 18.Kümelere Göre Merkez-Merkez Dışı Sayıları Ve Oranları

Parametreler			Kümeler			Toplam
			1	2	3	
İl Merkezi	Merkez	n	18.175	15.728	25.956	59.859
		Merkeze göre %	%30,4	%26,3	%43,4	%100,0
		Kümeye göre %	%78,9	%76,0	%75,2	%76,5
		Toplam %	%23,2	%20,1	%33,2	%76,5
	Merkez Dışı	n	4.869	4.957	8.554	18.380
		Merkeze göre %	%26,5	%27,0	%46,5	%100,0
		Kümeye göre %	%21,1	%24,0	%24,8	%45,6
		Toplam %	%6,2	%6,3	%10,9	%45,6
Toplam		n	23.044	20.685	34.510	78.239
		Merkeze göre %	%29,5	%26,4	%44,1	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0	%100,0
		Toplam %	%29,5	%26,4	%44,1	%100,0

Tablo 19. Merkez-Merkez Dışı Olmasına Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	106,349	2	0,000
Benzerlik Oranı	107,704	2	0,000
Geçerli Gözlem	78.239		

Tablo 19 ’ya göre hastaneye başvuran hastaların kümelerdeki dağılımlarına göre 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak anlamlı bir farklılık vardır. Hastaların il merkezlerinden gelip gelmemeleri durumunu gösteren Tablo 18’de, hastaların çoğunluğunun il merkezinden geldiği görülmektedir. Ayrıca merkez dışından gelen hastaların çoğunluğu üçüncü kümede yoğunlaşmıştır.

Tablo 20.Kümelere Göre Sivas-Sivas Dışı Sayıları ve Oranları

Parametreler			Kümeler			Toplam
			1	2	3	
Sivas	Sivas	n	20.780	18.351	28.230	67.361
		İle göre %	%30,8	%27,2	%41,9	%100,0
		Kümeye göre %	%90,2	%88,7	%81,8	%86,1
		Toplam %	%26,6	%23,5	%36,1	%86,1
	Sivas Dışı	N	2.264	2.334	6.280	10.878
		İle göre %	%20,8	%21,5	%57,7	%100,0
		Kümeye göre %	%9,8	%11,3	%18,2	%13,9
		Toplam %	%2,9	%3,0	%8,0	%13,9
Toplam		n	23.044	20.685	34.510	78.239
		Merkeze göre %	%29,5	%26,4	%44,1	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0	%100,0
		Toplam %	%29,5	%26,4	%44,1	%100,0

Tablo 21.Sivas-Sivas Dışı Olmasına Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	970,481	2	0,000
Benzerlik Oranı	968,815	2	0,000
Geçerli Gözlem	78.239		

Tablo 21 ’den de anlaşılacağı gibi, oluşturulan kümelerde hastaların Sivas ili dışından olup olmaması 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak farklıdır. Tablo 20’de ki değerlere göre, hastaların çoğunluğu Sivas ili sınırlarından gelmiştir. Bu oran tüm hastalar içinde %86,1’lik bir kısmı oluşturmaktadır. Sivas ili dışından gelen hastaların çoğunluğu üçüncü kümeye atanmıştır.

5.3.YOĞUNLUĞA DAYALI KÜMELEME

Yoğunluğa dayalı kümeleme algoritmaları, diğer kümeleme analizlerinde olan uzaklığa dayalı çözümleme sistemini kullanmamaktadır. Analizin yapısı verilerin yoğunlukta olduğu bölgeleri belirlemek ve kümeleri bu bölgelere göre tanımlamaktır. Bu algoritmada küme sayısının özellikle belirtilmesine gerek yoktur.

Analiz sonucunda veri seti iki kümeye ayrılmış olup, algoritma işlemi 14 iterasyonda sonlanmıştır. Bu analize göre ortaya çıkan küme merkezleri ve standart sapma değerleri Tablo 22'de verilmektedir.

Tablo 22.Küme Merkezleri Ve Standart Sapmalar

Parametre	1.Küme		2.Küme	
	Merkez	Standart Sapma	Merkez	Standart Sapma
Yaş	27,9897	±6,2844	**52,2188**	±7,2822
Poliklinik Tedavi	7,0813	±9,3058	**12,1754**	±16,0401
Yatarak Tedavi	0,6348	±1,5753	**1,3069**	±2,4741
Farklı Servis	3,4179	±2,6981	**4,691**	±3,718
Ameliyat	0,6109	±1,8911	**1,1436**	±3,1494
Farklı Teşhis	4,4701	±4,5777	**6,6488**	±6,8313
Yatılan Gün Sayısı	3,804	±14,7329	**10,7197**	±27,6846
Tedavi Ücreti	1143,3001	±3.414,3695	**3.187,5895**	±7.733,818

Tablo 22'den de anlaşılacağı gibi küme merkezlerinin tamamı ikinci kümede daha büyük değerler almıştır. Aynı şekilde ikinci kümenin küme merkezlerinin standart sapmaları birinci kümeye göre daha fazladır. Buradan yola çıkarak ikinci kümenin değişkenliğinin daha fazla olduğu söylenebilir. Kümelerin hacimleri ve veri setindeki oranları Tablo 23'te gösterilmiştir.

Tablo 23.Kümelere Göre Gözlem Sayısı ve Oranları

Kümeler	Veri Sayısı	Oran
1	48624	%62,1
2	29615	%37,9

Elde edilen kümelere göre parametrelerin ortalamaları ve standart sapmaları hesaplanmış olup Tablo 24'te verilmiştir.

Tablo 24.Parametrelere Göre Küme Ortalamaları ve Standart Sapmaları

Parametre	1.Küme		2.Küme	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma
Yaş	30,02	±8,007	52,73	±8,730
Poliklinik Tedavi	5,65	±5,795	15,34	±18,163
Yatarak Tedavi	0,44	±0,821	1,73	±2,986
Farklı Servis	3,10	±2,192	5,42	±4,077
Ameliyat	0,42	±1,021	1,54	±3,807
Farklı Teşhis	3,81	±3,232	8,08	±7,727
Yatılan Gün Sayısı	2,06	±5,414	14,68	±33,073
Tedavi Ücreti	724,62	±1.090,744	4.198,90	±8.935,773

Tablo 24 incelendiğinde, ikinci kümedeki parametrelerin tümünde ortalamaların birinci kümeye göre daha büyük olduğu gözükmektedir. K-means uygulamasında olduğu gibi burada da parametrelerdeki değerlerin bir kümede yoğunlaşması söz konusudur. İkinci kümeye sağlık hizmetlerini daha çok kullanan ve birinci kümeyi de daha az kullanan gruplar olarak isimlendirmek mümkündür.

Yapılan yoğunluğa dayalı algoritmada da tutarlılığı test etmek amacıyla parametrelerin farklılıkları araştırılmıştır. Öncelikle parametrelerin normal dağılıma uyup uymadığını araştırmak amacıyla Kolmogorov-Smirnov Z testi sonuçlarına bakmak gerekmektedir.

Tablo 25. Normallik Sınaması Sonuçları

Parametre	1.Küme		3.Küme	
	K-S Değeri	Önemlilik Seviyesi	K-S Değeri	Önemlilik Seviyesi
Yaş	22,743	0,000	16,198	0,000
Poliklinik Tedavi	46,601	0,000	36,989	0,000
Yatarak Tedavi	89,767	0,000	48,309	0,000
Farklı Servis	44,574	0,000	26,494	0,000
Ameliyat	95,950	0,000	59,042	0,000
Farklı Teşhis	44,240	0,000	30,910	0,000
Yatılan Gün Sayısı	79,012	0,000	56,540	0,000
Tedavi Ücreti	56,410	0,000	54,978	0,000

Tablo 25'e göre kümelere ayrılmış olana parametrelerin 0,05 ve 0,01 anlamlılık düzeylerin de istatistiksel olarak normal dağılıma uymadığı olmadığı belirlenmiştir. Kümelerin tutarlılığını ölçmek amacıyla parametrik olmayan testlerin uygulanması gerekmektedir. Bu amaçla parametrelerin dağılımlarının farklılığı için Kruskal-Wallis H testi ve ortalamasın farkı içinde Mann-Whitney U testi uygulanması gerekmektedir.

Tablo 26. Kruskal-Wallis Testi İstatistikleri ve Önemlilik Düzeyleri

Parametreler	K-W İstatistiği	P-Değeri
Yaş	46.940,901	0,000
Poliklinik Tedavi	7.704,216	0,000
Yatarak Tedavi	8.750,327	0,000
Farklı Servis	6.763,425	0,000
Ameliyat	3.437,694	0,000
Farklı Teşhis	7.691,485	0,000
Yatılan Gün Sayısı	9.395,281	0,000
Tedavi Ücreti	13.673,258	0,000

Tablo 26'da Kruskal-Wallis H testi değeri ve p değerleri verilmiştir. Bu analizin sonucuna göre veri seti iki kümeye ayrıldıktan sonra bölümlendirmede parametrelerin dağılımları birbirleri ile benzeşmemektedir. Yani parametrelerin aynı oldukları hipotezi 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak reddedilmektedir. Kümeler ilgili tüm başlıklarda farklıdır denilebilmektedir. Dağılımların farklılığı gibi ortalamalarında farklılığı test edilmelidir. Bu amaçla yapılan Mann-Whitney U testi sonuçları aşağıdaki tabloda yer almaktadır.

Tablo 27.Mann-Whitney U Testi İstatistikleri Ve Önemlilik Düzeyleri

Parametre	M-W İstatistiği	P-Değeri
Yaş	6E+007	0,000
Poliklinik Tedavi	5E+008	0,000
Yatarak Tedavi	5E+008	0,000
Farklı Servis	5E+008	0,000
Ameliyat	6E+008	0,000
Farklı Teşhis	5E+008	0,000
Yatılan Gün Sayısı	5E+008	0,000
Tedavi Ücreti	4E+008	0,000

Tablo 27'de görülen Mann-Whitney U testi sonuçları, kümelere ayrılmış olan parametrelerin ortalamasının birbirlerinden önemli ölçüde farklı olduğunu göstermektedir. Test sonuçlarına göre 0,05 ve 0,01 anlamlılık düzeylerinde parametrelerin farklı olduğu söylenebilmektedir.

Çalışmanın Cumhuriyet Üniversitesi Hastanesine gelen hastaları içermesi sebebiyle hastaların cinsiyeti, il merkezinde veya il merkezi dışında olması, ayrıca Sivas ilinden gelip gelmemesi durumlarının da çalışma için önemli olduğu düşünülmektedir. Hastaların belirtilen özelliklerine göre kümelerde nasıl bir yapıda olduğu Ki-Kare bağımsızlık testi ile incelenmiştir.

Tablo 28. Kümelere Göre Cinsiyet Sayıları ve Oranları

Parametreler			Kümeler		Toplam
			1	2	
Cinsiyet	Erkek	n	22.275	13.416	35691
		Cinsiyete göre %	%62,4	%37,6	%100,0
		Kümeye göre %	%45,8	%45,3	%45,6
		Toplam %	%28,5	%17,1	%45,6
	Kadın	n	26.349	16.199	42.548
		Cinsiyete göre %	%61,9	%38,1	%100,0
		Kümeye göre %	%54,2	%54,7	%54,4
		Toplam %	%33,7	%20,7	%54,4
Toplam		n	48.624	29.615	78.239
		Cinsiyete göre %	%62,1	%37,9	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0
		Toplam %	%62,1	%37,9	%100,0

Tablo 29.Cinsiyete Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	1.925	1	0,165
Benzerlik Oranı	1.925	1	0,165
Geçerli Gözlem	78.239		

Elde edilen kümelerle cinsiyet arasında 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak bir fark olmadığı sonucu Tablo 29'da görülmektedir. Yani yoğunluğa dayalı kümeleme analizine göre cinsiyet kullanılan parametrelerde duyarsızdır. Tablo 28'de ki sonuçlarda kümelerdeki kadın ve erkek dağılımlarının neredeyse eşit olması bu durumu açıklamaktadır. Birinci kümedeki erkek oranı %54,2 ve ikinci kümedeki erkek

oranı da %54,7'dir. Ancak birinci kümedeki erkek oranı tüm veri setindeki kişilerin %33,7'si ve ikinci kümedeki erkek oranı %20,7 olarak hesaplanmıştır.

Tablo 30.Kümelere Göre Merkez-Merkez Dışı Sayıları ve Oranları

Parametreler			Kümeler		Toplam
			1	2	
İl Merkezi	Merkez Dışı	n	820	17.560	18.380
		Merkeze göre %	%4,8	%95,2	%100,0
		Kümeye göre %	%1,6	%61,5	%22,4
		Toplam %	%1,0	%22,4	%77,6
	Merkez	N	48.883	10.976	59.859
		Merkeze göre %	%81,7	%18,3	%100,0
		Kümeye göre %	%98,4	%38,5	%22,4
		Toplam %	%62,5	%14,0	%77,6
Toplam		n	49.703	28.536	78.239
		Cinsiyete göre %	%63,5	%36,5	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0
		Toplam %	%63,5	%36,5	%100,0

Tablo 31.Merkez-Merkez Dışı Olmasına Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	36.954,395	1	0,000
Benzerlik Oranı	40.399,976	1	0,000
Geçerli Gözlem	78.239		

Tablo 31'e göre hastaneye başvuran hastaların kümelerdeki dağılımlarına göre 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak anlamlı bir farklılık vardır. Hastaların il merkezlerinden gelip gelmemeleri durumunu gösteren Tablo 30'da,

hastaların çoğunluğunun il merkezinden geldiği görülmektedir. Ayrıca merkez dışından gelen hastaların çoğunluğu ikinci kümede yoğunlaşmıştır.

Tablo 32.Kümelere Göre Sivas-Sivas Dışı Sayıları ve Oranları

Parametreler			Kümeler		Toplam
			1	2	
Sivas	Sivas	n	40.825	26.536	67.361
		İle göre %	%60,6	%39,4	%100,0
		Kümeye göre %	%84,0	%89,6	%86,1
		Toplam %	%52,2	%33,9	%86,1
	Sivas Dışı	N	7.799	3.079	10.878
		İle göre %	%71,7	%28,3	%100,0
		Kümeye göre %	%16,0	%10,4	%13,9
		Toplam %	%10,0	%3,9	%13,9
Toplam		n	48.624	29.615	78.239
		Cinsiyete göre %	%62,1	%37,9	%100,0
		Kümeye göre %	%100,0	%100,0	%100,0
		Toplam %	%62,1	%37,9	%100,0

Tablo 33.Sivas-Sivas Dışı Olmasına Göre Ki-Kare Bağımsızlık Testi Sonuçları

	Test İstatistiği	Serbestlik Derecesi	P-Değeri
Pearson Ki-Kare	36.954,395	1	0,000
Benzerlik Oranı	40.399,976	1	0,000
Geçerli Gözlem	78.239		

Tablo 33'den de anlaşılacağı gibi, oluşturulan kümelerde hastaların Sivas ili dışından olup olmaması 0,05 ve 0,01 anlamlılık düzeylerinde istatistiksel olarak farklıdır. Tablo 32'de ki değerlere göre, hastaların çoğunluğu Sivas ili sınırlarından

gelmiştir. Bu oran tüm hastalar içinde %86,1'lik bir kısmı oluşturmaktadır. Sivas ili dışından gelen hastaların çoğunluğu birinci kümeye atanmıştır.

5.4. BULGULARIN DEĞERLENDİRİLMESİ VE SONUÇLAR

2011 yılında Cumhuriyet Üniversitesi Hastanesi'ne gelmiş olan hastalar üzerinden yapılan çalışmada veri madenciliğinde sıkça kullanılan yöntemlerden birisi olan K-Means kümeleme ve veri setindeki gürültülü veriler ile oldukça iyi sonuçlar üretebilen Yoğunluk Tabanlı Kümeleme yöntemleri uygulanmıştır.

K-Means kümeleme yönteminde, veri setinden herhangi bir örneklem alınmaz, veri seti olduğu gibi analize dahil edilir. Bu yöntemde, küme sayısını otomatik olarak belirlemek amacıyla bir mekanizma içermez. Dolayısıyla küme sayısını araştırmacı belirler. Böylelikle küme sayısını istediği gibi değiştirecek olan araştırmacı analiz noktasında bazı istatistiklere dikkat etmelidir. Bunlar, küme içi maksimum hata kareler ve kümeler arası benzemezlik istatistikleridir. Çalışmada küme sayısı ikiden ona kadar olacak şekilde hesaplanmıştır. Küme sayısı dokuz ve on olunca benzemezlik istatistiğinin değişmediği gözlenmiştir. Ancak küme içi maksimum hata kareler toplamı düşmektedir. Bu istatistiğin düşmesi iyi bir sonuç gibi gözükse de küme merkezlerinin giderek birbirine yaklaşması ve uç değerlerin varlığı ile küme ayrımının belirsizleşmesi sorunlarını da beraberinde getirmektedir. Bu sebeple küme sayısının üç olmasına karar verilmiştir.

Kümeler incelendiğinde hastaneden en çok hizmet alan grubun ikinci grup olduğu gözükmektedir. Bu durumda, analize katılan tüm parametrelerde ikinci kümedeki ortalamaların değerleri diğer kümlere göre daha yüksek olduğunu göstermektedir. Ortalama yaşta ikinci kümenin ortalaması daha yüksek olup neredeyse 65 yaş olan üst sınırdadır. Sadece üçüncü kümedeki, yani yaş olarak diğer kümedeki hastalardan daha genç olan hastalar kullanılan parametrelerin tamamında genel ortalamanın aşağısında kalmıştır. Birinci kümede sadece yatılan gün sayısı ve ücret

parametreleri genel ortalamanın aşağısındadır. Bunlardan ücret parametresi genel ortalamanın sadece 14,34 TL aşağısında olması önemli bir fark olarak görülmemektedir. Ücret parametresindeyse, birinci kümedeki hastaların genel ücret ortalamasının yaklaşık %2,8 daha azı maliyete katlanmakta oluğu sonucu ortaya çıkmıştır, buradan yola çıkarak hastaların üçüncü kümede % 52,5 daha az ve ikinci kümede %88,4 daha fazla sağlık ücretiyle ödedikleri yorumu söylenebilir.

Kümeler kendi arasında incelendiğinde ikinci kümedeki hastaların en az bir defa yatarak tedavi olduğu, yine en az bir defa ameliyat oldukları belirlenmiştir. Ayrıca ikinci gruptaki bir hastanın bir seferlik yatarak tedavisinin yaklaşık 8,5 gün sürdüğü buna karşılık birinci grupta ki bir hasta 6,9 gün ve üçüncü gruptaki bir hasta da 5,6 gün yatarak tedavi olduğu anlaşılmıştır. Genel ortalamaya bakıldığında bu sürenin 7,2 gün olduğu hesaplanmıştır. İkinci kümedeki bir hastanın polikliniğe gelişinde farklı bir teşhis konmasının olasılığı %54,2, birinci kümede aynı durumun olasılığı %57,3 ve üçüncü kümede %65,5 olarak hesaplanmıştır.

Kümeleme analizi sonucunda ortaya çıkan kümeler ile hastaların demografik verilerinin karşılaştırılması hastaların davranışlarını ortaya koymaya yardımcı olmuştur. Hastaların demografik verileri kümeler arasında bazı farklılıklara sebep olmaktadır.

Oluşturulan çapraz tablolarda küme içi oran olarak üçüncü kümedeki kadın oranını birinci ve ikinci kümeye göre daha fazla olduğu (birinci ve ikinci kümede %53,7; üçüncü kümede %55,2) görülmektedir. Bu durum kümelerin farklılaşması için yeterli olmaktadır. Bu sonuçlara göre hasta sayısının artmasıyla kümelerin cinsiyete parametresine göre farklılaşmasının ortadan kalkacağı düşünülmektedir.

Hastaların %76,5'lik kısmı il merkezlerinde yaşamaktadır. Birinci kümedeki hastaların %78,9'luk kısmı şehir merkezlerinde yaşamaktadır ki bu oran diğer gruplara göre daha yüksektir. Oran olarak şehir dışından gelen hastaların daha fazla olduğu küme %24,8 ile üçüncü kümedir. Hastaların yaşadıkları yerlerde kümeler arasında farklılık arz etmektedir.

Benzer şekilde hastaların üniversite hastanesine il dışından gelip gelmemesi incelenmektedir. Hastaların çoğunluğunun Sivas ili sınırlarından geldiği (%86,1) anlaşılmıştır. Birinci kümedeki hastaların çoğunluğu (%90,2) Sivas ilinden gelmiştir. Hastaların Sivas ili sınırlarından gelip gelmemesi arasındaki oran farkının en az olduğu küme üçüncü kümedir. Üçüncü kümedeki hastaların %81,8'lik kısmı Sivas ili sınırları içinden ve %18,2'lik kısmı da Sivas ili dışından gelmiştir. Bu parametreye göre kümeler arasında bir istatistiksel olarak bir farklılık olduğu bulunmuştur.

Yukarıdaki yorumlara bakılarak, kümeleme analizi sonuçlarının hastaların demografik verilerine göre açıklayıcılığının yeterliği olmadığı söylenebilir. Hastaların cinsiyetinin, yaşadığı yerin veya ilin kümelemede kullanılan parametrelere göre belirgin bir farklılığı olduğu fakat bu farklılığın hasta sayısı arttıkça azalacağı veya yok olacağı düşünülmektedir.

K-Means kümeleme yönteminin yanı sıra, veri setine Yoğunluk Tabanlı Kümeleme Algoritması da uygulanmıştır. Bu algoritma veri setini iki kümeye bölmüştür. Hastaların %62,1'i birinci kümeye ve %37,9'u ikinci kümeye ayrılmıştır. İkinci kümedeki hastalar tüm parametrelerde genel ortalamanın üstündedir. Aynı zamanda ikinci kümeye düşen hastalarda ilgili parametrelerde standart sapmaların da birinci kümedeki hastalara göre daha büyük olduğu görülmektedir. Birinci kümedeki hastaların yaş ortalaması 30,02 ve ikinci kümedeki hastaların yaş ortalaması 52,73 olarak hesaplanmıştır. Benzer şekilde ikinci kümedeki hastalar birinci kümedekilere göre %271 defa daha ayaktan tedavi olmuşlardır. Bir hastanın ortalama sağlık hizmeti ücretinin 1953 TL olarak hesaplanması göz önüne alınacak olunursa, birinci kümedeki hastaları genel ortalamanın yaklaşık 1228 TL daha az ve ikinci kümedeki hastalarında yaklaşık 2245TL daha fazla sağlık hizmetlerine ödeme yaptıkları anlaşılmaktadır.

Yoğunluk tabanlı kümeleme algoritması sonucunda kümeleri sağlık hizmetine daha fazla ihtiyaç duyan ya da talep eden grup, birinci kümeyi; sağlık hizmetlerine daha az ihtiyaç duyan ya da talep eden grup, ikinci kümeyi temsil etmektedir. Birinci kümeye dahil olan hastaların ortalama olarak 7,1 defa ayaktan tedavi aldıkları buna rağmen yeni

bir başvuru olması durumunda farklı bir teşhis ile karşılaşmaları olasılığı yaklaşık %63 olarak hesaplanmıştır. Aynı durum ikinci kümede incelenirse, hastaların ortalama olarak 12,2 defa ayaktan tedavi aldıkları ve buna karşılık yeni bir başvuruda bulunduklarında farklı bir teşhis ile karşılaşma olasılıkları %54,6 olarak karşımıza çıkmaktadır. Buradan da yaşı ilerlemiş hastalarda hastalıkların önceden belirlenmiş olduğu, başvuruların kronik hastalıkların yol açtığı sağlık sorunları olduğu düşünülmektedir. Yatarak tedavi edilen hastalar birinci kümede ortalama olarak 6 gün ve ikinci kümede de ortalama olarak 8,2 yatarak tedavi olmuşlardır.

Her iki yöntemde de sonuçlar birbirinden farklıdır. Ancak ortak olarak küme ayrımları yapılırken hizmet yoğunluğuna göre ayrım yapıldığı anlaşılmıştır. K-Means yönteminde olduğu gibi, yoğunluk Tabanlı algoritmada da kümeler hastaların demografik verileriyle incelenmiştir.

Yoğunluk tabanlı kümeleme analizi sonuçlarında ayrılan kümeler cinsiyet parametresi ile ilişkilendirildiğinde, kümelerde cinsiyet ayrımın herhangi bir farklılık olmadığı gözlenmiştir. Ancak kümeler ve hastaların il merkezinde ya da il merkezi dışında yaşamaları arasında kurulan çapraz tabloda, kümeler arasındaki farklılığın istatistiksel olarak anlamlı olduğu sonucuna varılmıştır. İl merkezi dışında yaşayan hastaların çoğunluğu (%95,2) ikinci kümede toplanmıştır. Birinci kümedeki hastaları ortalama sağlık hizmeti ücretinin daha az olduğu ve il merkezi dışında yaşayan hastaların çoğunluğunun ikinci kümede olduğu düşünülerek; il merkezi dışındaki hastaların sağlık hizmetlerine daha çok para harcadıkları söylenebilmektedir. Benzer şekilde hastaların Sivas ili sınırları içinden gelen hastaların küme içi oranı ikinci kümede birinci kümeye göre (1. Kümede %89,4; 2. Kümede %84) daha fazladır.

İncelenen iki yöntemde de sonuçlar incelenmiş olup, K-Means algoritmasında küme sayısının araştırmacı tarafından belirlenmesi eksiklik olarak düşünülmektedir. Ayrıca bu yöntemde verilerin kanonik bir şekle uyması analiz için daha uygun olacağı ve gürültülü değerlerden etkilenmesi hususları da analizin doğruluğunu etkilemektedir. Ancak yoğunluk tabanlı kümeleme algoritmasının verilerin şeklinden ya da veri

setindeki gürültülü verilerden etkilenmemesi sonuçların doğruluğu konusunda araştırmanın güvenirliliğini arttıran bir niteliktedir.

Kümeleme analizleri sonucunda oluşan hasta yapıları aşağıdaki tabloda verilmiştir (Tablo 34).

Tablo 34.Kümelere Göre Hasta Yapıları

<table>
<tr><th colspan="2">K-Means Kümeleme Sonuçlarına Göre Hasta Yapıları</th><th colspan="2">Yoğunluk Tabanlı Kümeleme Algoritması Sonuçlarına Göre Hasta Yapıları</th></tr>
<tr><td>1.Küme</td><td>• Orta Yaşlı
• Ortalama Sağlık Hizmet Alan
• İl Merkezinde Yaşayan
• Sivas'ta Yaşayan</td><td rowspan="2">1.Küme</td><td rowspan="2">▪ Genç
▪ Ortalama Altı Sağlık Hizmeti Alan
▪ İl Merkezinde Yaşayan
▪ Sivas Dışında Yaşayan</td></tr>
<tr><td>2.Küme</td><td>• Yaşlı
• Ortalama Üstü Sağlık Hizmeti Alan</td></tr>
<tr><td>3.Küme</td><td>• Genç
• Ortalama Altı Sağlık Hizmeti Alan
• Kadın
• İl Merkezi Dışında Yaşayan
• Sivas Dışından Gelen</td><td>2.Küme</td><td>▪ Yaşlı
▪ Ortalama Üstü Sağlık Hizmeti Alan
▪ İl Merkezi Dışında Yaşayan
▪ Sivas'ta Yaşayan</td></tr>
</table>

Her iki uygulamanın sonucunda da kümeler için yaş ve hastaların sağlık hizmetinden ne kadar faydalandıkları ayırıcı etken olmuştur. K-Means kümeleme analizinde kadınların diğer kümelere göre daha çok üçüncü kümede yoğun olarak bulunduğu gözlenmiştir. Yoğunluk tabanlı kümeleme algoritmasında ise birinci kümedeki hastaların Sivas dışında ve il merkezlerinde yaşayan hastalar olduğu, ikinci kümede ise bu durumun tam tersi olduğu görülmektedir.

KAYNAKÇA

Abubaker, M. B., (2011). Efficient Data Clustering Algorithms (Yayımlanmamış Yüksek Lisans Tezi).Islamic University of Gaza/Faculty of Engineering, Palestine.

Adèr, H. J., ve Mellenbergh, G. J. (2008). Advising on research methods: A consultant's companion..Huizen, Netherland. Johannes van Kessel Publishing

Akın, Y. K. (2008). Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi. (yayınlanmamış Doktora Tezi). Marmara Üniversitesi/Sosyal Bilimler Enstitüsü, İstanbul.

Aldrich, J.,(2005). Fisher and Regression. Statistical Sciences, 20(4), 401-417.

Allen, F. H., Doyle, M. J. ve Taylor, R. (1991). Automated Conformational Analysis from Crystallographic Data. Symmetry-Modified Jarvis-Patrick and Complete-Linkage Clustering Algorithms for Three-Dimensional Pattern Recognition. *International Union of Crystallography, 47(1),* 41-49.

Altıntaş, T. (2006). *Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi.* (Yayınlanmamış Yüksek Lisans Tezi). Sakarya Üniversitesi/Fen Bilimleri Enstitüsü, Sakarya.

Altun Ada, A. (2011). Kümeleme Analizi İle AB Ülkeleri ve Türkiye'nin Sürdürülebilir Kalkınma Açısından Değerlendirilmesi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 29,* 319-332.

Ankerst, M., Breunig, M. M., Kriegel H. P. ve Sander, J. (1999). *OPTICS: Ordering Points to Identify the Clustering Structure.* ACM SIGMOD'99 International Conference on Management of Data, New York, USA, 1-3 June.

Atbaş, A. C. G. (2008). Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma. (Yayınlanmamış Yüksek Lisans Tezi). Ankara Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Awad, M., Khan, L., Thuraisingham, B. ve Wang, L. (2009). *Design and Implementation of Data Mining Tools.* Florida, USA, Auerbach Publications.

Barioni, M. C. N., Razente, H. L., Traina, A. J. M. ve Traina, C. Jr. (2006). An Efficient Approach to Scale up K-Medoid Based Algorithms in Large Databases. *Brazilian Symposium on Database-SBBD,* 265-279.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. Kogan, J., Nicholas, C ve Teboulle, M. (Eds). *Grouping Multidimensional Data* içinde (25-71). Berlin, Germany, Springer

Borah, B. ve Bhattacharyya, D. K. (2004). *An Improved Sampling-Based DBSCAN for Large Spatial Databases.* 8th International Conference on Spoken Language Processing, Jeju Island, South Korea, 4-8 October.

Bruno, G. ve Fiori, A. (2011). Microarray Data Mining Issues and Prospects. Vipin Kumar (Ed.). *Knowledge Discovery Practices and Emerging Applications of Data Mining* içinde (23-47), New York, USA, Information Science Reference.

Chakrabarti, S.,Cox. E., Frank, Eibe., Güting, R. H., Han J., Kamber, M., Lighstone, S. S., Nadeau, T. P., Neapolitan R. E., Pyle, D., Refaat, M., Schneider, M., Teorey, T. J. ve Witten I. H. (2009). Data Mining Know It All. Massachusetts, USA, Elsevier.

Chavent, M., Lechevallier, Y. ve Briant, O. (2007). DIVCLUS-T: A Monothetic Divisive Hierarchical Clustering Method. *Computational Statistical & Data Analysis, 52,* 687-701.

Chen, M. S., Han, J. ve Yu, P. S. (1996). Data Mining: An Overview From a Database Perspective. *IEEETransactions on Knowledge Data Engineering, 8(6),* 866-883.

Chidananda Gowda, K. ve Ravi T.V. (1995). Agglomerative Clustering of Symbolic Objects Using the Concepts of Both Similarity and Dissimilarity. *Pattern Recognition Letters,16,* 647-652.

Cohen, G. L. ve Shannon, A. G. (1981). John Ward's Method for The Calculation of Pi. *Historia Mathematica, 8,* 133-144.

Doğan, Ş. (2007). *Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi.* (Yayınlanmamış Yüksek Lisans Tezi). Fırat Üniversitesi/Fen Bilimleri Enstitüsü, Elazığ.

Doorn, J. H.,& Rivero, L. C. (2002). Database integrity: challenges and solutions.London, England, IGI Global.

Dua, S. ve Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity.* Florida, USA, Auerbach Publications.

Duan, L., Xu, L., Guo, F., Lee, J. ve Yan, B. (2007). A Local-Density Based Spatial Clustering Algorithm With Noise. *Information Systems, 32,* 978-986

Dutta, M., Kakoti Mahanta, A. ve Pujari, A. K. (2005). QROCK: A Quick Version of The ROCK Algorithm for Clustering of Categorical Data. *Pattern Recognition Letters, 26,* 2364-2373.

Elkan, C. (2003). *Usşng the Triangle Inequatily to Accelerate K-Means.* 20th International Conference on Machine Learning, Washington DC, USA, 21-24 August.

Elmasri, R. ve Navathe, S. B. (2003). *Fundamentals of Database Systems (4th Edition).* Boston, USA, Pearson Education Inc.

Enders, C. K. (2010). *Applied Missing Data Analysis.* London, United Kingdom, The Guilford Press.

Ertuğrul, İ., Organ, A., ve Şavlı, A. (2013). Veri Madenciliği Uygulamasına İlişkin PAÜ Hastanesinde Hasta Profilinin Belirlenmesi. *Pamukkale University Journal of Engineering Sciences,* 19(2).97-103

Ester, M., Kriegel, H. P., Sander, J. ve Xu, X. (1996). *A Density-Based Algorithm for Discovering Cluster in Large Spatial Databases with Noise*. 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, 2-4 August.

Frahling, G.ve Sohler, C. (2005). Coresets in Dynemic Geometric Data Streams. 37. Annual ACM Symposium on Theory of Computing. New York, USA, 22-24 May

Freanti, P., Virmajoki, O. ve Hautamaeki, V. (2006). Fast Agglomerative Clustering Using a K-Nearest Neighbor Graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(11),* 1875-1881.

Fu, Y.,Sandhu, K. ve Shih, M. Y., (1999). *Clustering of Web Users Based on Access Patterns*. 1999 KDD Workshop on Web Mining, San Diego, USA, 15-18 August

Garcia, M. H., Ullman, J. D. ve Widom, J. (2008). *Database Systems: The Complete Book Second Edition.* New Jersey, USA, Pearson Education Inc.

Gardner, M. (1997). Hydras, Eggs, and Other Mathematical Mystifications. Martin Gardner (Ed.). Taxicap Geometry içinde (159-175). New York, USA, Springer.

Ghahramani, Z. (2004). Unsupervised Learning. Bousquet, O., Raetsch, G. ve Luxburg, U. (Ed.), *Advances Lectures on Machine Learning,* (3-32). Verlag, Springer.

Ghanti, V., Gehrke, J ve Ramakrishnan, R. (1999). Mining Very Large Databases. *Computer, 3,* 38-45.

Göral, M. A. (2007). *Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli.* (yayınlanmamış Yüksek Lisans Tezi). İstanbul Teknik Üniversitesi/Fen Bilimleri Enstitüsü, İstanbul.

Grünberg,T. *Bilgi ve Bilim Felsefesi*. http://dergiler.ankara.edu.tr/dergiler/34/969/11930.pdf

Guha, S.,Rastogi, Rastogi, R. ve Shim, K., (1998). CURE: An Efficent Clustering Algorithm for Large Databases, ACM SIGMOD Record, 27(2), 73-84.

Gupta, M., ve Shrivastava, V. (2013). *Review of various Techniques in Clustering. International Journal*, 3(2), 134-137

Han, J., Kamber, M. ve Pei, J. (2012). *Data Mining Concepts and Techniques (3. Baskı).* Boston, USA, Elsevier.

Hinneburg, A. ve Keim, D. A. (1999). *Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering.* 25th Very Large Database Conference, Edinburgh, Scotland, 7-10 September.

http://binedir.com/blogs/veri-madenciligi/image_6e9f15ed.png, 26.08.2012

http://en.wikipedia.org/wiki/Cluster_analysis, 10.06.2013

http://en.wikipedia.org/wiki/Data_mining, 06.06.2013

http://hastane.dicle.edu.tr/index.php?option=com_content&view=article&id=125:icd, 11.08.2012

http://lh6.ggpht.com/_e7UyIXjsjN8/SbU6lyv5LII/AAAAAAAADUs/UQdsMwEIr6c/s800/sting.JPG, 20.09.2013

http://okul.selyam.net/pars_docs/refs/25/24645/24645_html_m4a91cb91.png, 15.08.2012

http://scikit-learn.org/stable/_images/plot_cluster_comparison_11.png, 14.08.2013

http://scikit-learn.org/stable/_images/plot_mean_shift_11.png, 12.08.2013

http://www.iszekam.net/image.axd?picture=2009%2F5%2F3-11%28k-meds%29.jpg, 24.06.2013

http://www.yazgelistir.com/Makaleler/Resimler/1000000753_image002.jpg, 01.02.2013

http://www.yazgelistir.com/Makaleler/Resimler/1000001858_DataMining1.png, 05.06.2013

Imhoff, C., Galemmo, N. ve Geiger, J. G. (2003). *Mastering Data Warehouse Design Relation and Dimensional Techniques.* Indianapolis, USA, Wiley Publishing.

Işık, M. ve Çamurcu, A. Y. (2007). K-Means, K-Medoids ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. İstanbul Ticaret Üniversitrsi Fen Bilimleri Dergisi. 11(1), 31-45

Jackson, S. L. (2009). *Research Methods and Statistics: A Critical Thinking Approach (3. Baskı).* San Francisco, USA, Wadsworth Cengage Learning.

Jiang, D., Pei, J. ve Zhang A. (2003). *DHC: A Density-Based Hierarchical Clustering Method for Time Series Gene Expression Data.* Third IEEE Symposium on BioInformatics and BioEngineering, Washington, DC, USA, 12 March.

Kandartzic, M. (2011). *Data Mining Concepts, Models, Methods and Algorithms (Second Edition).* New Jersey, USA, Wiley Publishing.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. ve Wu, A. Y. (2002). An Efficient K-Means Clustering Algorithm: Analysis and Implementation. Pattern Analysis and Machine Intelligence, 24(7), 881-892

Karahoca, A. (Ed.) (2012). *Advances in Data Mining Knowledge Discovery and Applications.* Rijeka, Croatia, InTech.

Kaur, H., ve Wasan, S. K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. Journal of Computer Science, 2(2).194-200

Kogan, J.,Nicholas, C. ve Teboulle. M., (Ed.) (2006). Grouping Multidimensional Data: Recent Advances in Clustering.Huizen, Netherlands: Springer.

Koh, H. C.,& Tan, G. (2011). Data mining applications in healthcare. Journal of Healthcare Information Management—Vol, 19(2), 65-75.

Koperski, K., Han, J. ve Adhikary, J., (1998).Mining Knowledge in Geographical Data,Communications of the ACM, 26(1), 65-74.

Korkmaz, A. (2005). Olasılık Kuramının Doğuşu. *Ankara Üniversitesi Siyasal Bilgiler Fakültesi Dergisi, 60(2).* 171-193.

Koyuncugil, A. S. ve Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. *Bilişim Teknolojileri Dergisi, 2(2),* 21-32.

Krigel, H. P., Kröger, P., Sander, J. ve Zimek, A. (2011). Density-Based Clustering. *Wire's Data Mining and Knowledge Discovery,* 3, 231-240.

Krishnapuram, R., Joshi, A. ve Yi, L. (1999). *A Fuzzy Relative oh The K-Medoids Algorithm With Application to Web Document and Snippet Clustering.* International Fuzzy Systems Conference, Seoul, Korea, 22-25 August.

Li, X. ve Luo, M. (2009). *An Improved WaveCluster Algorithm Based on ICA.* 5th International Conference on Wireless Communications, Networking and Mobile Computing, Beijing, China, 24-26 September.

Lin, D.,ve Wu, X. (2009). Phrase Clustering For Discriminative Learning. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL. Singapore City, Singapore, 2-7 August.

Lu, Y., Sun, Y., Xu, G. ve Liu, G. (2005). A Grid-Based Clustering Algorithm for High-Dimensional Data Streams, Alhajj, R. (Ed.), *Advance Data Mining and Applications* içinde(824-831). Berlin, Germany, Springer.

Ma, E. W. M. ve Chow, T. W. S. (2004). A new Shifting Grid Clustering Algorithm. The Journal of The Pattern Recognition Society. 37, 503-514

Mogull, Robert G. (2004). Second-Semester Applied Statistics. New York, USA Kendall/Hunt Publishing

Moody, D. L. ve Kortink, M. A. R. (2000). *From Enterprise Models to Dimensional Modls: A Methodology for Data Warehouse and Data Mart Design.* Design and Management of Data Warehouses 2nd International Workshop, Stockholm, Sweden, 5-6 June.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Greg Miller, W., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S. ve Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in biology and medicine, 36(12), 1351-1377.

Nagadevara, V. (2004). Application of neural prediction models in healthcare. In Proceedings of the 2nd International conference on e-Governance ICEG (pp. 139-149).

Nakip, M. (2013). *Pazarlamada Araştırma Teknikleri (3.Baskı).* Ankara, Türkiye, Seçkin

Ng, R. T. ve Han, J. (2002). CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEETransactions on Knowledge and Data Engineering, 14(5),* 1003-1016.

Nisbet, R., Elder, J. ve Miner G. (2009). *Handbook od Statistical Analysis Data Mining Applications.* Toronto, Canada, Elsevier.

Omrani, A., Santhisree, K. ve Damodaram, A. (2011). *Clustering Sequential Data With OPTICS.* 3rd International Conference on Communication Software and Networks, Xi'an, China, 27-29 May.

Öz, B., Taban, S. ve Kar, M. (2009). Kümeleme Analizi İle Türkiye ve AB Ülkelerinin Beşeri Sermaye Göstergeleri Açısından Karşılaştırılması. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 10,* 1-30.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri.* İstanbul, Türkiye, Papatya Yayıncılık.

Penrose, M. D. (1995). Single Linkage Clustering and Continuum Percolation. *Journal of Multivariate Analysis, 53,* 94-109.

Pilevar, A. H. ve Sukumar,M. (2005). GCHL: A Grid-Clustering Algorithm for High-Dimensional Very Large Spatial Data Bases. *Pattern Recognition Letters, 26,* 999-1010.

Prasad, P. K. ve Pandurangan, C. (2005). Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databased, Alhajj, R. (Ed.), *Advance Data Mining and Applications* içinde (146-157). Berlin, Germany, Springer.

Rafsanjani, M. K., Varzaneh, Z. A. ve Chukanlo, N. E. (2012). A Survey Hierarchical Clustering Algorithms. *The Journal of Mathematics and Computer Science, 5(3),* 229-240.

Rajagopal, S. ve Selvi, T. S. (2006). *Semantic Grid Service Discovery Approach Using Clustering of Service Ontologies.* IEEE Region 10 Conference TENCON 2006, Hong Kong, Hong Kong, 14-17 November.

Ramakrishnan, R. ve Gehrke, J. (2000). *Database Management Systems (Second Edution).* California, USA, McGraw-Hill.

Rao., A. R. ve Srinivas, V. V. (2005). Regionalization of Watersheds by Hybrid-Cluster Analysis. *Journal of Hydrology*, 318, 37-56

Sander, J. (1998). *Generalized Density-Based Clustering for Spatial Data Mining (Yayınlanmamış Doktora Tezi).* Ludwig Maximilians Universitaed/Fakultaet für Mathematik und Informatik, München, Germany.

Sheikholeslami, G., Chatterjee, S. ve Zhang, A. (2000). WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases. *The Very Large Data Bases Journal, 8,* 289-304.

Sibson, R. (1972). SLINK: An Optimally Efficient Algorithm fort he Single-Link Cluster Method. *The Computer Journal, 16(1),* 30-34.

Silahtaroğlu, G. (2008). *Kavram ve Algoritmalarıyla Temel Veri Madenciliği.* İstanbul, Türkiye, Papatya Yayıncılık.

Silberschatz, A., Korth H. ve Sudarshan, S. (2010). *Database Systems Concepts (6th Edition).* Columbus, USA, McGraw-Hill.

Taşkın, Ç. ve Emel, G. G. (2010). Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları İle Perakendecilik Sektöründe Bir Uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 15(3),* 395-409.

Tekerek, A. (2011). *Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları.* XIII. Akademik Bilişim Konferansı, Malatya, Türkiye, 2-4 Şubat.

Terlemez, L. (2008). *Eş İşlem Stratejisi Yöntemiyle İMKB'de Portföy Oluşturmada Veri Madenciliği Uygulaması.* (Yayınlanmamış Doktora Tezi). Anadolu Üniversitesi/Fen Bilimleri Enstitüsü, Eskişehir.

Thalheim, B. (2000). *Entity-Relational Modeling Foundation of Database Technology.* Berlin, Germany, Springer.

Thomsen, E. (2002). *OLAP Solutions Building Multidimensional Information Systems Second Edition.* New York, USA, Wiley.

Tuna, M. F. (2013). *Pazarlama Kapsamında Coğrafi Bilgi Sistemlerinin Konut Fiyatlarının Belirlenmesinde Kullanımı: Ankara İlinde Bir Uygulama* (Yayınlanmamış Yüksek Lisans Tezi). Cumhuriyet Üniversitesi/Sosyal Bilimler Enstitüsü, Sivas

Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), 966-970

Vatansever, M. (2008). *Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması.* (Yayınlanmamış Yüksek Lisans Tezi). Yıldız Teknik Üniversitesi/Fen Bilimleri Enstitüsü, İstanbul.

Vattani, A. (2011). K-Means Requires Exponential MAny Iterations Even in the Plane. Discrete Computer Genom. 45, 596-616

Wang, W., Yang, J. ve Muntz, R. (1997). *STING: A Statistical Information Grid Approach to Spatial Data Mining.* 23th Very Large Database Conference, Athens, Greece, 20 February.

Witten, I. H., Frank, E. ve Hall, M. A. (2011). *Data Mining: Machine Learning Tools and Techniques.* New York, USA, Morgan Kaufmann Publisher.

Yaari, Y. (1999). *Segmentation of Expository Texts by Hierarchical Agglomerative Clustering.* 2nd International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 24 September.

Ye, Q., Gao, W. ve Zeng, W. (2003). *Color Image Segmentation Using Density-Based Clustering.* IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, Hong Kong, 6-10 April.

Yıldırım, A. A. ve Özdoğan, C. (2011). Parallel WaveCluster: A Linear Scaling Parallel Clustering Algorithm Implementation With Application to Very Large Datasets. *Journal of Parallel and Distributed Computing, 71,* 955-962.

Young, R., ve Johnson, D. R. (2010). 'Imputing the Missing Y's: Implications for Survey Producers and Survey Users. In 64th Annual Conference of the American Association for Public Opinion Research. may 14-17

Zhang, T., Ramarkrishnan, R. ve Livny, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery, 1,* 141-182.

Zhang, Y. P., Sun, J. Z., Zhang, Y. ve Zhang, X. (2004). *Parallel Implementatin of Clarans Using PVM.* Third International Conference on Machine Learning and Cybernetics, Shangai, China, 26-29 August.

Zhao, Y. ve Karypis G. (2002). *Comparison of Agglomerative and Partitional Document Clustering Algorithms (02-014).* Minneapolis, USA, Department of Computer Science and Engineering University of Minesota.