

VERİ MADENCİLİĞİ

(Data Mining)

(Veri Madenciliğine Giriş)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Ders Bilgileri

- EME4214 Veri Madenciliği
- Ders ile ilgili duyurular
 - <http://kergun.baun.edu.tr/>
- Kaynaklar
 - İTÜ Veri Madenciliği Ders Notları, Şule Gündüz Öğdücü
 - Veri Madenciliği Yöntemleri, Yalçın Özkan.
 - Veri Madenciliği: Kavram ve Algoritmaları, Gökhan Silahtaroğlu.
 - Veri Madenciliği(Kavram ve Teknikler), Aysan Şentürk.
- Başarı Notu
 - Vize (%40)
 - Final (%60)

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
 - Sınıflandırma
 - Kümeleme
 - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

VERİ MADENCİLİĞİNE GİRİŞ

İçerik

- Veri madenciliği ve bilgi keşfinin tanımı
 - Veri madenciliğinin tarihçesi
 - Veri madenciliğinin uygulama alanları
 - Veri madenciliğinde temel kavramlar
 - Veri kaynakları
 - Veri madenciliği modellerinin gruplanması
 - Veri ambarları
 - Veri madenciliğinde sorunlar
-

Veri Madenciliği Giriş

- İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızlı artışı ve bilginin fazlalaşması sebebiyle öncelikle, genelde Veri Tabanlarında Bilgi Keşfi olarak adlandırılan yeni bir paradigma ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana **Veri Madenciliği** denilmektedir.

Veri Madenciliği Tanımları

(1/2)

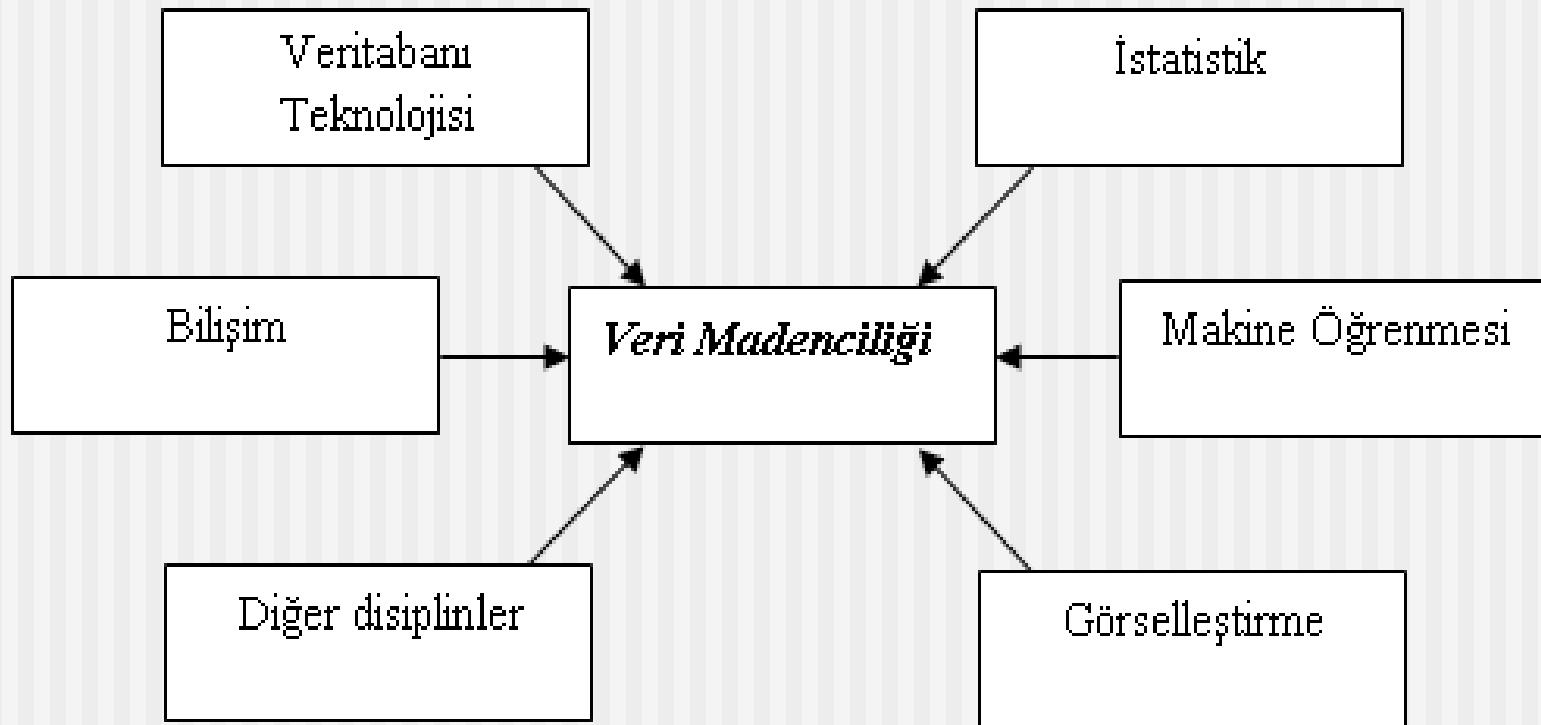
- Veri Madenciliği(Data Mining): Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak **bağıntı** ve **kuralların** aranmasıdır. (*Knowledge Discovery in Databases*)
- Daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır.
- Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.
- *Yapısal* veritabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabılır örüntülerin tanımlanması işlemidir.

Veri Madenciliği Tanımları

(2/2)

- Bu tanımlamalardan da anlaşıldığı üzere veriler arasındaki ilişkileri ortaya koymak ve gerekiğinde ileriye yönelik tahminlerde bulunmak veri madenciliği çalışmaları sayesinde mümkün olmaktadır. Bunun anlamı, veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi ortaya çıkarma süreci olarak değerlendirilmesidir. Bu açıdan bakıldığından veri madenciliği işinin kurumların Karar Destek Sistemleri için önemli bir yere sahip olduğu söylenebilir.
- Veri madenciliği çalışmaları, *sınıflandırma*, *ilişki kurma*, *kümeleme*, *regresyon*, *veri özetleme*, *değişikliklerin analizi*, *sapmaların tespiti* gibi belirli sayıda teknik yaklaşımları içerir.

Veri Madenciliği ile İlişkili Diğer Disiplinler



Veri Madenciliğinin Tarihçesi (1/4)

- Data FishingData Dredging: 1960
 - istatistikçiler
- Data Mining: 1990
 - veritabanı kullanıcıları, ticari
- Knowledge Discovery in Databases (KDD): 1989
 - Yapay zeka, makine öğrenmesi toplulukları
- Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction,...

Veri Madenciliğinin Tarihçesi (2/4)

- Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. O dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenilmiştir. Bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilmiştir.

Veri Madenciliğinin Tarihçesi (3/4)

- 1990'lı yıllara gelindiğinde Veri Madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atıldı. Bu camianın amacı, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktı. Bu noktadan sonra bilimadamları veri madenciliğine çeşitli yaklaşımalar getirmeye başladılar. Bu yaklaşımların kökeninde istatistik, makine öğrenmesi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar yatmaktadır.

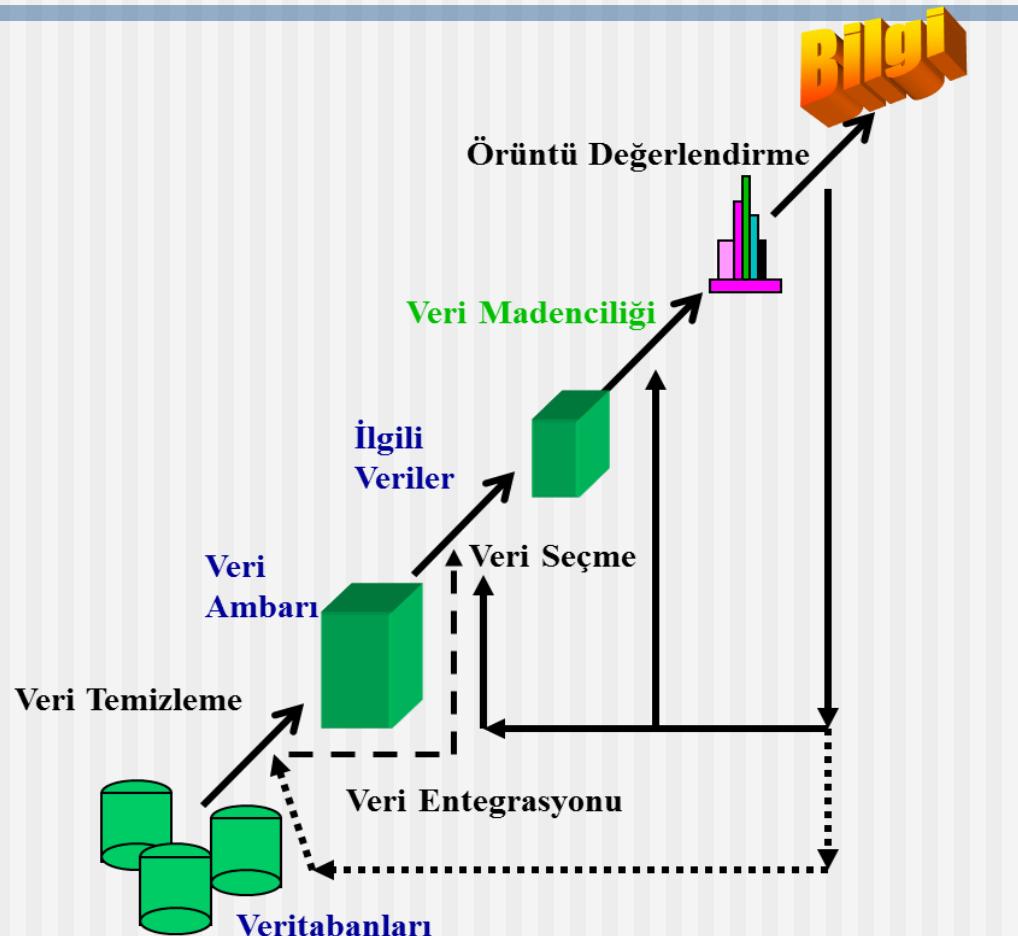
Veri Madenciliğinin Tarihçesi (4/4)

- İstatistik, süre gelen zaman içerisinde verilerin değerlendirilmesi ve analizleri konusunda hizmet veren bir yöntemler topluluğuydu. Bilgisayarların veri analizi için kullanılmaya başlamasıyla istatistiksel çalışmalar hız kazandı. Hatta bilgisayarın varlığı daha önce yapılması mümkün olmayan istatistiksel araştırmaları mümkün kııldı. 1990lardan sonra istatistik, veri madenciliği ile ortak bir platforma taşındı. Verinin, yiğinlar içerisinde çekip çıkarılması ve analizinin yapılarak kullanıma hazırlanması sürecinde veri madenciliği ve istatistik sıkı bir çalışma birlikteliği içine girmiş bulundular.
- Bunun yanısıra veri madenciliği, veritabanları ve makine öğrenimi disipliniyle birlikte yol aldı. Günümüzdeki Yapay Zeka çalışmalarının temelini oluşturan makine öğrenimi kavramı, bilgisayarların bazı işlemlerden çıkarsamalar yaparak yeni işlemler üretmesidir. Önceleri makineler, insan öğrenimine benzer bir yapıda inşa edilmeye çalışıldı. Ancak 1980lerden sonra bu konuda yaklaşım değişti ve makineler daha spesifik konularda kestirim algoritmaları üretmeye yönelik inşa edildi. Bu durum ister istemez uygulamalı istatistik ile makine öğrenim kavramlarını, veri madenciliği altında bir araya getirdi.

Bilgi Keşfi

- Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir.
- Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.
- Veri madenciliği teknikleri veriyi belli bir modele uydurur.
 - veri içindeki örüntüleri bulur
 - örüntü: veri içindeki herhangi bir yapı
- Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.
- Büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak
- Bulunan bilgi
 - gizli,
 - önemli,
 - önceden bilinmeyen,
 - yararlı olmalı.

Bilgi Keşfi



Bilgi Keşfinin Aşamaları

- Veri Temizleme : Gürültülü ve tutarsız verileri çıkarmak
- Veri Bütünleştirme: Birçok data kaynağını birleştirebilmek
- Veri Seçme : Yapılacak olan analiz ile ilgili olan verileri belirlemek
- Veri Dönüşümü : Verinin veri madenciliği yöntemine göre hale dönüşümünü gerçekleştirmek
- Veri Madenciliği : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin uygulanması
- Örüntü Değerlendirme: Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumunu

Veri Madenciliği Uygulama Alanları

- Veritabanı analizi ve karar verme desteği
 - Pazar araştırması
 - Hedef Pazar, müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi
 - Risk analizi
 - Kalite kontrolü, rekabet analizi, öngörü
 - Sahtekarlıkların saptanması
- Diğer Uygulamalar
 - Belgeler arası benzerlik (haber kümeleri, e-posta)
 - Sorgulama sonuçları

Veri Madenciliği Uygulama Alanları

Bilim	İş Hayatı	Web	Devlet
<ul style="list-style-type: none">• Astronomi• Biyoinformatik• İlaç keşfi	<ul style="list-style-type: none">• Reklam• CRM (Müşteri İlişkileri Yönetimi) ve Müşteri Modelleme• E-ticaret• Yatırım değerlendirme ve karşılaştırma• Sağlık• Üretim• Spor/eğlence• Telekom (telefon ve iletişim)• Hedef pazarlama	<ul style="list-style-type: none">• Metin Madenciliği (haber grubu, email, dokümanlar)• Web analizi• Arama motorları	<ul style="list-style-type: none">• Terörle Mücadele• Kanun Yaptırımı• Vergi• Kaçakçılarının Profilinin Çıkarılması

Uygulamalar

- Hangi promosyonu ne zaman uygulamalıyım?
- Hangi müşteri aldığı krediyi geri ödemeyebilir?
- Bir müşteriye ne kadar kredi verilebilir?
- Sahtekarlık olabilecek davranışlar hangileridir?
- Hangi müşteriler yakın zamanda kaybedilebilir?
- Hangi müşterilere promosyon yapmalıyım?
- Hangi yatırım araçlarına yatırım yapmalıyım?

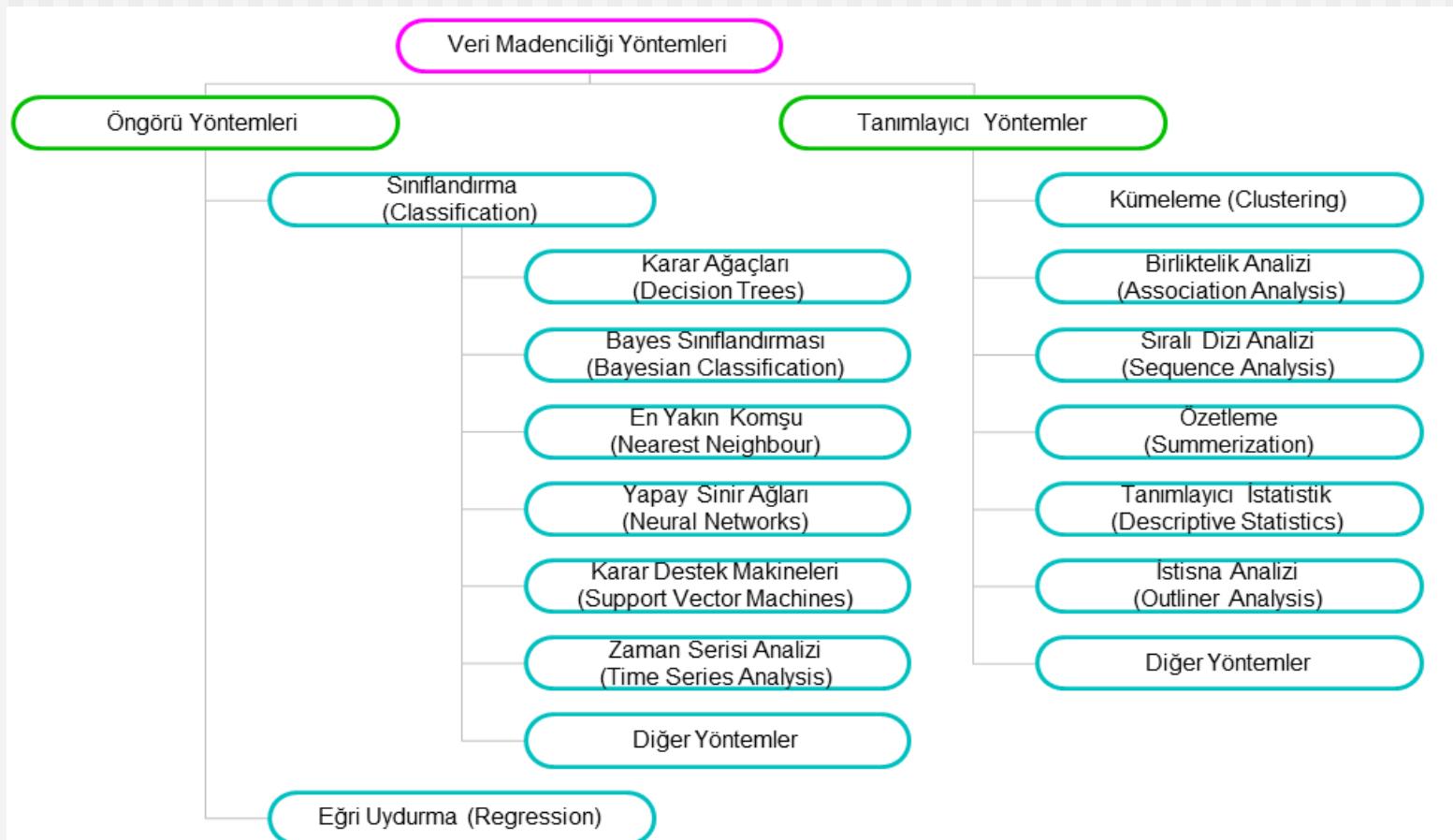
Veri Kaynakları

- Veri dosyaları
- Veritabanı kaynaklı veri kümeleri
 - ilişkisel veritabanları, veri ambarları
- Gelişmiş veri kümeleri
 - duraksız veri (data stream), algılayıcı verileri (sensor data)
 - zaman serileri, sıralı diziler (biyolojik veriler)
 - çizgeler, sosyal ağ (social networks) verileri
 - konumsal veriler (spatial data)
 - çokul ortam veritabanları (multimedia databases)
 - nesneye dayalı veritabanları
 - www

Veri Madenciliği Algoritmaları

- amaç: veriyi belli bir modele uydurmak
 - tanımlayıcı
 - En iyi müşterilerim kimler?
 - Hangi ürünler birlikte satılıyor?
 - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
 - kestirime dayalı
 - Kredi başvurularını risk gruplarına ayırma
 - Şirketle çalışmayı bırakacak müşterileri öngörme
 - Borsa tahmini
- seçim: veriye uygun en iyi modeli seçmek için kullanılan kriter
- arama: veri üzerinde arama yapmak için kullanılan teknik

Veri Madenciliği Yöntemleri



Veri Madenciliği İşlevleri

(1/2)

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
 - Danışmanlı (Gözetimli) öğrenme
 - Örütü tanıma
 - Kestirim
- Eğri uydurma (Regression): Veriyi gerçek değerli bir fonksiyona dönüştürür.
- Zaman serileri incelemesi (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- İstisna Analizi (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

Veri Madenciliği İşlevleri

(2/2)

- Kümeleme (Clustering): Benzer verileri aynı grupta toplama
 - Danışmansız (Gözetimsiz) öğrenme
- Özetteleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
 - Genelleştirme (Generalization)
 - Nitelendirme (Characterization)
- İlişkilendirme kuralları (Association Rules)
 - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

Veri Madenciliğinde Temel Kavramlar

- Veri (*Data*)
- Enformasyon (*Information*)
- Bilgi (*Knowledge*)
- Bilgelik (*Wisdom*)

Veri (Data)

(1/2)

- Veri kelimesi Latince'de "gerçek, reel" anlamına gelen "datum" kelimesine denk gelmektedir. "Data" olarak kullanılan kelime ise çoğul "datum" manasına gelmektedir. Her ne kadar kelime anlamı olarak gerçeklik temel alınsa da her veri her daim somut gerçeklik göstermez. Kavramsal anlamda veri, kayıt altına alınmış her türlü olay, durum, fikirdir. Bu anlamıyla değerlendirildiğinde çevremizdeki her nesne bir veri olarak algılanabilir.

Veri (Data)

(2/2)

- Veri, oldukça esnek bir yapıdadır. ■ Temel olarak varlığı bilinen, işlenmemiş, ham haldeki kayıtlar olarak adlandırılırlar. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. Ancak bu durum her zaman geçerli değildir. İşlenerek farklı bir boyut kazanan bir veri, daha sonra bu haliyle kullanılmak üzere kayıt altına alındığında, farklı bir amaç için veri halini koruyacaktır. Bu konuya daha iyi açıklayabilmek için enformasyon kavramını incelemek gerekmektedir.
- a. Bir araştırmacıın, bir tartışmanın, bir muhakemenin temeli olan ana öge.
b. Bir sanat eserine veya bir edebî esere temel olan ana ilkeler:
"Bir romanın verileri."
c. Bilgi, data.
d. Matematik: Bir problemde bilinen, belirtilmiş anlatımlardan bilinmeyeni bulmaya yarayan şey.
e. Bilişim: Olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimli gösterimi.

Enformasyon (Information)

- Enformasyon, veri kavramının tanımından yola çıkıldığında, adreslemedeki ikinci safhadır. Yani verilerin ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış, işlenmiş halidir. Bu haliyle enformasyon, potansiyel olarak içinde bilgi barından bir veri halindedir.
 - Belli bir alanda ve belli bir toplumda bilgi ve haberlerin yayılmasına olanak sağlayan araçların tümüne verilen isimdir.
 - Enformasyon, genel olarak insanın dış dünyaya ilişkisinde, belirsizlik düzeyini azaltan her tür uyarın şeklinde tanımlanabilir. Daha özel olarak ise, formatlanmış ve yapılandırılmış veriler bütünü olarak tanımlanabilir.
- Yaygın anlamda enformasyon terimi, "haber" (ing. news, alm. nachrichf) veya mesaj terimiyle eşanlamlıdır.
- Veriler enformasyona dönüştürülerek kullanışlı hale getirilirler. Bu yönyle enformasyon anlam katılmış verilerdir.

Bilgi (Knowledge)

- Bilgi, bu süreçteki üçüncü aşamadır. Enformasyonun, bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir. Dolayısıyla bireyin algılama yeteneği, yaratıcılık, deneyim gibi kişisel nitelikleri de bu süreci doğrudan etkilemektedir.
- «İnsan aklının erebileceği olgu, gerçek ve ilkeler bütünü, malumat» olarak sözlüğümüzde tanımlanan bilgi, bilişim dilinde kurallardan yararlanarak kişinin veriye yönelttiği anlam demektir.
- Felsefi olarak ise insanların maddi ve toplumsal anıksal etkinliğinin ürünü olarak tanımlanmaktadır. Enformasyonun daha yüksek biçimi olarak bilginin tüm modelleri altında yatan, bilginin ham maddelerinden onlara anlam eklerek ortaya çıkarılması gereği düşüncesidir. Bilgiden, farklı enformasyon parçacıkları arasındaki ilişkiler anlaşılmalıdır. Örneğin bir kişiyi sadece bir T.C kimlik numarasının temsil edebileceği bilgisine sahip olunmalıdır.

Bilgelik (Wisdom)

- Bilgelik ulaşılmasına çalışan noktadır ve bu kavramların zirvesinde yer alır. Bilgilerin kişi tarafından toplanıp bir sentez haline getirilmesiyle ortaya çıkan bir olgudur. ■ Yetenek, tecrübe gibi kişisel nitelikler birer bilgelik elemanıdır.
 - Neyin bilindiğinin (bilgi) ve en iyinin ne olduğunu (sosyal ve etnik faktörler) dikkate alınarak en uygun davranışın sergilenebilmesi demektir. Belirli bir alanı veya alanları anlamak için daha geniş ve genelleştirilmiş kuralları ve şemaları temsil etmesiyle bilgiden ayrıılır.
- Bilgelik bilginin teferruatlı ve hassas kullanımını gerektirir. Bilgelik karar alma ve kararın uygulanması sırasında tecrübe edilir.

Bilgi Piramidi

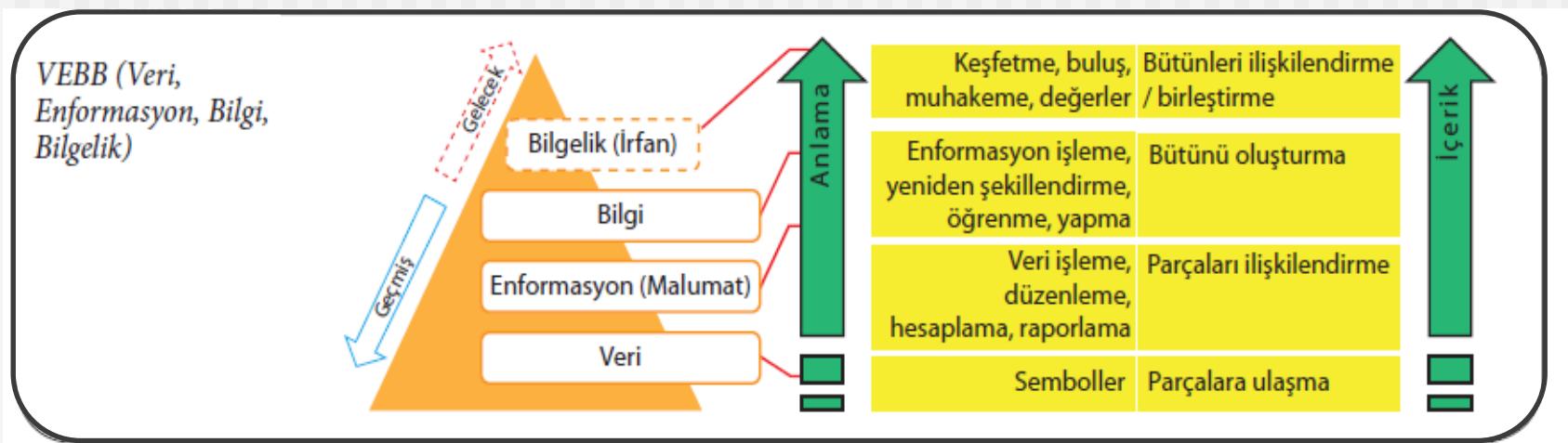
Bilgi piramidi hiyerarşisi incelenecək olursa bilgiye ulaşmanın kolay olmadığı görülür. Yeni teknolojiler enformasyona ulaşmayı daha kolay hale getirmektedir buna karşın, doğru ve güvenilir, yeterli enformasyona ulaşmak zordur. Eğer ulaşılan enformasyon hatalı ya da eksik ise doğal olarak elde edilecek bilgi ve uygulama sonuçları da sağlıklı olmayacağındır.



Bilgi Piramidi

- Bilgeliğe ulaşabilmek için geçilmesi gereken yollar bilgi piramidinin aşamalarına benzemektedir. Veriden bilgeliğe kadar olan yükselme sırasında, gözlemlerden iletişime varan boyutlarda değişiklik gerekmekte ve bilge olana sağlanacak değerin buradan çıkacağı varsayılmaktadır. Bilgelik için gereken şartlara bakıldığından ise, hem bağlam hem de anlayış açısından, gerçekleştirilmesi gereken bir bakış açısının ortaya çıktığı görülmektedir. Bilgelik, deneyimlerin düşünme becerilerine dahil edilmesi ile oluşmaktadır

Veri, Enformasyon, Bilgi, Bilgelik Piramidi (Bilgi Piramidinin Geliştirilmiş Hali)



Kaynak: Temel Bilgi Teknolojileri-I , AÖF Yayıncı

Veri Ambarı

- **Veritabanı:** birbirleriyle ilişkili bilgilerin depolandığı alanlardır.
- **Veri Ambarı:** ilişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işlemel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar.
- **Data Mart:** veri ambarlarının alt kümeleridir. Veri ambarları bir iş probleminin tamamına yönelik bir bakış sağlarken, data mart'lar sadece belli bir kısma bakış sağlarlar. Veri pazarları ile veriye hızlı erişim sağlayabiliriz. İkinci olarak, verinin gruplanmamış yapıda olması ve farklı iş birimlerinin farklı verileri görmesidir. Bu da bize gereksiz bir iş yükü ve güvenlik sorununa neden olmaktadır. İşte tam bu noktada, veri pazarları konuya, bölümlere uygun, veri ambarının küçük bir kopyası halinde çözüm sunmaktadır.

Veri Ambarı

- Amaca yönelik
- Birleştirilmiş
- Zaman değişkenli
- Değişken değil

Veri Ambarları: Amaca Yönelik

- Müşteri, ürün, satış gibi belli konular için düzenlenebilir.
- Verinin incelenmesi ve modellenmesi için oluşturulur.
- Konuya ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar.

Veri Ambarları: Birleştirilmiş

- Veri kaynaklarının birleştirilmesiyle oluşturulur.
 - Canlı veri tabanları, dosyalar.
- Veri temizleme ve birleştirme teknikleri kullanılır.
 - Değişik veri kaynakları arasındaki tutarlılık sağlanır.

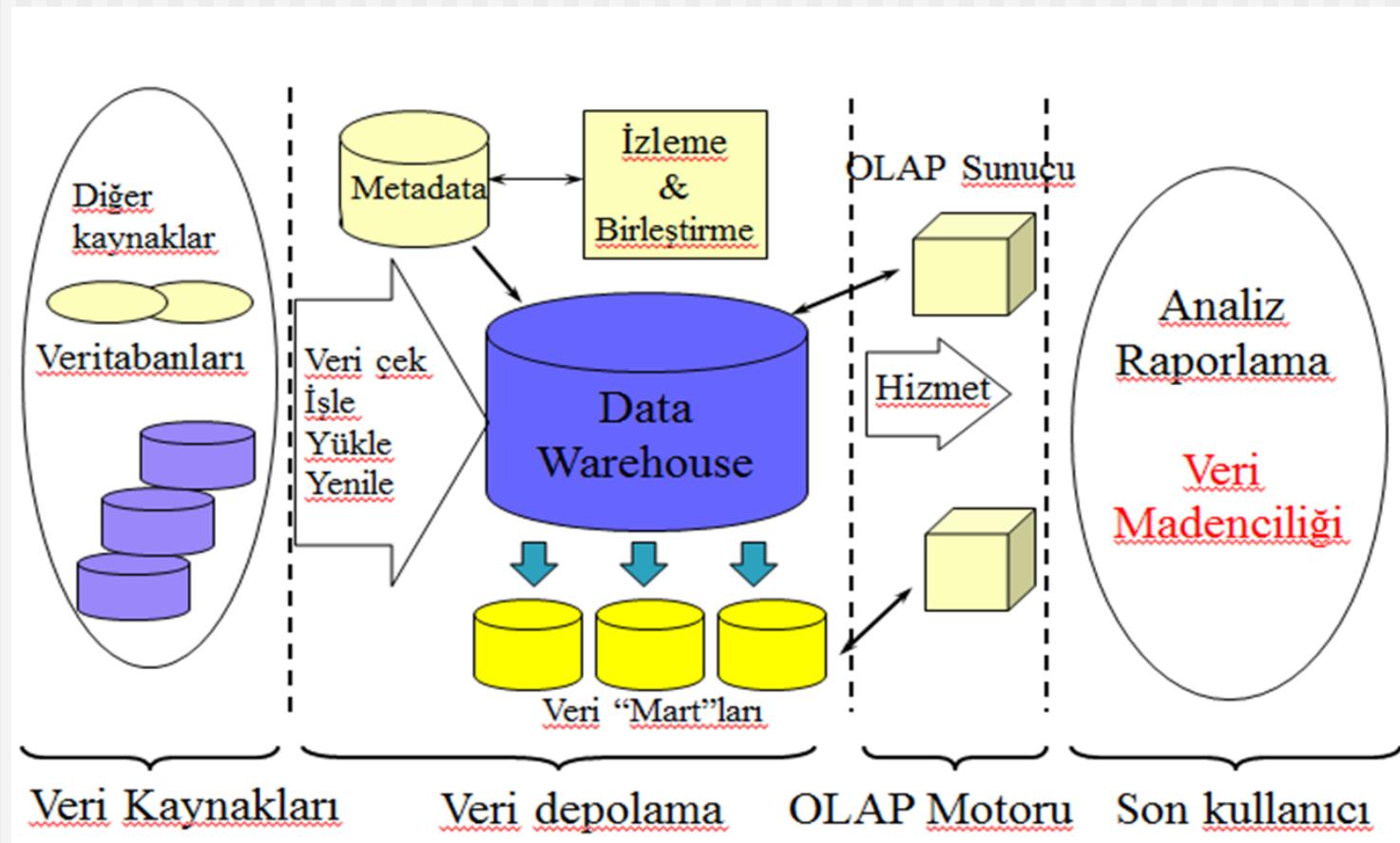
Veri Ambarları: Zaman Değişkenli

- Zaman değişkeni canlı veri tabanlarına göre daha uzundur.
 - Canlı veri tabanları: Güncel veriler bulunur (en çok geçmiş 1 yıl)
 - Veri ambarları: Geçmiş hakkında bilgi verir (geçmiş 5-10 yıl)

Veri Ambarları: Değişken Değil

- Canlı veritabanlarından alınmış verinin fiziksel olarak başka bir ortamda saklanması.
- Canlı veritabanlarındaki değişimin veri ambarlarını etkilememesi.

Veri Ambarı Mimarisi

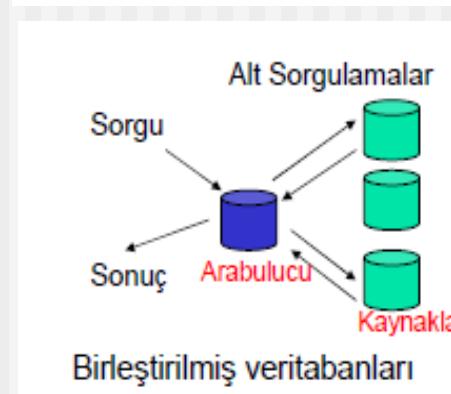
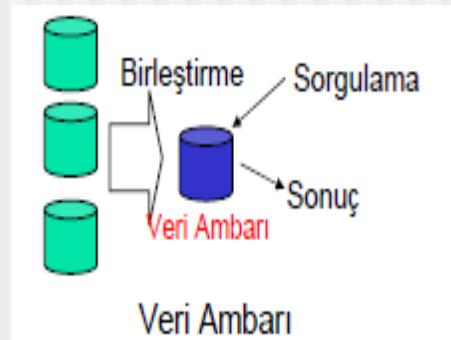


Veri Kaynakları

- İki yaklaşım:
 - sorgulamalı
 - veri ambarı

Veri Ambarı & Birleşmiş Veritabanları

- Veritabanlarının birleştirilmesi:
 - Farklı veritabanları arasında bir arabulucu katman
 - Sorgulamalı
 - Bir sorgulamayı her veritabanı için alt sorgulamalara ayır
 - Sonucu birleştir
- Veri ambarı:
 - Veri daha sonra kullanılmak üzere birleştirilip veri ambarında saklanıyor.



Veri Madenciliği & OLAP

- OLAP (On-Line Analytical Processing)
 - Veri ambarlarının işlevi
 - Veriyi inceleme ve karar verme
 - OLTP (On-Line Transaction Processing) saatler sürebilen işlemler
- OLAP avantajları
 - Daha geniş kapsamlı sonuçlar
 - Daha kısa süreli işlem
- OLAP dezavantajları
 - Kullanıcı neyi nasıl soracağını bilmesi gerekiyor
 - Genelde veriden istatistiksel inceleme yapmak için kullanılır.
- OLAP NE sorusuna cevap verir, veri madenciliği NEDEN sorusuna cevap verir.

Veri Madenciliğinde Sorunlar (1/3)

- Gizlilik ve sosyal haklar
 - Kişilere ait verilerin toplanarak, kişilerden habersiz ve izinsiz olarak kullanılması
 - Veri madenciliği yöntemleri ile bulunan sonuçların izinsiz olarak açıklanması (/paylaşılması)
 - Gizlilik ve veri madenciliği politikalarının düzenlenmesi
- Kullanıcı Arabirimi
 - Görüntüleme
 - Sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi
 - Bilginin sunulması
 - Etkileşim
 - Veri madenciliği ile elde edilen bilginin kullanılması
 - Veri madenciliği yöntemine müdahale etmek
 - Veri madenciliği yönteminin sonucuna müdahale etmek
- Veri madenciliği yöntemi
- Başarım ve ölçüklenebilirlik

Veri Madenciliğinde Sorunlar (2/3)

- Veri madenciliği yöntemi
 - Farklı tipte veriler üzerinde çalışabilme
 - Farklı seviyelerde kullanıcı ile etkileşim halinde olabilme
 - Uygulama ortamı bilgisini kullanabilme
 - Veri madenciliği ile elde edilen sonucu anlaşılır şekilde sunabilme
 - Gürültülü ve eksik veri ile çalışabilme (ve iyi sonuç verebilme)
 - Değişen veya eklenen verileri kolayca kullanabilme
 - Örütü değerlendirme: önemli örüntüler bulma

Veri Madenciliğinde Sorunlar (3/3)

- Başarım ve ölçülebilirlik
 - Kullanabilirlik ve ölçülebilirlik
 - Zaman karmaşıklığı ve yer karmaşıklığı kabul edilebilir
 - Örnekleme yapabilme
 - Paralel ve dağıtık yöntemler
 - Artımlı veri madenciliği
 - Parçala ve çöz
- Veri kaynağı

VERİ MADENCİLİĞİ

(Veri Önüşleme-1)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
 - Sınıflandırma
 - Kümeleme
 - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

Veri Önisleme

- Veri
- Veri Önisleme
 - Veriyi Tanıma
 - Veri temizleme
 - Veri birleştirme
 - Veri dönüşümü
 - Veri azaltma
- Benzerlik ve farklılık

VERİ ÖNİŞLEME

Veri Nedir?

- Nesneler ve nesnelerin niteliklerinden oluşan küme
 - kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir
 - bir insanın yaşı, ortamın sıcaklığı...
 - boyut (dimension), özellik (feature, characteristic) olarak da kullanılır.
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolan dimesi
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Boşanmış	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Boşanmış	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
 - aynı nitelik farklı değer kümelerinden değer alabilir
 - ağırlık: kg, lb(libre, ağırlık ölçüsü)
 - farklı nitelikler aynı değer kümelerinden değer alabilirler
 - ID, yaş: her ikisi de sayısal

İstatistiksel Veri Türleri

- **1- Nümerik Veriler** : Sayısal-Nümerik-Nicel Veriler de denmektedir. Boy, Yaş gibi süreklilik arzeden değerler Nümerik verilerdir. "Daha fazla" ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:
 - a) Sürekli Nümerik Veriler: Yaşı, Sıcaklık
 - b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı
- **2-Nominal Veriler** : Kategorik bir veri çeşididir. "Daha fazla" ifadesi ile kullanılmazlar. İkiye ayrılır:
 - a) Binary Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı
 - b) İkiiden Çok Kategorili: Medeni Durum-Renk-Irk-Şehir, İsim, Forma Numarası
 - Örneğin forma numarası oyuncunun seviyesi ile ilgili bir bilgi içermez.
- **3-Ordinal Veriler** : Ordinal veriler de yine kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. "Daha fazla" ifadesi ile kullanılabilirler ancak ne kadar daha fazla olduğunun ölçüsünü veremezler. Örneğim: Eğitim Düzeyi, Sosyoekonomik ölçek skorları gibi. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.
- **4-Ratio Veriler** : Nümerik verilere benzerler. 100 santigrat derece, 50 santigrat derecenin iki katı denilemez ama derece kelvine çevrilirse 60 kelvin 30 kelvinin 2 misli sıcak denilebilir. Oran verilebilir veri türlerine Ratio veriler denir. Burada kelvin derece ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir.

Nitelik Türleri

- Belli aralıkta yer alan değişkenler (interval)
 - sıcaklık, tarih
- İkili değişkenler (binary)
 - cinsiyet
- Ayrık ve sıralı değişkenler (nominal, ordinal, ratio scaled)
 - göz rengi, posta kodu

Problem

- Gerçek uygulamalarda toplanan veri kirli
 - eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
 - meslek = " "
 - gürültülü: hatalar var
 - maaş= "-10"
 - tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz
 - yaş= "35", d.tarihi: "03/10/2004"
 - önceki oylama değerleri: "1,2,3", yeni oylama değerleri: "A,B,C"
 - bir kaynakta nitelik değeri 'ad', diğerinde 'isim'

Verinin Gürültülü Olma Nedenleri

- Eksik veri kayıtlarının nedenleri
 - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
 - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
 - İnsan, yazılım ya da donanım problemleri
- Gürültülü (hatalı) veri kayıtlarının nedenleri
 - Hatalı veri toplama gereçleri
 - İnsan, yazılım ya da donanım problemleri
 - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
 - Verinin farklı veri kaynaklarında tutulması
 - İşlevsel bağımlılık kurallarına uyulmaması

Sonuç

- Veri güvenilmez
 - Veri madenciliği sonuçlarına güvenilebilir mi?
 - Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir.
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.

Veri Önisleme

■ Veri temizleme

- Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıklarını giderme

■ Veri birleştirme

- Farklı veri kaynağındaki verileri birleştirme

■ Veri dönüşümü

- Normalizasyon ve biriktirme

■ Veri azaltma

- Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

Veriyi Tanıma

Veriyi Tanımlayıcı Özellikler

- Amaç: Veriyi daha iyi anlamak
 - Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım
- Verinin dağılım özellikleri
 - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
 - verinin dağılımı
 - kutu grafiği çizimi ve sıklık derecesi incelemesi

Merkezi Eğilimi Ölçme

Ortalama:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

- ağırlıklı ortalama
- kırpılmış ortalama: Uç değerleri kullanmadan hesaplama

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

Ortanca (median): Verinin tümü kullanılarak hesaplanır

- veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerin ortalaması

Mod

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum f)_l}{f_{\text{median}}} \right) c$$

- Veri içinde en sıklıkla görülen değer
- Unimodal, bimodal, trimodal
- deneysel formül: $\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$

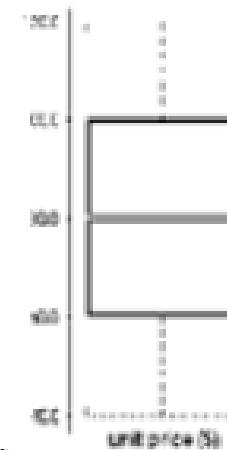
Verinin Dağılımını Ölçme

Çeyrek, aykırılıklar, kutu grafiği çizimi

- Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
 - Q1: ilk %25, Q3: ilk %75
- Dörtlü aralık (Inter-quartile Range): $IQR = Q3 - Q1$
- Five Number Summary: min, Q1, median, Q3, max
- Kutu Grafiği Çizimi:
 - Q1 ve Q3 aralığında bir kutu
 - kutu içinde ortanca noktayı gösteren bir çizgi
 - kutudan min ve max değerlere birer uzantı
- Aykırılıklar: $1,5 \times IQR$ değerinden küçük/büyük olan değerler

Varyans ve standart sapma

$$= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$



Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
 - Eksik nitelik değerlerini tamamlama
 - Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
 - Tutarsızlıkların giderilmesi

Eksik Veri

- Veri için bazı niteliklerin değerleri her zaman bilinemeyebilir.
- Eksik veri
 - diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi
 - bazı nitelik değerleri hatalı olması dolayısıyla silinmesi
 - yanlış anlama sonucu kaydedilmeme
 - veri girişi sırasında bazı nitelikleri önemsiz görme

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
 - Eksik nitelik değerlerini elle doldur
 - Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,...)
 - Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
 - Aynı sınıfı ait kayıtların nitelik değerlerinin ortalaması ile doldur
 - Olasılığı en fazla olan nitelik değeriyle doldur
-

Gürültülü Veri

- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
 - hatalı veri toplama gereçleri
 - veri girişi problemleri
 - veri iletimi problemleri
 - teknolojik kısıtlar
 - nitelik isimlerinde tutarsızlık

Gürültülü Veri nasıl düzeltılır?

- Gürültüyü yok etme

- Bölmeleme

- veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür

- Kümeleme

- aykırılıkları belirler

- Eğri uydurma

- veriyi bir fonksiyona uydurarak gürültüyü düzeltir.

Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
 - Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür
 - Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.
 - her bölge ortalamaya ya da bölmenin en alt ve üst sınırlarıyla temsil edilir.

Bölme genişliği: 3

1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

Ortalamayla düzeltme:

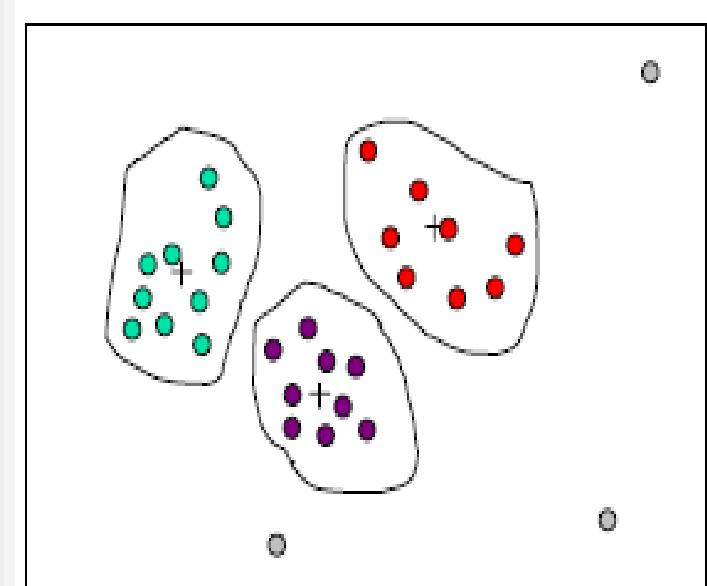
1. Bölme: 9, 9, 9
2. Bölme: 22, 22, 22
3. Bölme: 29, 29, 29

Alt-üst sınırla düzeltme:

1. Bölme: 4, 4, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 25, 34

Kümeleme

- Benzer veriler aynı kümede olacak şekilde gruplanır
- Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.

Veri Birleştirme

Veri Birleştirme

- Farklı kaynaklardan verilerin tutarlı olarak birleştirilmesi
- Şema birleştirilmesi
 - Aynı varlıkların saptanması
 - meta veri kullanılır
- Nitelik değerlerinin saptanması tutarsızlığının
- Aynı nitelik için farklı kaynaklarda farklı değerler olması
- Farklı metrikler kullanılması

Gereksiz Veri

Farklı veri kaynaklarından veriler birleştirilince gereksiz (fazla) veri oluşabilir

- aynı nitelik farklı kaynaklarda farklı isimle
- bir niteliğin değeri başka bir nitelik kullanılarak hesaplanabilir
 - korelasyon hesaplaması: sayısal nitelikler
 - =0: nitelikler bağımsız, >0: pozitif korelasyon, <0: negatif korelasyon

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} \quad \bar{A} = \frac{\sum A}{n} \quad \sigma_A = \sqrt{\frac{\sum (A - \bar{A})^2}{n-1}}$$

- korelasyon hesaplaması: aynık nitelikler (chi-square test)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
- Seçilen algoritmaya uygun olmayabilir
 - Veri belirleyici değil
- Çözüm
 - Veri düzeltme
 - Bölmeleme
 - Kümeleme
 - Eğri Uydurma
 - Biriktirme
 - Genelleme
 - Normalizasyon
 - Nitelik oluşturma

Normalizasyon

- min-max normalizasyon
- z-score normalizasyon
- ondalık normalizasyon

VERİ MADENCİLİĞİ

(Veri Ön İşleme-2)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
 - Sınıflandırma
 - Kümeleme
 - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

Veri Önisleme

- Veri
- Veri Önisleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
- Seçilen algoritmaya uygun olmayabilir
 - Veri belirleyici değil
- Çözüm
 - Veri düzeltme
 - Bölmeleme
 - Kümeleme
 - Eğri Uydurma
 - Biriktirme
 - Genelleme
 - Normalizasyon
 - Nitelik oluşturma

Normalizasyon

- min-max normalizasyon ■ ondalık normalizasyon
- min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.
- z-score normalizasyon ■ Ondalık ölçekte ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilebilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçekte menin formülü aşağıdaki şekildedir:
- z-Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin herhangi bir değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir.

Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new_max_A - new_min_A) + new_min_A$$

- z-score normalizasyon

$$v' = \frac{v - mean_A}{stand_dev_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j} \quad j: \text{Max}(|v'|) < 1 \text{ olacak şekildeki en küçük tam sayı}$$

Nitelik Oluşturma

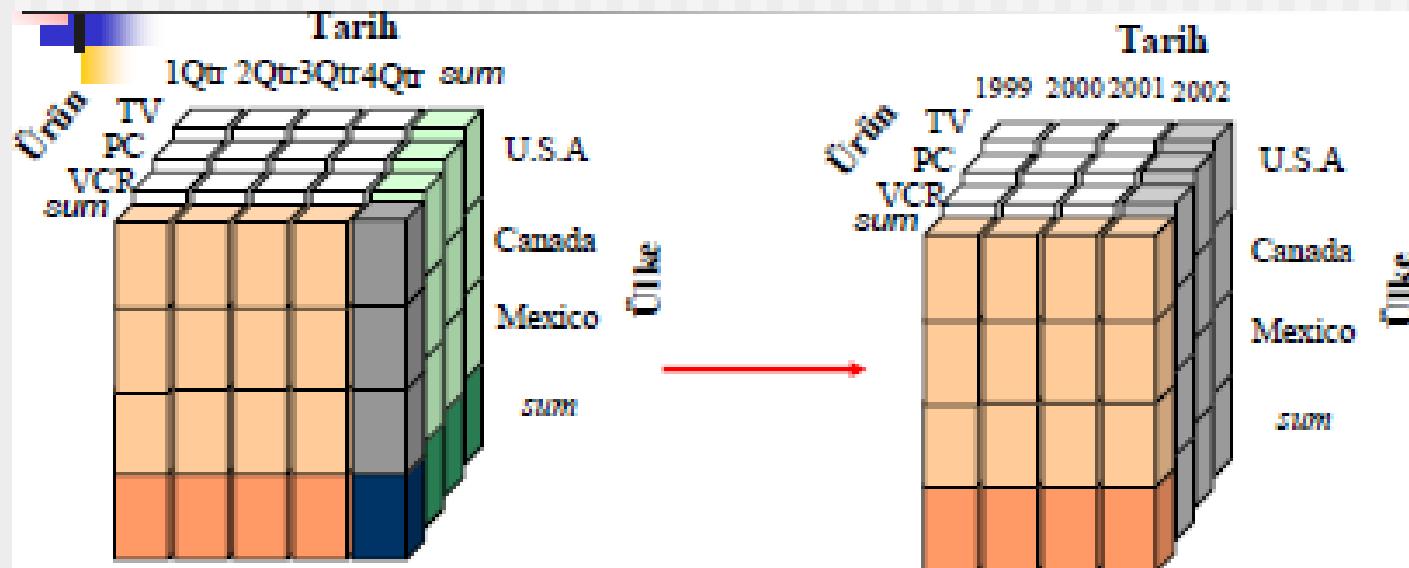
- Yeni nitelikler yarat
 - orjinal niteliklerden daha önemli bilgi içersin
 - alan=boy x en
 - veri madenciliği algoritmalarının başarımı daha iyi olsun

Veri Azaltma

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
 - veriyi azaltma başarımı artırır
 - sonucun (nerdeyse) hiç değişmemesi gereklidir
- Veri azaltma
 - nitelik birleştirme
 - nitelik azaltma
 - veri sıkıştırma
 - veri ayrıştırma ve kavram oluşturma
 - veri küçültme
 - eğri uydurma
 - kümeleme
 - histogram
 - örneklemeye

Nitelik Birleştirme



- Sorgulama için gerekli olan boyutlar kullanılıyor.

Nitelik Seçme - Nitelik Azaltma

■ Nitelik Seçme

- Nitelikler kümesinin bir altkümesi seçilerek veri madenciliği işlemi yapılır.

■ Nitelik azaltma

- d boyutlu veri kümesi $k < d$ olacak şekilde k boyuta taşınır.

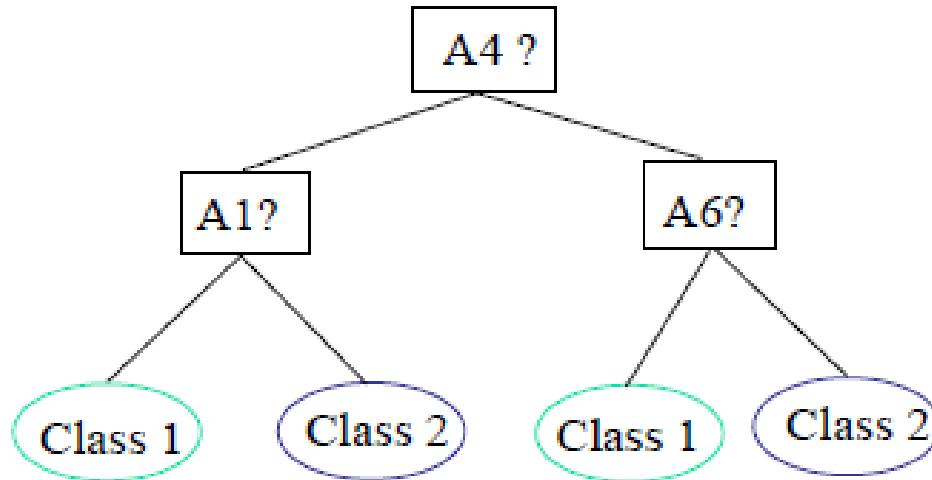
Nitelik Seçme

- Nitelik seçme
 - Veri madenciliği uygulaması için gerekli olan niteliklerin seçilmesi
 - Nitelikler altkümesi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalı
 - Veri madenciliği işlemi yer ve zaman karmaşıklığını azaltma
- Sistemin başarımını artırma
 - Sezgisel yöntemler kullanılarak nitelikler seçilebilir.
 - İstatistiksel anlamlılık testi (statistical significance)
 - bilgi kazancı (information gain)
 - karar ağaçları

Örnek

Başlangıç nitelikler kümesi:

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$



Seçilen nitelik kümesi: $\{A_1, A_4, A_6\}$

Nitelik Azaltma

- Çok boyutlu veriyi daha küçük boyutlu uzaya taşıma
- d nitelikten oluşan n adet veri $D=\{x_1, x_2, \dots, x_n\}$ k boyutlu uzaya taşınır:

$$x_i \in \mathbb{R}^d \rightarrow y_i \in \mathbb{R}^k (k \ll d)$$

- Veri kümesinde yer alan bütün nitelikler kullanılır
 - Niteliklerin doğrusal kombinasyonu
- Niteliklerin ayırcılığına artırma

Veri Sıkıştırma

- Verinin boyutunu azaltır
 - daha az saklama ortamı
 - veriye ulaşmak daha çabuk
- Kayıplı ve kayıpsız veri sıkıştırma
 - bazı yöntemler bazı veri tiplerine uygun
 - her veri tipi için kullanılan yöntemler de var
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

Veri Ayırıştırma

- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır.
- Sürekli bir nitelik değerini bölerek her aralığı etiketler.
- Verinin değeri, bulunduğu aralığın etiketi ile değişir.
- Veri boyutu küçülür.
- Kavram oluşturmak için kullanılır.

Kavram Oluşturma

- Sayısal veriler
 - çok geniş aralıkta olabilir
 - değerleri çok sık değişebilir
- Sayısal veriler için kavram oluşturma
 - bölmeleme
 - histogram
 - kümeleme
 - entropi

Veri Küçültme

- Veriyi farklı şekillerde gösterme
 - parametrik
 - eğri uydurma
 - parametrik olmayan
 - histogram
 - kümeleme
 - örnekleme

Histogram ile Veri Küçültme

- Verinin dağılımı
- Veriyi bölerek her bölüm için veri değerini gösterir (toplam, ortalama)
 - eşit genişlik (equi-width): bölmelerin genişliği eşit
 - eşit yükseklik (equi-height): her bölmedeki veri sayısı eşit
 - v-optimal: en az varyansı olan histogram $\sum(\text{count}_b * \text{value}_b)$
 - MaxDiff: bölme genişliğini kullanıcı belirler

Kümeleme ile Veri Küçültme

- Veri kümelere ayrılır
- Veri kümeleri temsil eden örnekler (küme merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı.
- Hiyerarşik kümeleme yöntemleri kullanılabilir.

Örnekleme ile Veri Küçültme

- Büyük veri kümesini daha küçük bir alt kümeye ile temsil etme
- Alt kümeye nasıl seçiliyor?
 - yerine koymadan örnekleme (SRSWOR)
 - yerine koyarak örnekleme (SRSWR)
 - kümeye örnekleme (yerine koymadan veya koyarak)
 - katman örnekleme (katman: nitelik değerine göre grup)

Benzerlik ve Farklılık

Benzerlik ve Farklılık

■ Benzerlik

- iki nesnenin benzerliğini ölçen sayısal değer
- nesneler birbirine daha benzer ise daha büyük
- genelde 0-1 aralığında değer alır

■ Farklılık

- iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
- nesneler birbirine daha benzer ise daha küçük
- en küçük farklılık genelde 0
- üst sınır değişebilir.

Uzaklık Çeşitleri

- Öklid(Euclid)
- Minkowski
- Manhattan

Öklid Uzaklığı

- Veri kumesi

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Uzaklık matrisi

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

- Öklid uzaklığı (Euclidean Distance) nesneler arası farklılığı bulmak için kullanılır.
 - p adet niteliği (boyutu) olan i ve j nesneleri arasındaki uzaklık

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

q: positif tam sayı

- q=1 → Manhattan uzaklığı

Uzaklık Özellikleri

- $q=1 \Rightarrow$ Manhattan Uzaklığı
- $q=2 \Rightarrow$ Öklid Uzaklığı
- Uzaklık ölçütünün sağlaması gereken özellikler:
 1. $d(i,j) \geq 0$
 2. $d(i,i) = 0$
 3. $d(i,j) = d(j,i)$
 4. $d(i,j) \leq d(i,h) + d(h,j)$
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

$$d(i, j) = \sqrt{w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + \dots + w_p |x_{ip} - x_{jp}|^2}$$

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
- 1. $\text{sim}(i,j) \geq 0$
- 2. $\text{sim}(i,j) = \text{sim}(j,i)$

İkili Değişkenler Arası Benzerlik

- İkili bir değişkenin 0 veya 1 olarak iki değeri olabilir.
- Bir olasılık tablosu oluşturulur:

		Nesne j	
		0	1
Nesne i	0	M_{00}	M_{01}
	1	M_{10}	M_{11}

M_{00} : i nesnesinin 0, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{10} : i nesnesinin 1, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{01} : i nesnesinin 0, j nesnesinin 1 olduğu niteliklerin sayısı
 M_{11} : i nesnesinin 1, j nesnesinin 1 olduğu niteliklerin sayısı

- Yalın uyum katsayısı (simple matching coefficient): ikili değişkenin simetrik olduğu durumlarda

$$sim(i, j) = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard katsayısı (İkili değişkenin asimetrik olduğu durumlar):

$$d(i, j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Kosinüs Benzerliği

- d_1 ve d_2 iki doküman. Kosinüs benzerliği

$$\cos(d_1, d_2) = d_1 \bullet d_2 / \|d_1\| \|d_2\|$$

$d_i \bullet d_j$: iki dokümanın vektör çarpımı

$\|d_i\|$: d_i dokümanının uzunluğu

- Örnek

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

VERİ MADENCİLİĞİ

(Sınıflandırma Yöntemleri)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
 - Sınıflandırma
 - Kümeleme
 - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

İçerik

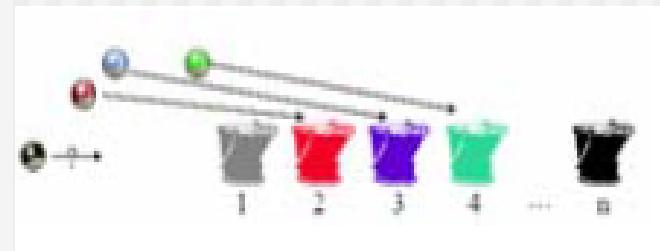
- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcılar
 - Bayes ağları

Sınıflandırma (Classification)

- Sınıflandırma (classification) problemi:
 - nesnelerden oluşan veri kümesi (**öğrenme kümesi**):
 $D=\{t_1, t_2, \dots, t_n\}$
 - her nesne niteliklerden oluşuyor, niteliklerden biri **sınıf bilgisi**
- Sınıf niteliğini belirlemek için diğer nitelikleri kullanarak bir **model** bulma
- Öğrenme kümesinde yer almayan nesneleri (**test kümesi**) mümkün olan en iyi şekilde doğru sınıflara atamak
- sınıflandırma=ayrık değişkenler için öngörüde (prediction) bulunmak.

Sınıflandırma (Classification)

- Amaç: Yeni bir kayıt geldiğinde, bu kaydı geliştirilen modeli kullanılarak mümkün olduğunca doğru bir sınıfa atamak.
 - verinin dağılımına göre bir model bulunur
 - bulunan model, başarımı belirlendikten sonra niteliğin gelecekteki
 - ya da bilinmeyen değerini tahmin etmek için kullanılır
 - model başarımı: doğru sınıflandırılmış sınıma kümesi örneklerinin oranı
- Veri madenciliği uygulamasında:
 - ayrık nitelik değerlerini tahmin etmek: sınıflandırma
 - sürekli nitelik değerlerini tahmin etmek: öngörü



- Sınıflandırma: hangi topun hangi sepete koyulabileceği
- Öngörü: Topun ağırlığı

Danışmanlı & Danışmansız Öğrenme

- Danışmanlı (Gözetimli, Supervised) öğrenme=
sınıflandırma
 - Sınıfların sayısı ve hangi nesnenin hangi sınıfıfta olduğu biliniyor.



- Danışmansız (Gözetimsiz, Unsupervised) öğrenme=
kümeleme (clustering)
 - Hangi nesnenin hangi sınıfıfta olduğu bilinmiyor. Genelde sınıf sayısı bilinmiyor.



Sınıflandırma Uygulamaları

- Kredi başvurusu değerlendirme
- Kredi kartı harcamasının sahtekarlık olup olmadığına karar verme
- Hastalık teşhisini
- Ses tanıma
- Karakter tanıma
- Gazete haberlerini konularına göre ayırma
- Kullanıcı davranışları belirleme

Sınıflandırma için Veri Hazırlama

- Veri dönüşümü:
 - Sürekli nitelik değeri ayrık hale getirilir
 - Normalizasyon $([-1, \dots, 1], [0, \dots, 1])$
- Veri temizleme:
 - gürültüyü azaltma
 - gereksiz nitelikleri silme

Sınıflandırma İşlemi

- Sınıflandırma işlemi üç aşamadan oluşur:
 1. Model oluşturma
 2. Model değerlendirme
 3. Modeli kullanma

Sınıflandırma İşlemi: Model Oluşturma

■ 1. Model Oluşturma:

- Her nesnenin sınıf etiketi olarak tanımlanan niteliğinin belirlediği bir sınıfta olduğu varsayıılır
- Model oluşturmak için kullanılan nesnelerin oluşturduğu veri kümesi öğrenme kümesi olarak tanımlanır
- Model farklı biçimlerde ifade edilebilir
 - IF – THEN – ELSE kuralları ile
 - Karar ağaçları ile
 - Matematiksel formüller ile

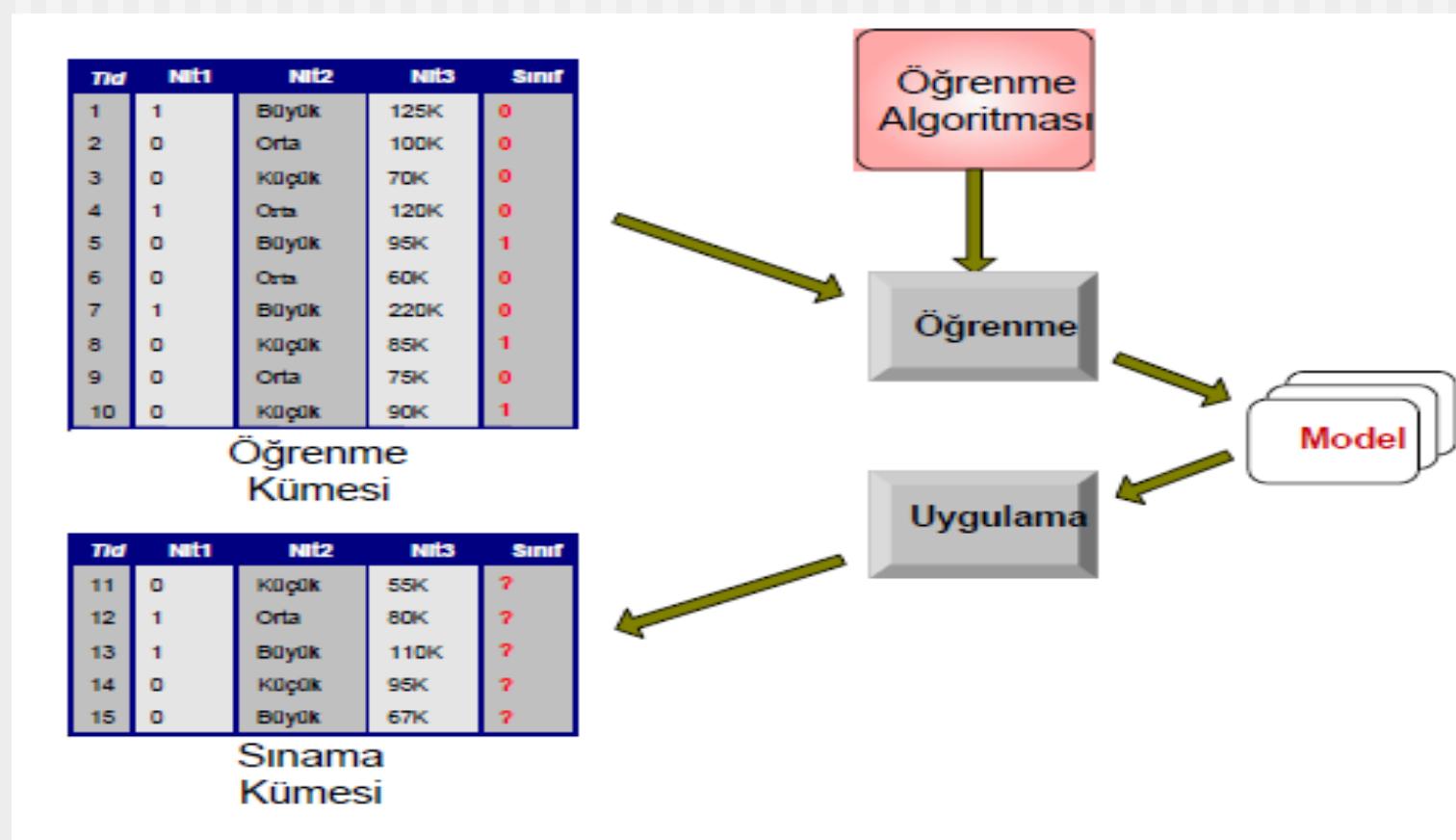
Sınıflandırma İşlemi: Model Değerlendirme

- 2. Model Değerlendirme:
- Modelin başarımı (doğruluğu) sınavma kümesi örnekleri kullanılarak belirlenir.
- Sınıf etiketi bilinen bir sınavma kümesi örneği model kullanılarak belirlenen sınıf etiketiyle karşılaştırılır.
- Modelin doğruluğu, doğru sınıflandırılmış sınavma kümesi örneklerinin toplam sınavma kümesi örneklerine oranı olarak belirlenir.
- Sınavma kümesi model öğrenirken kullanılmaz.

Sınıflandırma İşlemi: Modeli Kullanma

- 3. Modeli kullanma:
 - Model daha önce görülmemiş örnekleri sınıflandırmak için kullanılır
 - Örneklerin sınıf etiketlerini tahmin etme
 - Bir niteliğin değerini tahmin etme

Örnek



Sınıflandırıcı Başarımını Değerlendirme

Doğru sınıflandırma başarısı

- Hız
 - modeli oluşturmak için gerekli süre
 - sınıflandırma yapmak için gerekli süre
- Kararlı olması
 - veri kümesinde gürültülü ve eksik nitelik değerleri olduğu durumlarda da iyi sonuç vermesi
- Ölçeklenebilirlik
 - büyük miktarda veri kümesi ile çalışabilmesi
- Anlaşılabilir olması
 - kullanıcı tarafından yorumlanabilir olması
- Kuralların yapısı
 - birbiriyle örtüşmeyen kurallar

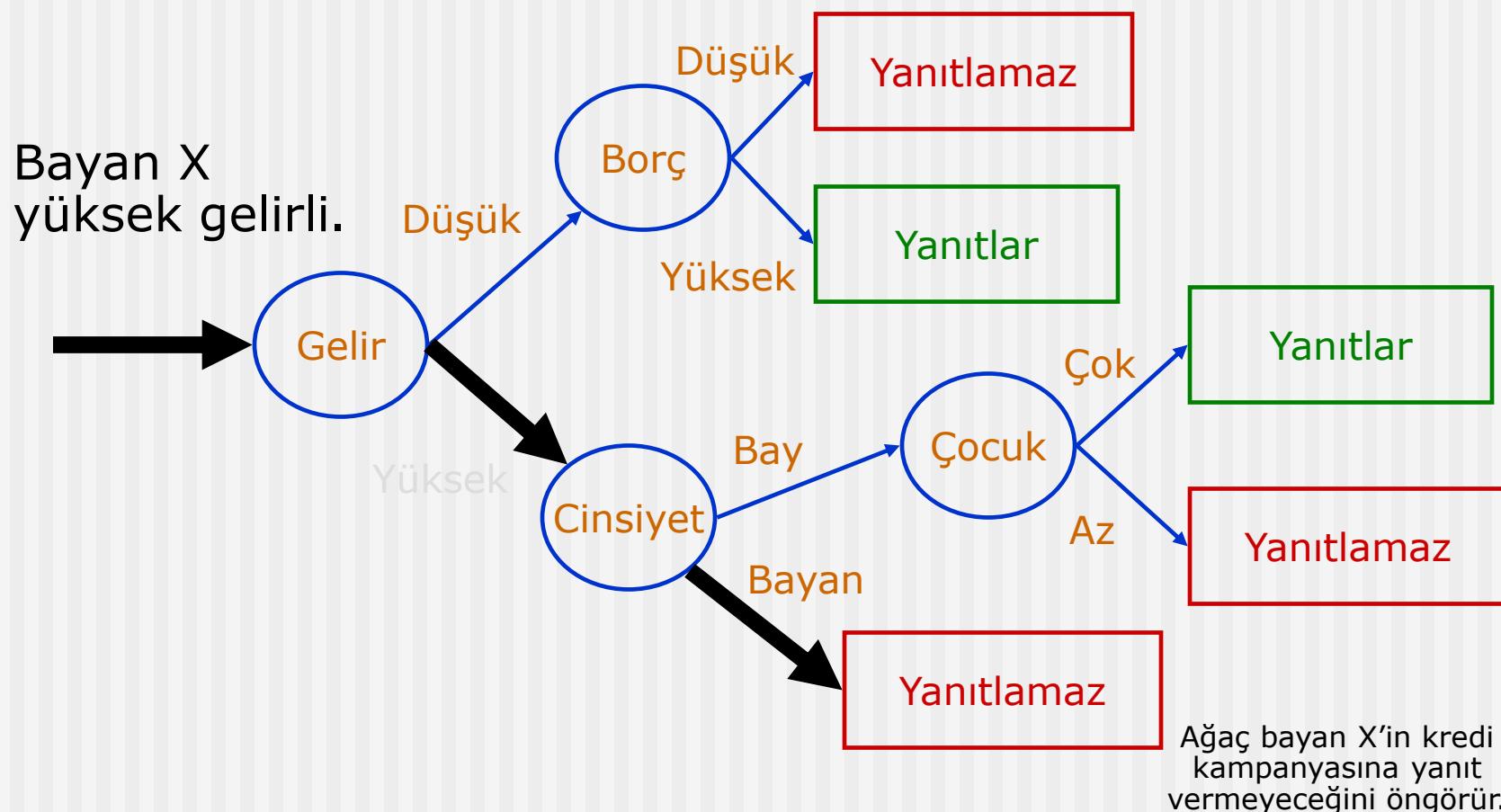
Sınıflandırma Yöntemleri

- Karar ağaçları (decision trees)
- Yapay sinir ağları (artificial neural networks)
- Bayes sınıflandırıcılar (Bayes classifier)
- İlişki tabanlı sınıflandırıcılar (association-based classifier)
- k-en yakın komşu yöntemi (k- nearest neighbor method)
- Destek vektör makineleri (support vector machines)
- Genetik algoritmalar (genetic algorithms)
- ...

Karar Ağaçları

- Karar Ağacı
 - Yaygın kullanılan öngörü yöntemlerinden bir tanesidir.
 - Ağaçtaki her düğüm bir özellikteki testi gösterir.
 - Düğüm dalları testin sonucunu belirtir.
 - Ağaç yaprakları sınıf etiketlerini içerir.
- Karar ağacı çıkarımı iki aşamadan oluşur
 - Ağaç inşası
 - Başlangıçta bütün öğrenme örnekleri kök düğümdedir.
 - Örnekler seçilmiş özelliklere tekrarlamalı olarak göre bölünür.
 - Ağaç Temizleme (Budama) (Tree pruning)
 - Gürültü ve istisna kararları içeren dallar belirlenir ve kaldırılır.
- Karar ağacı kullanımı: Yeni bilinmeyen örneğin sınıflandırılması
 - Bilinmeyen örneğin özellikleri karar ağacında test edilerek sınıfı bulunur.

Bir Kredi Kartı Kampanyasında Yeni Bir Örneğin Sınıflandırılması



Karar Ağacı Yöntemleri

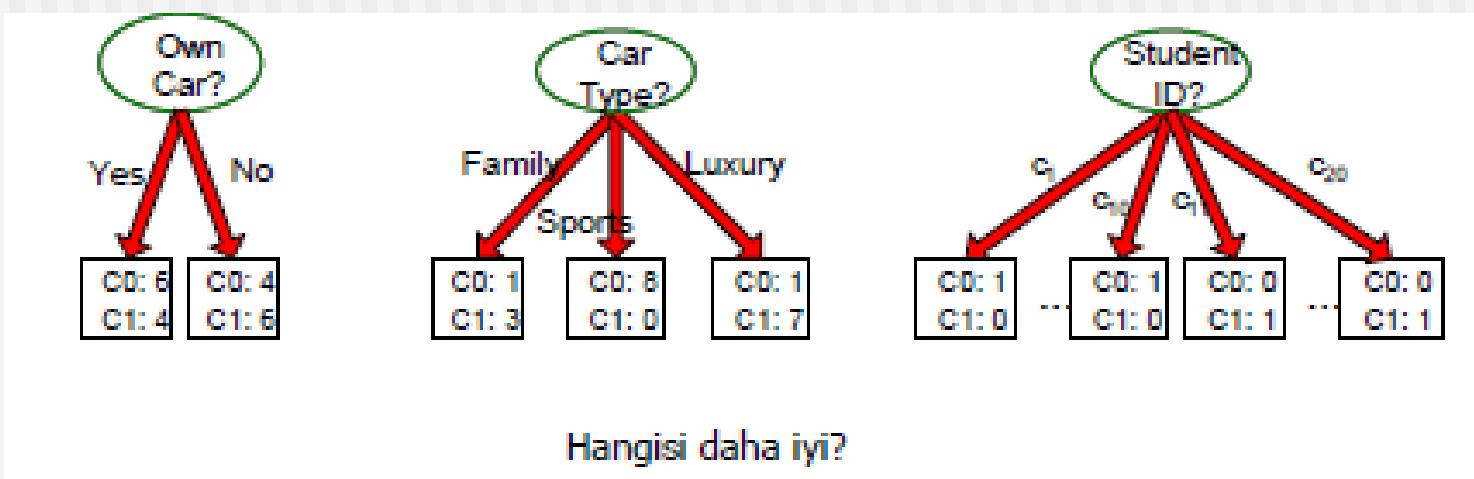
- Karar ağacı oluşturma yöntemleri genel olarak iki aşamadan oluşur:
 - 1. ağaç oluşturma
 - en başta bütün öğrenme kümesi örnekleri kökte
 - seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor.
 - 2. ağaç budama
 - öğrenme kümesindeki gürültülü verilerden ve sınıma kümesinde hataya neden olan dalları silme (sınıflandırma başarımını artırır)

Karar Ağacı Oluşturma

- Yinelemeli işlem
 - ağaç bütün verinin oluşturduğu tek bir düğümle başlıyor
 - eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor
 - eğer değilse örnekleri sınıflara en iyi bölecek olan nitelik seçiliyor
 - işlem sona eriyor
 - örneklerin hepsi (çoğunluğu) aynı sınıfa ait
 - örnekleri bölecek nitelik kalmamış
 - kalan niteliklerin değerini taşıyan örnek yok

Örnekleri En İyi Bölten Nitelik Hangisi?

- Bölmeden önce:
 - 10 örnek C0 sınıfında
 - 10 örnek C1 sınıfında



En iyi Bölme Nasıl Belirlenir?

- “Greedy” (aç gözlü) yaklaşım
 - çoğunlukla aynı sınıfı ait örneklerin bulunduğu düğümler tercih edilir
- Düğümün kalitesini ölçmek için bir yöntem



En İyi Bölen Nitelik Nasıl Belirlenir?

- İyilik Fonksiyonu (Goodness Function)
- Farklı algoritmalar farklı iyilik fonksiyonları kullanabilir:
 - bilgi kazancı (information gain): ID3, C4.5
 - bütün niteliklerin ayrık değerler aldığı varsayılıyor
 - sürekli değişkenlere uygulamak için değişiklik yapılabilir
 - gini index (IBM IntelligentMiner)
 - her nitelik ikiye bölünüyor
 - her nitelik için olası bütün ikiyi bölmeler sınavılıyor

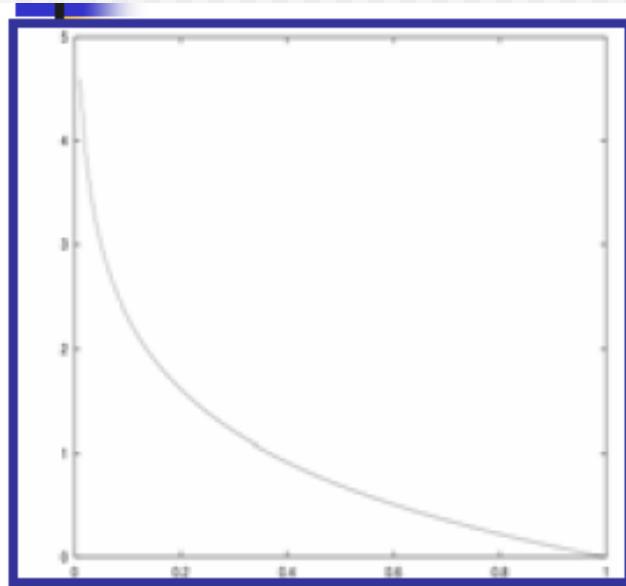
Bilgi / Entropi

- p_1, p_2, \dots, p_s toplamları 1 olan olasılıklar. Entropi (Entropy)

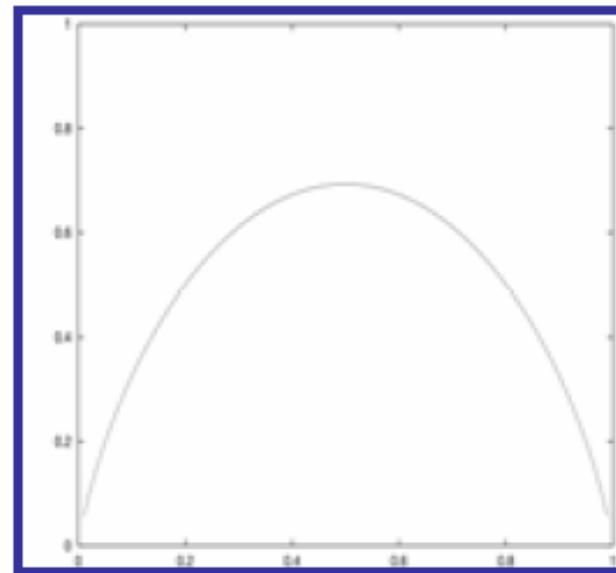
$$H(p_1, p_2, \dots, p_s) = -\sum_{i=1}^s p_i \log(p_i)$$

- Entropi rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığını gösterir.
- Sınıflandırmada
 - olayın olması beklenen bir durum
 - entropi=0

Entropi



$\log(p)$



$H(p, 1-p)$

- örnekler aynı sınıfı aitse entropi=0
- örnekler sınıflar arasında eşit dağılmışsa entropi=1
- örnekler sınıflar arasında rastgele dağılmışsa $0 < \text{entropi} < 1$

Örnek

- S veri kümesinde 14 örnek: C0 sınıfına ait 9, C1 sınıfına ait 5 örnek.
- Entropi
- $H(p_1, p_2, \dots, p_s) = -\sum_{i=1}^s p_i \log(p_i)$
- $H(p_1, p_2) = - (9/14) \log_2 (9/14) - (5/14) \log_2 (5/14)$
 $= 0.940$

Bilgi Kazancı (ID3 / C4.5)

- Bilgi kuramı kavramlarını kullanarak karar ağacı oluşturulur. Sınıflandırma sonucu için en az sayıda karşılaştırma yapmayı hedefler.
- Ağaç bir niteliğe göre dallandığında entropi ne kadar düşer?
- A niteliğinin S veri kümesindeki bilgi kazancı

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$Values(A)$, A niteliğinin alabileceği değerler, S_v , $A=v$ olduğu durumda S 'nin altkümesi.

VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
 - Sınıflandırma
 - Kümeleme
 - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

İçerik

■ Sınıflandırma yöntemleri

■ Karar ağaçları ile sınıflandırma

- Entropi Kavramı
 - ID3 Algoritması
 - C4.5 Algoritması
- 
- Entropiye dayalı algoritmalar

- Twoing Algoritması
 - Gini Algoritması
- 
- Sınıflandırma ve regresyon ağaçları (CART)

- k-en yakın komşu algoritması
- 
- Bellek tabanlı algoritmalar

Karar Ağaçları ile Sınıflandırma

- Sınıflandırma problemleri için yaygın kullanılan yöntemdir.
- Sınıflandırma doğruluğu diğer öğrenme metotlarına göre çok etkindir.
- Öğrenmiş sınıflandırma modeli ağaç şeklinde gösterilir ve karar ağaçları (decision tree) olarak adlandırılır.
- Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak en üst yapı kök ve bunların arasında kalan yapılar dal olarak isimlendirilir.

Karar Ağaçlarında Dallanma Kriterleri

- Karar ağaçlarında en önemli sorunlardan birisi hangi kökten itibaren bölümlemenin veya dallanmanın hangi kriter'e göre yapılacağıdır. Aslında her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir.
- Bu algoritmalar şu şekilde grüplendirilebilir.
 - ID3 ve C4.5, entropiye dayalı sınıflandırma algoritmalarıdır.
 - Twoing ve Gini, CART (Classification And Regression Trees) sınıflandırma ve regresyon ağaçlarına dayalı sınıflandırma algoritmalarıdır.
 - k-en yakın komşu algoritması bellek tabanlı sınıflandırma yöntemleri arasında yer almaktadır.

Entropi

(1/3)

- Entropi, rastgele değere sahip bir değişken veya bir sistem için belirsizlik ölçütüdür.
- Enformasyon, rassal bir olayın gerçekleşmesi halinde ortaya çıkan bilgi ölçütüdür.
- Bir süreç için entropi, tüm örnekler tarafından içerenen enformasyonun beklenen değeridir.
- Eşit olasılık durumlara sahip sistemler yüksek belirsizliğe sahiptirler.
- Shannon, bir sistemdeki durum değişikliğinde, entropideki değişimin enformasyon boyutunu tanımladığını öne sürmüştür.
- Buna göre bir sistemdeki belirsizlik arttıkça, bir durum gerçekleştiğinde elde edilecek enformasyon boyutu da artacaktır.

Entropi

(2/3)

- Shannon bilgiyi bitlerle ifade ettiği için, logaritmayı 2 tabanında kullanmıştır.
- S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \dots, m_n\}$ olmak üzere n mesaj üretildiğini varsayalım. Tüm mesajlar birbirinden bağımsız üretilmektedir ve m_i mesajlarının üretilme olasılıkları p_i 'dir. $P = \{p_1, p_2, \dots, p_n\}$ olasılık dağılımına sahip mesajları üreten S kaynağının entropisi $H(S)$ şu şekildedir.

$$H(S) = - \sum_{i=1}^n p_i \log_2(p_i)$$

Entropi

(3/3)

- Bir paranın havaya atılması olayı rassal X sürecini göstersin. Yazı ve tura gelme olasılıkları eşit olduğundan elde edilecek entropi,

$$H(X) = - \sum_{i=1}^n p_i \log_2(p_i) = - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1$$

Örnek

- Aşağıdaki 8 elemanlı S kümesi verilsin.
- $S = \{\text{evet}, \text{hayır}, \text{evet}, \text{hayır}, \text{hayır}, \text{hayır}, \text{hayır}, \text{hayır}\}$
- “evet” ve “hayır” için olasılık,

- $p(\text{evet}) = \frac{2}{8}, p(\text{hayır}) = \frac{6}{8}$

$$H(S) = -\left(\frac{2}{8} \log_2 \frac{2}{8} + \frac{6}{8} \log_2 \frac{6}{8}\right) = 0.81128$$

ID3 Algoritması

(1/4)

- Karar ağaçları yardımcıla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir. Bunlar arasında ID3 ve C4.5 algoritması yer almaktadır.
- ID3(Iterative Dichotomiser 3) algoritması sadece *kategorik* verilerle çalışmaktadır.
- Karar ağaçları çok boyutlu veriyi belirlenmiş bir niteliğe göre parçalara böler.
- Her adımda verinin hangi özelliğine göre ne tür işlem yapılacağına karar verilir.
- Oluşturulabilecek tüm ağaçların kombinasyonu çok fazladır.
- Karar ağaçlarının en az düğüm ve yaprak ile oluşturulması için farklı algoritmalar kullanılarak bölme işlemi yapılır.

ID3 Algoritması

(2/4)

■ Karar AĞACINDA ENTROPI

- Bir eğitim kümesindeki sınıf niteliğinin alacağı değerler kümesi T , her bir sınıf değeri C_i olsun.
- T sınıf değerini içeren küme için P_T sınıfların olasılık dağılımı,

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_k|}{|T|} \right)$$

şeklinde ifade edilir.

- T sınıf kümesi için ortalama entropi değeri ise

$$H(T) = - \sum_{i=1}^n p_i \log_2(p_i)$$

şeklinde ifade edilir.

ID3 Algoritması

(3/4)

- Karar ağaçlarında bölümlemeye hangi düğümden başlanacağı çok önemlidir.
- Uygun düğümden başlanmazsa ağaçın içerisindeki düğümlerin ve yaprakların sayısı çok fazla olacaktır.
- Bir risk kümesi aşağıdaki gibi tanımlansın. $C_1 = \text{"var"}$, $C_2 = \text{"yok"}$
 - $RISK = \{var, var, var, yok, var, yok, yok, var, var, yok\}$

$$|C_1| = 6 \quad |C_2| = 4 \quad p_1 = 6/10 = 0,6 \quad p_2 = 4/10 = 0,4$$

$$P_{RISK} = \left(\frac{6}{10}, \frac{4}{10} \right)$$

$$H(RISK) = - \sum_{i=1}^n p_i \log_2(p_i) = - \left(\frac{6}{10} \log_2 \frac{6}{10} + \frac{4}{10} \log_2 \frac{4}{10} \right) = 0,97$$

■ Dallanma için niteliklerin seçimi

- Öncelikle sınıf niteliğinin entropisi hesaplanır.

$$H(T) = -\sum_{i=1}^n p_i \log_2(p_i)$$

- Sonra özellik vektörlerinin sınıfa bağımlı entropileri hesaplanır.

$$H(X_k) = -\sum_{i=1}^n \frac{|T_i|}{|X_k|} \log \frac{|T_i|}{|X_k|} \quad H(X, T) = \sum_{k=1}^n \frac{|X_k|}{|X|} H(X_k)$$

- Son olarak sınıf niteliğinin entropisinden tüm özellik vektörlerinin entropisi çıkartılarak her özellik için kazanç ölçüyü hesaplanır.

$$\text{Kazanç}(X, T) = H(T) - H(X, T)$$

- En büyük kazanca sahip özellik vektörü o iterasyon için dallanma düğümü olarak seçilir.

Örnek

- Aşağıdaki tablo için karar ağacı oluşturulsun.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(T) = H(RISK) = -\sum_{i=1}^n p_i \log_2(p_i) = -\left(\frac{5}{10} \log_2 \frac{5}{10} + \frac{5}{10} \log_2 \frac{5}{10} \right) = 1$$

Örnek

$$H(BORÇ_{YÜKSEK}) = -\left(\frac{3}{3}\log_2 \frac{3}{3} + \frac{0}{3}\log_2 \frac{0}{3}\right) = 0$$

$$H(BORÇ_{DUSUK}) = -\left(\frac{5}{7}\log_2 \frac{5}{7} + \frac{2}{7}\log_2 \frac{2}{7}\right) = 0,863$$

$$\begin{aligned} H(BORÇ, RISK) &= \frac{3}{10}H(BORÇ_{YÜKSEK}) + \frac{7}{10}H(BORÇ_{DUSUK}) \\ &= \frac{3}{10}(0) + \frac{7}{10}(0,863) = 0,64 \end{aligned}$$

$$Kazanç(BORÇ, RISK) = 1 - 0,64 = 0,36$$

Örnek

$$H(GELIR_{YÜKSEK}) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) = 0,971$$

$$H(GELIR_{DUSUK}) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0,971$$

$$\begin{aligned} H(GELIR, RISK) &= \frac{5}{10} H(GELIR_{YÜKSEK}) + \frac{5}{10} H(GELIR_{DUSUK}) \\ &= \frac{5}{10} (0,971) + \frac{5}{10} (0,971) = 0,971 \end{aligned}$$

$Kazanç(GELIR, RISK) = 1 - 0,971 = 0,029$

Örnek

$$H(STATU_{ISVEREN}) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0,971$$

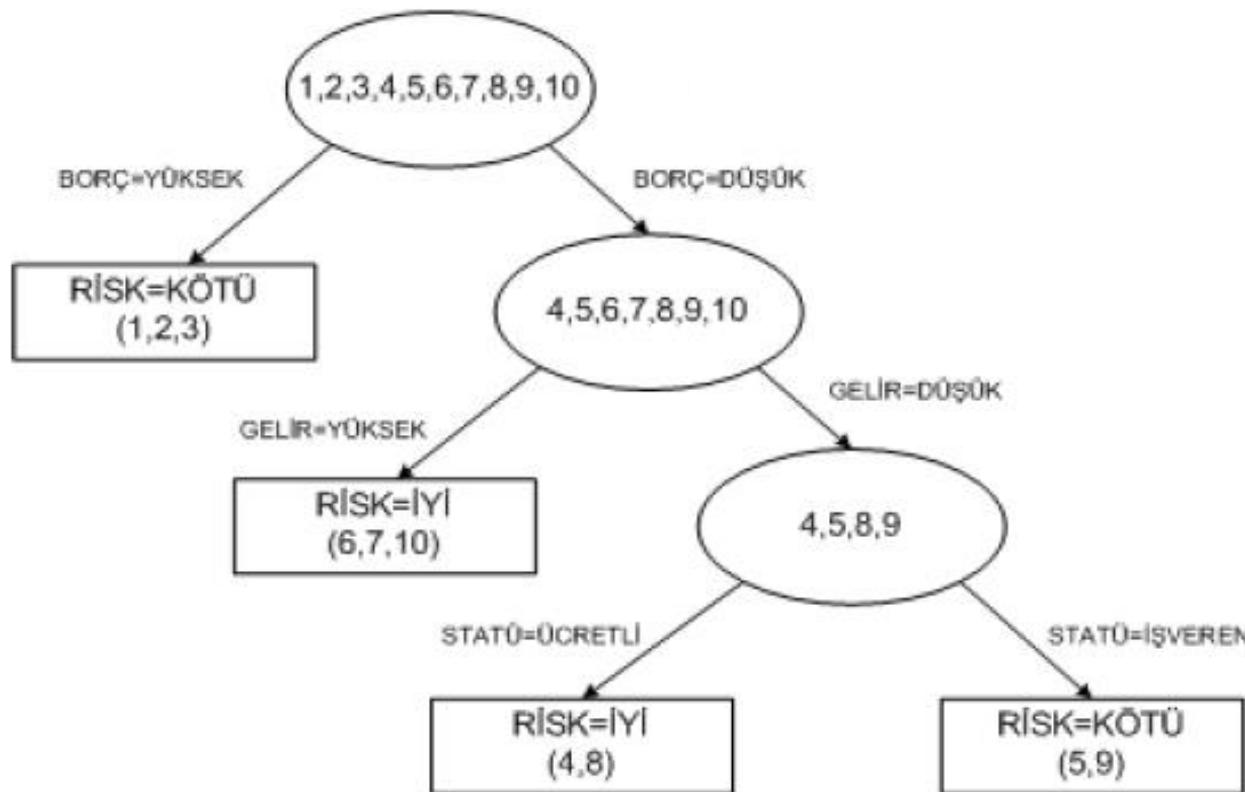
$$H(STATU_{DUSUK}) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0,971$$

$$\begin{aligned} H(STATU, RISK) &= \frac{5}{10} H(STATU_{YÜKSEK}) + \frac{5}{10} H(STATU_{DUSUK}) \\ &= \frac{5}{10} (0,971) + \frac{5}{10} (0,971) = 0,971 \end{aligned}$$

$$Kazanç(STATU, RISK) = 1 - 0,971 = 0,029$$

İlk dallanma için uygun seçim BORÇ niteliğidir.

Örnek



Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(BORÇ = YÜKSEK) **İSE** (RİSK = KÖTÜ)
- **2.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = YÜKSEK) **İSE** (RİSK = İYİ)
- **3.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = ÜCRETLİ) **İSE** (RİSK = İYİ)
- **4.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = İŞVEREN) **İSE**(RİSK = KÖTÜ)

Uygulama: Hava problemi örneği

Eğitim kümesi				
HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	Hayır
güneşli	sıcak	yüksek	kuvvetli	Hayır
bulutlu	sıcak	yüksek	hafif	Evet
yağmurlu	ılık	yüksek	hafif	Evet
yağmurlu	soğuk	normal	hafif	Evet
yağmurlu	soğuk	normal	kuvvetli	Hayır
bulutlu	soğuk	normal	kuvvetli	Evet
güneşli	ılık	yüksek	hafif	Hayır
güneşli	soğuk	normal	hafif	Evet
yağmurlu	ılık	normal	hafif	Evet
güneşli	ılık	normal	kuvvetli	Evet
bulutlu	ılık	yüksek	kuvvetli	Evet
bulutlu	sıcak	normal	hafif	Evet
yağmurlu	ılık	yüksek	kuvvetli	Hayır

Uygulama: Hava problemi

- $OYUN = \{hayır, hayır, hayır, hayır, hayır, evet, evet, evet, evet, evet, evet, evet\}$
- C1, sınıfı "**hayır**", C2, sınıfı ise "**evet**"
- $P1=5/14$, $P2=9/14$

$$H(OYUN) = - \left(\frac{5}{14} \log_2 \frac{5}{14} + \frac{9}{14} \log_2 \frac{9}{14} \right) = 0.940$$

Adım1: Birinci dallanma

ISI niteliği için kazanç ölçütü:

$$|ISI_{so\check{g}uk}| = 4$$

$$|ISI_{ihk}| = 6$$

$$|ISI_{sicak}| = 4$$

$$H(X, T) = \sum_{i=1}^n \frac{|T_i|}{|T|} H(T_i)$$

$$H(ISI, OYUN) = \frac{4}{14} H(ISI_{so\check{g}uk}) + \frac{6}{14} H(ISI_{ihk}) + \frac{4}{14} H(ISI_{sicak})$$

ISI	OYUN
so\check{g}uk	evet
so\check{g}uk	hayır
so\check{g}uk	evet
so\check{g}uk	evet
ihk	evet
ihk	hayır
ihk	evet
ihk	evet
ihk	evet
ihk	hayır
sicak	hayır
sicak	hayır
sicak	evet
sicak	evet

$$H(ISI_{so\check{g}uk}) = -\left(\frac{1}{4} \log_2 \frac{1}{4} + \frac{3}{4} \log_2 \frac{3}{4}\right) = 0.811$$

$$H(ISI_{ihk}) = -\left(\frac{2}{6} \log_2 \frac{2}{6} + \frac{4}{6} \log_2 \frac{4}{6}\right) = 0.918$$

$$H(ISI_{sicak}) = -\left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4}\right) = 1.00$$

$$H(ISI, OYUN) = \frac{4}{14}(0.811) + \frac{6}{14}(0.918) + \frac{4}{14}(1.00) = 0.911$$

$$\begin{aligned} \text{Kazanç}(ISI, OYUN) &= H(OYUN) - H(ISI, OYUN) \\ &= 0.940 - 0.911 = 0.029 \end{aligned}$$

Adım1: Birinci dallanma

HAVA niteliği için kazanç ölçütü:

$$|HAVA_{\text{güneşli}}| = 5 \quad |HAVA_{\text{yağmurlu}}| = 5 \quad |HAVA_{\text{bulutlu}}| = 4$$

$$H(HAVA, OYUN) = \frac{5}{14} H(HAVA_{\text{güneşli}}) + \frac{4}{14} H(HAVA_{\text{bulutlu}}) + \frac{5}{14} H(HAVA_{\text{yağmurlu}})$$

$$H(HAVA_{\text{güneşli}}) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0.971$$

$$H(HAVA_{\text{yağmurlu}}) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) = 0.971$$

$$H(HAVA_{\text{bulutlu}}) = -\left(\frac{4}{4} \log_2 \frac{4}{4}\right) = 0$$

HAVA	OYUN
güneşli	hayır
güneşli	hayır
güneşli	hayır
güneşli	evet
güneşli	evet
yağmurlu	evet
yağmurlu	evet
yağmurlu	hayır
yağmurlu	evet
yağmurlu	hayır
bulutlu	evet
bulutlu	evet
bulutlu	evet
bulutlu	evet

$$H(HAVA, OYUN) = \frac{5}{14} H(HAVA_{\text{güneşli}}) + \frac{4}{14} H(HAVA_{\text{bulutlu}}) + \frac{5}{14} H(HAVA_{\text{yağmurlu}})$$

$$H(HAVA, OYUN) = \frac{5}{14} (0.971) + \frac{4}{14} (0) + \frac{5}{14} (0.971) = 0.694$$

$$\begin{aligned} \text{Kazanç}(HAVA, OYUN) &= H(OYUN) - H(HAVA, OYUN) \\ &= 0.940 - 0.693 = 0.247 \end{aligned}$$

Adım1: Birinci dallanma

NEM niteliği için kazanç ölçütü:

$$|NEM_{yüksek}| = 7$$

$$|NEM_{normal}| = 7$$

$$H(NEM, OYUN) = \frac{7}{14} H(NEM_{yüksek}) + \frac{7}{14} H(NEM_{normal})$$

$$H(NEM_{yüksek}) = -\left(\frac{4}{7} \log_2 \frac{4}{7} + \frac{3}{7} \log_2 \frac{3}{7}\right) = 0.985$$

$$H(NEM_{normal}) = -\left(\frac{1}{7} \log_2 \frac{1}{7} + \frac{6}{7} \log_2 \frac{6}{7}\right) = 0.592$$

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	evet
yüksek	evet
yüksek	hayır
yuksek	evet
yüksek	hayır
normal	evet
normal	hayır
normal	evet
norma!	evet
normal	evet
normal	evet
normal	evet

$$H(NEM, OYUN) = \frac{7}{14} H(NEM_{yüksek}) + \frac{7}{14} H(NEM_{normal})$$

$$H(NEM, OYUN) = \frac{7}{14} (0.985) + \frac{7}{14} (0.592) = 0.789$$

$$\begin{aligned}Kazanç(NEM, OYUN) &= H(OYUN) - H(NEM, OYUN) \\&= 0.940 - 0.789 = 0.151\end{aligned}$$

Adım1: Birinci dallanma

RÜZGAR niteliği için kazanç ölçütü:

$$|RÜZGAR_{hafif}| = 8$$

$$|RÜZGAR_{kuvvetli}| = 6$$

$$H(RÜZGAR, OYUN) = \frac{8}{14} H(RÜZGAR_{hafif}) + \frac{6}{14} H(RÜZGAR_{kuvvetli})$$

$$H(RÜZGAR_{hafif}) = -\left(\frac{2}{8} \log_2 \frac{2}{8} + \frac{6}{8} \log_2 \frac{6}{8}\right) = 0.811$$

$$H(RÜZGAR_{kuvvetli}) = -\left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6}\right) = 1.00$$

RÜZGAR	OYUN
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır
kuvvetli	evet
kuvvetli	evet
kuvvetli	evet
kuvvetli	hayır

$$H(RÜZGAR, OYUN) = \frac{8}{14} H(RÜZGAR_{hafif}) + \frac{6}{14} H(RÜZGAR_{kuvvetli})$$

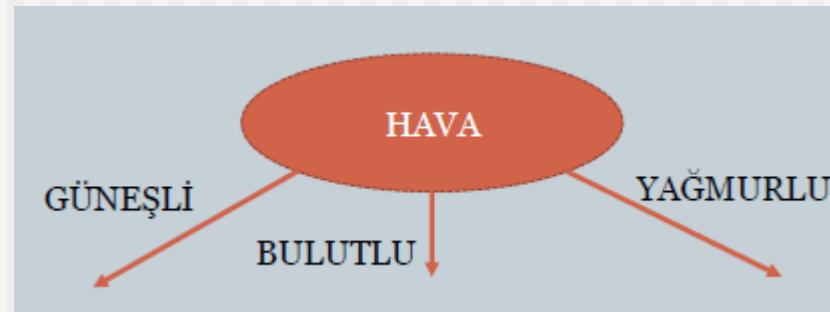
$$H(RÜZGAR, OYUN) = \frac{8}{14} (0.811) + \frac{6}{14} (1.00) = 0.892$$

$$\begin{aligned} \text{Kazanç}(RÜZGAR, OYUN) &= H(OYUN) - H(OYUN) \\ &= 0.940 - 0.892 = 0.048 \end{aligned}$$

Nitelik	Kazanç
HAVA	0.246
İSİ	0.029
NEM	0.151
RÜZGAR	0.048

Adım1: Birinci dallanma

- Birinci dallanma sonucu karar ağacı:



Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

HAVA=güneşli için gözlem değerleri				
HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	hayır
güneşli	sıcak	yüksek	kuvvetli	hayır
güneşli	ılık	yüksek	hafif	hayır
güneşli	soğuk	normal	hafif	evet
güneşli	ılık	normal	kuvvetli	evet

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

- Oyun için entropi:

$$H(OYUN) = - \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

ISI niteliği için kazanç ölçütü:

$$|ISI_{so\check{g}uk}| = 1$$

$$H(ISI_{so\check{g}uk}) = - \left(\frac{1}{1} \log_2 \frac{1}{1} \right) = 0$$

$$H(ISI_{sıcak}) = - \left(\frac{2}{2} \log_2 \frac{2}{2} \right) = 0$$

$$H(ISI_{ılık}) = - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1$$

$$H(ISI, OYUN) = \frac{1}{5}(0) + \frac{1}{5}(0) + \frac{1}{5}(1) = 0.4$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.4 = 0.570$$

ISI	OYUN
soğuk	evet
sıcak	hayır
sıcak	hayır
ılık	hayır
ılık	evet

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

NEM niteliği için kazanç ölçüttü:

$$H(NEM_{yüksek}) = -\left(\frac{3}{3} \log_2 \frac{3}{3}\right) = 0$$

$$H(NEM_{normal}) = -\left(\frac{2}{2} \log_2 \frac{2}{2}\right) = 0$$

$$H(NEM, OYUN) = \frac{3}{5}(0) + \frac{2}{5}(0) = 0$$

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	hayır
normal	evet
normal	evet

$$\text{Kazanç}(NEM, OYUN) = H(OYUN) - H(NEM, OYUN) = 0.970 - 0 = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

RÜZGAR niteliği için kazanç ölçütü:

$$H(RÜZGAR_{\text{hafif}}) = - \left(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3} \right) = 0.918$$

$$H(RÜZGAR_{\text{kuvvetli}}) = - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1$$

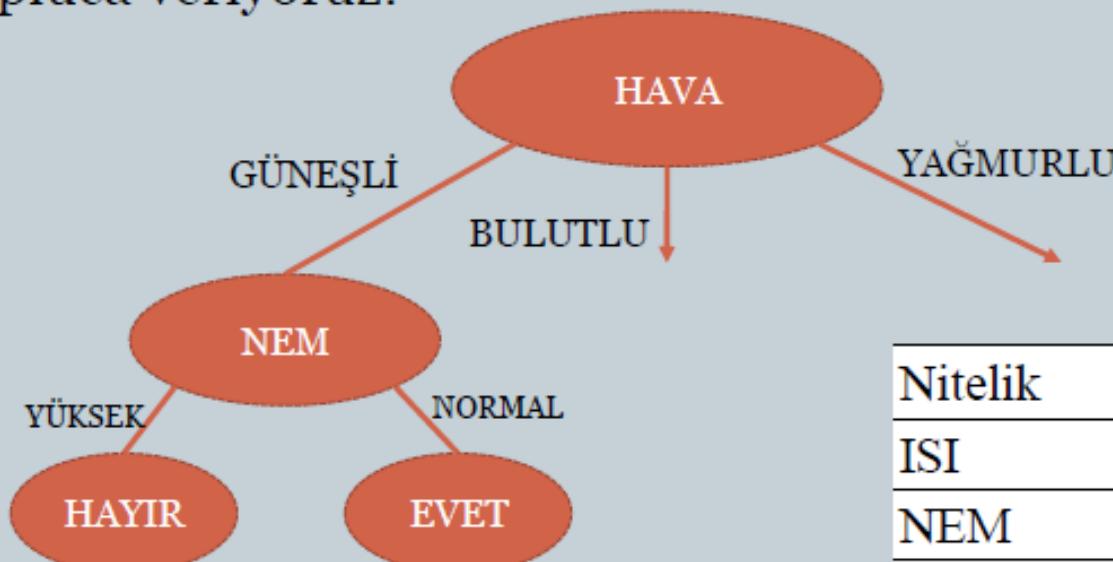
$$H(RÜZGAR, OYUN) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$$

RÜZGAR	OYUN
hafif	hayır
hafif	hayır
hafif	evet
kuvvetli	hayır
kuvvetli	evet

$$\text{Kazanç}(RÜZGAR, OYUN) = H(OYUN) - H(RÜZGAR, OYUN) = 0.970 - 0.951 = 0.019$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

Elde edilen kazanç ölçütlerini aşağıdaki tabloda topluca veriyoruz:



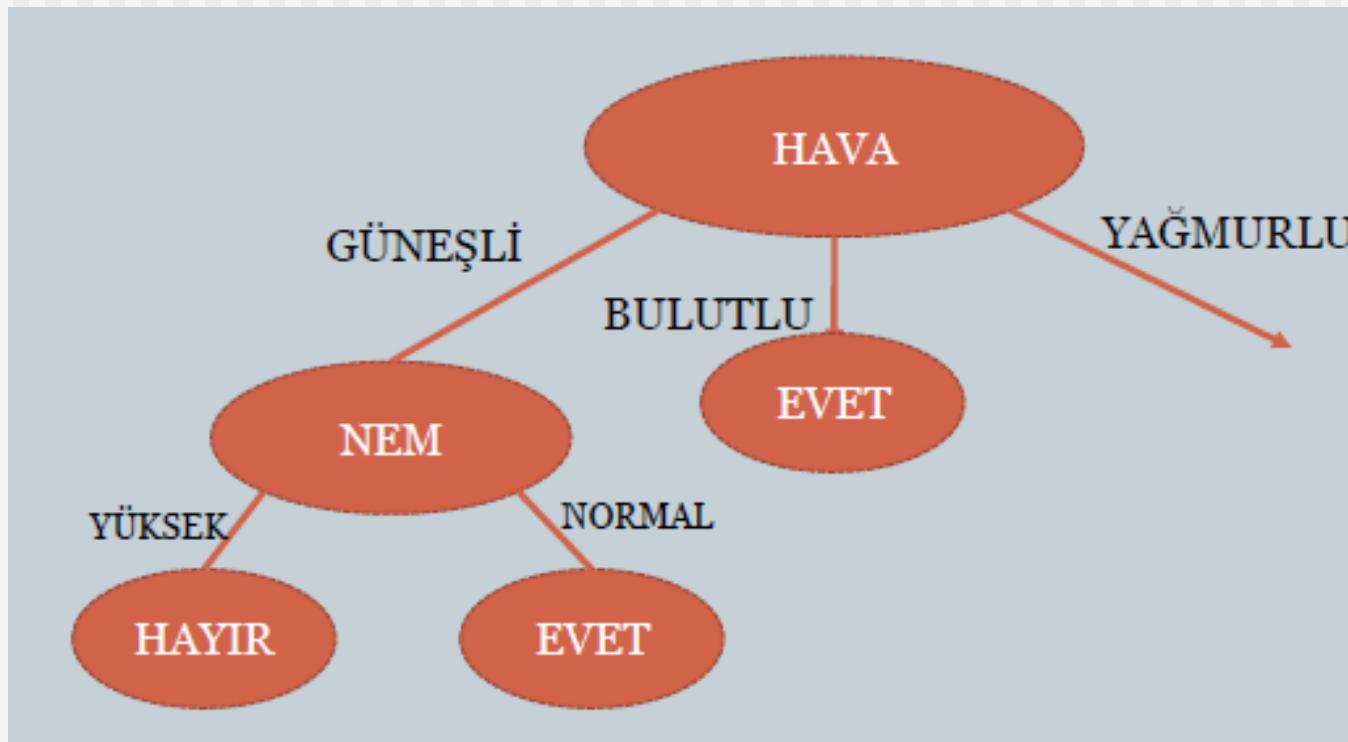
Nitelik	Kazanç
ISI	0.570
NEM	0.970
RÜZGAR	0.019

Adım 3: HAVA niteliğinin “bulutlu” değeri için dallanma:

Görüldüğü gibi tüm karar değerleri "**evet**" olduğu için herhangi bir analize gerek yoktur.

HAVA	ISI	NEM	RÜZGAR	OYUN
bulutlu	sıcak	yüksek	hafif	evet
bulutlu	soğuk	normal	kuvvetli	evet
bulutlu	ılık	yüksek	kuvvetli	evet
bulutlu	sıcak	normal	hafif	evet

Adım 3: HAVA niteliğinin “bulutlu” değeri için dallanma:



Adım 3:HAVA niteliğinin “yağmurlu” değeri için dallanma:

OYUN için entropi:

HAVA	ISI	NEM	RÜZGAR	OYUN
yağmurlu	ılık	yüksek	hafif	evet
yağmurlu	soğuk	normal	hafif	evet
yağmurlu	soğuk	normal	kuvvetli	hayır
yağmurlu	ılık	normal	hafif	evet
yağmurlu	ılık	yüksek	kuvvetli	hayır

$$H(OYUN) = - \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.970$$

Adım 3:HAVA niteliğinin “yağmurlu” değeri için dallanma:

ISI niteliği için kazanç ölçütü:

$$|ISI_{so\check{g}uk}| = 2 \quad |ISI_{\check{ihk}}| = 3$$

$$H(ISI_{so\check{g}uk}) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1$$

$$H(ISI_{\check{ihk}}) = -\left(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}\right) = 0.918$$

$$H(ISI, OYUN) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.951$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.951 = 0.019$$

ISI	OYUN
so\check{g}uk	evet
so\check{g}uk	hayır
\check{ihk}	evet
\check{ihk}	evet
\check{ihk}	hayır

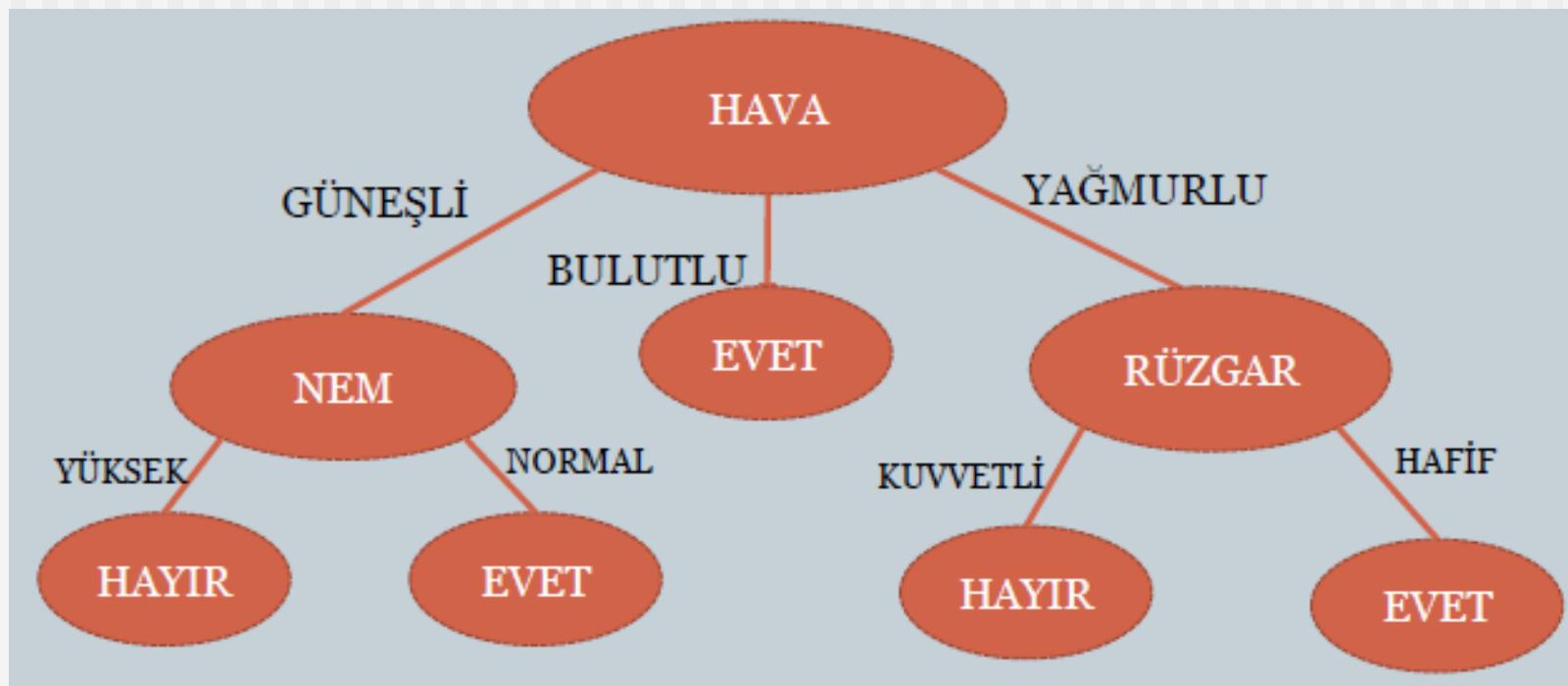
Adım 3:HAVA niteliğinin “yağmurlu” değeri için dallanma:

RÜZGAR niteliği için kazanç ölçütü:

$$|RÜZGAR_{hafif}| = 3 \quad |RÜZGAR_{güçlü}| = 2$$

RÜZGAR	OYUN
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır

Oluşturulan Karar Ağacı



C4.5 Algoritması

- C4.5 ile sayısal değerlere sahip nitelikler için karar ağacı oluşturmak için Quinlan tarafından geliştirilmiştir.
 - ID3 algoritmasından tek farkı nümerik değerlerin kategorik değerler haline dönüştürülmesidir.
 - En büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir.
 - Eşik değeri belirlemek için tüm değerler sıralanır ve ikiye bölünür.
 - Eşik değer için $[v_i, v_{i+1}]$ aralığının orta noktası alınabilir.
$$t_i = \frac{v_i + v_{i+1}}{2}$$
 - Nitelikteki değerler eşik değere göre iki kategoriye ayrılmış olur.
-

Örnek

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

Tabloda örneğe ait eğitim kümesi ele alındığında sayısal değerlere sahip olan **NİTELİK2** niteliğinin seçilmesi durumunda bilgi kazancının bulunması istenmektedir.

Örnek

Eşik değerinin belirlenmesi

- Nitelik 2 = {65, 70, 75, 80, 85, 90, 95, 96} için eşik değer $(80+85)/2 = 83$ alınmıştır.

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	doğru	sınıf1
a	90	doğru	sınıf2
a	85	yanlış	sınıf2
a	95	yanlış	sınıf2
a	70	yanlış	sınıf1
b	90	doğru	sınıf1
b	78	yanlış	sınıf1
b	65	doğru	sınıf1
b	75	yanlış	sınıf1
c	80	doğru	sınıf2
c	70	doğru	sınıf2
c	80	yanlış	sınıf1
c	70	yanlış	sınıf1
c	96	yanlış	sınıf1

NİTELİK2≤ 83
veya
NİTELİK2>83
testi uygulanarak
düzenleme
yapıldığında
yandaki tablo
elde edilir.

Örnek

$$H(SINIF) = -\left(\frac{5}{14} \log_2 \frac{5}{14} + \frac{9}{14} \log_2 \frac{9}{14} \right) = 0,940$$

Entropi değerleri
ve Bilgi kazancı
hesaplanır

$$H(NITELIK1_a) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) = 0,971$$

$$H(NITELIK1_b) = -\left(\frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4} \right) = 0$$

$$H(NITELIK1_c) = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0,971$$

$$\begin{aligned} H(NITELIK1, SINIF) &= \frac{5}{14} H(NITELIK1_a) + \frac{4}{14} H(NITELIK1_b) + \frac{5}{14} H(NITELIK1_c) \\ &= \frac{5}{14} 0,971 + \frac{4}{14} 0 + \frac{5}{14} 0,971 = 0,694 \end{aligned}$$

$$Kazanç(NITELIK1, SINIF) = 0,940 - 0,694 = 0,246$$

Örnek

$$H(NITELIK2_{ek}) = -\left(\frac{7}{9} \log_2 \frac{7}{9} + \frac{2}{9} \log_2 \frac{2}{9}\right) = 0,765$$

$$H(NITELIK2_b) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) = 0,971$$

$$\begin{aligned} H(NITELIK2, SINIF) &= \frac{9}{14} H(NITELIK2_{ek}) + \frac{5}{14} H(NITELIK1_b) \\ &= \frac{9}{14} 0,765 + \frac{5}{14} 0,971 = 0,836 \end{aligned}$$

$Kazanc(NITELIK\ 2, SINIF) = 0,940 - 0,836 = 0,104$

Örnek

$$H(NITELIK3_d) = -\left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6}\right) = 1$$

$$H(NITELIK3_y) = -\left(\frac{6}{8} \log_2 \frac{6}{8} + \frac{2}{8} \log_2 \frac{2}{8}\right) = 0,811$$

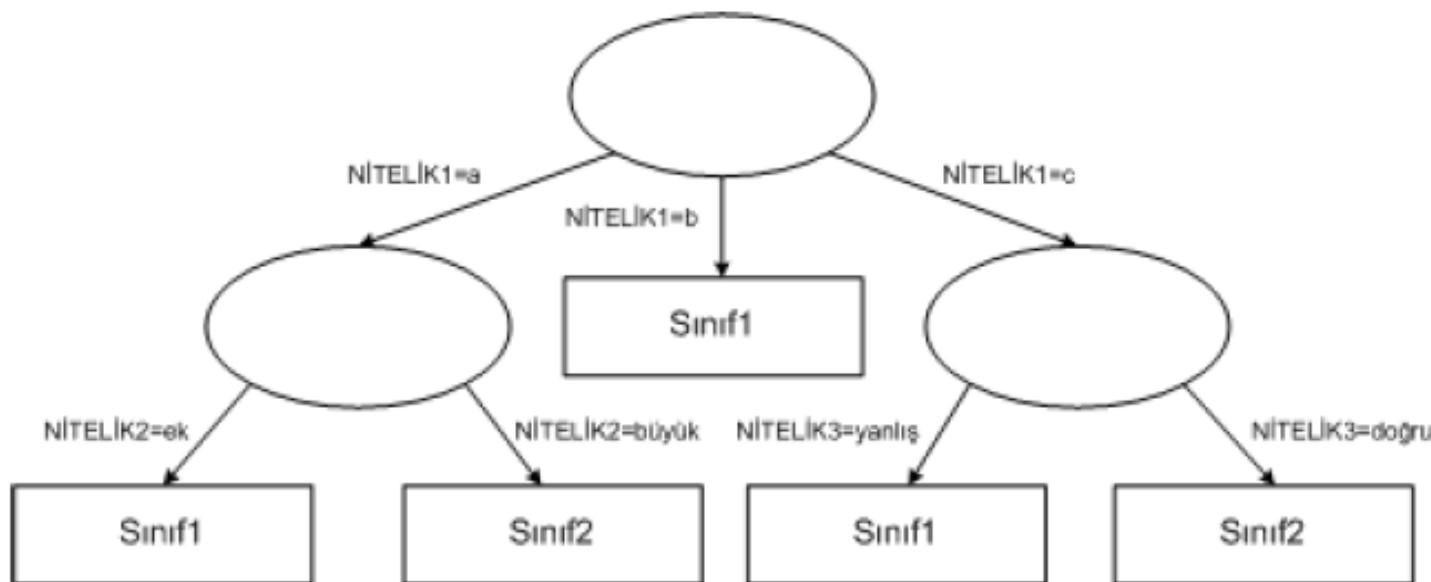
$$\begin{aligned} H(NITELIK3, SINIF) &= \frac{6}{14} H(NITELIK3_d) + \frac{8}{14} H(NITELIK3_y) \\ &= \frac{6}{14} 1 + \frac{8}{14} 0,811 = 0,892 \end{aligned}$$

$$Kazanç(NITELIK3, SINIF) = 0,940 - 0,892 = 0,048$$

$$Kazanç(NITELIK3, SINIF) < Kazanç(NITELIK2, SINIF) < Kazanç(NITELIK1, SINIF)$$

Örnek

Oluşturulan karar ağıacı



Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Eşit veya Küçük) **İSE**(SINIF = Sınıf1)
- **2.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Büyük) **İSE**(SINIF = Sınıf2)
- **3.EĞER**(NİTELİK1 = b) **İSE**(SINIF = Sınıf1)
- **4.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = yanlış) **İSE**(SINIF = Sınıf1)
- **5.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = doğru) **İSE**(SINIF = Sınıf2)

VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

İçerik

■ Sınıflandırma yöntemleri

■ Karar ağaçları ile sınıflandırma

- Entropi Kavramı
 - ID3 Algoritması
 - C4.5 Algoritması
- 
- Entropiye dayalı algoritmalar

- Twoing Algoritması
 - Gini Algoritması
- 
- Sınıflandırma ve regresyon ağaçları (CART)

- k-en yakın komşu algoritması
- 
- Bellek tabanlı algoritmalar

Sınıflandırma ve Regresyon Ağaçları (CART)

- Sınıflandırma ve regresyon ağaçları veri madenciliğinin sınıflandırma ile ilgili konuları arasında yer alır. Bu yöntem 1984'te Breiman tarafından ortaya atılmıştır. CART karar aacı, herbir karar düğümünden itibaren aacıın iki dala ayrılması ilkesine dayanır. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur.
- CART algoritmasında bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler gözönüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu bölünmeler üzerinde seçme işlemi uygulanır. Bu kapsamdaki iki algoritma bulunmaktadır.
 - Twoing Algoritması
 - Gini Algoritması

Twoing Algoritması

- Twoing algoritmasında eğitim kümesi her adımda iki parçaya ayrılarak bölümleme yapılır.
- Aday bölünmelerin sağ ve sol kısımlarının her birisi için nitelik değerinin ilgili sütundaki tekrar sayısı alınır.
- Aday bölünmelerin sağ ve sol kısımlarındaki her bir nitelik değeri için sınıf değerlerinin olma olasılığı hesaplanır.
- Her bölünme için uygunluk değeri en yüksek olan alınır.

$$\Phi(B|d) = 2 \frac{|B_{sol}|}{|T|} \frac{|B_{sag}|}{|T|} \sum_{j=1}^n abs\left(\frac{|Tsinif_j|}{|B_{sol}|} - \frac{|Tsinif_j|}{|B_{sag}|} \right)$$

- Burada, T eğitim kümesindeki kayıt sayısını, B aday bölünmeyi, d düğümü, Tsinif_j ise j.sınıf değerini gösterir.

Örnek

(1/8)

- Tabloda çalışanların maaş, deneyim, görev niteliklerine göre hedef niteliği olan memnun olma durumlarına ait 11 gözlem verilmiştir. Twoing algoritmasını kullanarak sınıflandırma yapınız.

PERSONEL	MAAŞ	DENEYİM	GÖREV	MEMNUN
1	NORMAL	ORTA	UZMAN	EVET
2	YÜKSEK	YOK	UZMAN	EVET
3	DÜŞÜK	YOK	YÖNETİCİ	EVET
4	YÜKSEK	ORTA	YÖNETİCİ	EVET
5	DÜŞÜK	ORTA	YÖNETİCİ	EVET
6	YÜKSEK	İYİ	YÖNETİCİ	EVET
7	DÜŞÜK	İYİ	YÖNETİCİ	EVET
8	YÜKSEK	ORTA	UZMAN	HAYIR
9	DÜŞÜK	ORTA	UZMAN	HAYIR
10	YÜKSEK	İYİ	UZMAN	HAYIR
11	DÜŞÜK	İYİ	UZMAN	HAYIR

Örnek

(2/8)

- Aday bölünmeler aşağıdaki gibidir.

BÖLÜNME	SOL	SAĞ
1	MAAŞ = NORMAL	MAAŞ = {DÜŞÜK, YÜKSEK}
2	MAAŞ = YÜKSEK	MAAŞ = {DÜŞÜK, NORMAL}
3	MAAŞ = DÜŞÜK	MAAŞ = {NORMAL, YÜKSEK}
4	DENEYİM = YOK	DENEYİM = {ORTA, İYİ}
5	DENEYİM = ORTA	DENEYİM = {YOK, İYİ}
6	DENEYİM = İYİ	DENEYİM = {YOK, ORTA}
7	GÖREV = UZMAN	GÖREV = YÖNETİCİ
8	GÖREV = YÖNETİCİ	GÖRE = UZMAN

Örnek

(3/8)

- MAAŞ = NORMAL için

$$P_{sol} = \frac{|B_{sol}|}{|T|} = \frac{1}{11} = 0,09$$

$$P_{(EVET|t_{sol})} = \frac{|T\text{sinif } EVET|}{|B_{sol}|} = \frac{1}{1} = 1$$

$$P_{(HAYIR|t_{sol})} = \frac{|T\text{sinif } HAYIR|}{|B_{sol}|} = \frac{0}{1} = 0$$

BÖLÜNME	B _{sol}	P _{Sol}	sinif _{EVET}	sinif _{HAYIR}	P(EVET t _{Sol})	P(HAYIR t _{Sol})
1	1	0,09	1	0	1	0
2	5	0,45	3	2	0,6	0,4
3	5	0,45	3	2	0,6	0,4
4	2	0,18	2	0	1	0
5	5	0,45	3	2	0,6	0,4
6	4	0,36	2	2	0,5	0,5
7	6	0,55	2	4	0,33	0,67
8	5	0,45	5	0	1	0

Örnek

(4/8)

- MAAŞ = {DÜŞÜK, YÜKSEK} için

$$P_{\text{sag}} = \frac{|B_{\text{sag}}|}{|T|} = \frac{10}{11} = 0,91$$

$$P_{(\text{EVET}|t_{\text{sag}})} = \frac{|\text{Tsinif}_{\text{EVET}}|}{|B_{\text{sag}}|} = \frac{6}{10} = 0,6$$

$$P_{(\text{HAYIR}|t_{\text{sag}})} = \frac{|\text{Tsinif}_{\text{HAYIR}}|}{|B_{\text{sag}}|} = \frac{4}{10} = 0,4$$

BÖLÜNME	B _{sag}	P _{sag}	\text{sınıf}_{\text{EVET}}	\text{sınıf}_{\text{HAYIR}}	P(EVET t _{Sag})	P(HAYIR t _{Sag})
1	10	0,91	6	4	0,6	0,4
2	6	0,55	4	2	0,67	0,33
3	6	0,55	4	2	0,67	0,33
4	9	0,82	5	4	0,56	0,44
5	6	0,55	4	2	0,67	0,33
6	7	0,64	5	2	0,71	0,29
7	5	0,45	5	0	1	0
8	6	0,55	2	4	0,33	0,67

Örnek

(5/8)

Uygunluk değeri (1. aday bölünme için)

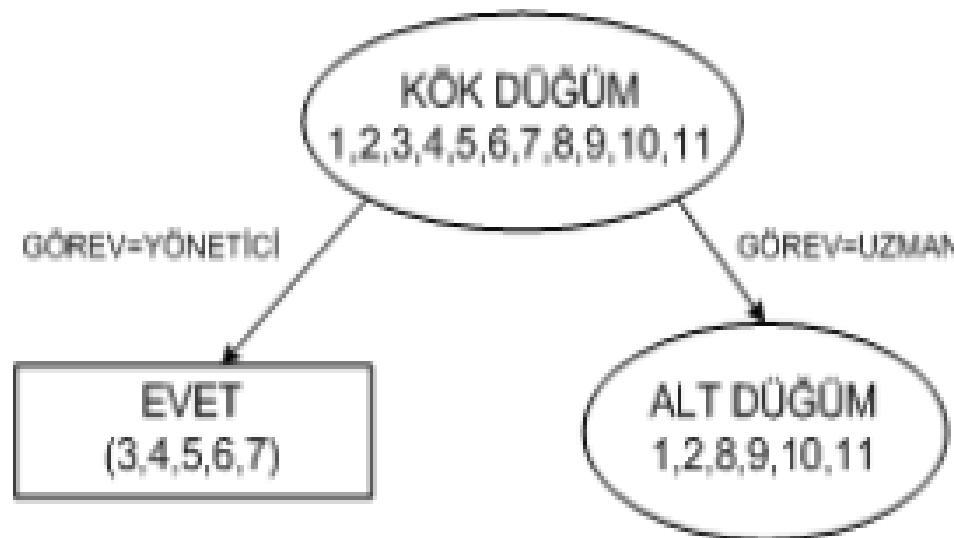
$$\Phi(1|d) = 2 \frac{|B_{sol}|}{|T|} \frac{|B_{sag}|}{|T|} \sum_{j=1}^n abs\left(\frac{|Tsinif_j|}{|B_{sol}|} - \frac{|Tsinif_j|}{|B_{sag}|} \right)$$
$$= 2(0,09)(0,91)[|1 - 0,6| + |0 - 0,4|] = 0,13$$

BÖLÜNME	P _{Sol}	P _{Sağ}	2P _{Sol} P _{Sağ}	$\Phi(B d)$
1	0,09	0,91	0,17	0,13
2	0,45	0,55	0,5	0,07
3	0,45	0,55	0,5	0,07
4	0,18	0,82	0,3	0,26
5	0,45	0,55	0,5	0,07
6	0,36	0,64	0,46	0,2
7	0,55	0,45	0,5	0,66
8	0,45	0,55	0,5	0,66

Örnek

(6/8)

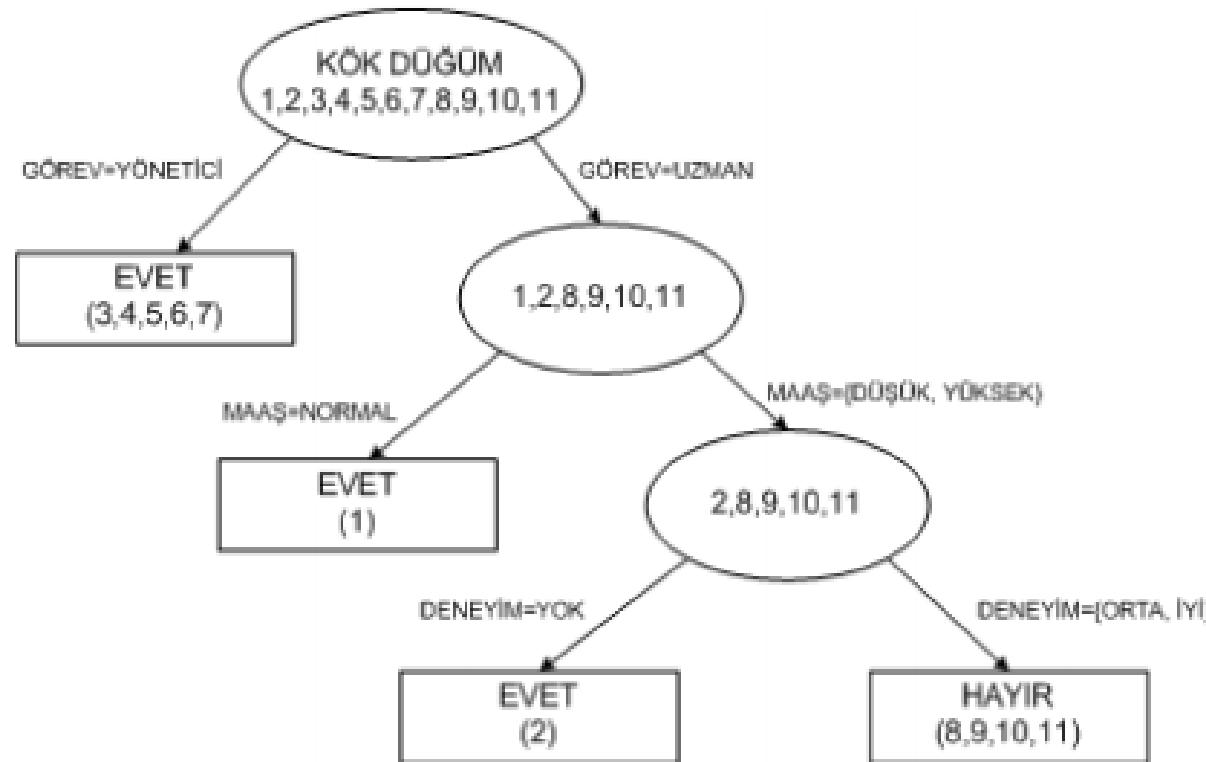
- Aynı işlemler ALT DÜĞÜM için tekrarlanır.



Örnek

(7/8)

- Sonuç karar ağacı.



■ Karar ağacından elde edilen kurallar

- 1. EĞER (GÖREV = YÖNETİCİ) İSE (MEMNUN = EVET)
- 2. EĞER (GÖREV = UZMAN) VE (MAAŞ = NORMAL) İSE (MEMNUN =EVET)
- 3. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM=YOK) İSE (MEMNUN = EVET)
- 4. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM = ORTA VEYA DENEYİM = İYİ) İSE (MEMNUN = HAYIR)

Gini Algoritması

- Gini algoritmasında nitelik değerleri iki parçaaya ayrılarak bölümleme yapılır.
- Her bölünme için $Gini_{sol}$ ve $Gini_{sag}$ değerleri hesaplanır.

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{|Tsinif_i|}{|B_{sol}|} \right)^2 \quad Gini_{sag} = 1 - \sum_{i=1}^k \left(\frac{|Tsinif_i|}{|B_{sag}|} \right)^2$$

- Burada, $Tsinif_i$ soldaki bölümdeki her bir sınıf değerini, $Tsinif_i$ sağdaki bölümdeki her bir sınıf değerini, $|B_{sol}|$ sol bölümdeki tüm değer sayısını, $|B_{sag}|$ sağ bölümdeki tüm değer sayısını gösterir.

$$Gini_j = \frac{1}{n} (|B_{sol}| Gini_{sol} + |B_{sag}| Gini_{sag})$$

- Her bölümlemeden sonra Gini değeri en küçük olan seçilir.

Örnek

(1/8)

SIRA	EĞİTİM	YAŞ	CİNSİYET	SONUÇ
1	ORTA	YAŞLI	ERKEK	EVET
2	İLK	GENÇ	ERKEK	HAYIR
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET
7	İLK	GENÇ	KADIN	HAYIR
8	ORTA	ORTA	ERKEK	EVET

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

Örnek

(2/8)

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

EĞİTİM için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0,444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5} \right)^2 + \left(\frac{1}{5} \right)^2 \right] = 0,320$$

Örnek

(3/8)

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

YAŞ için

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] = 0$$

$$Gini_{sag} = 1 - \left[\left(\frac{5}{6} \right)^2 + \left(\frac{1}{6} \right)^2 \right] = 0,278$$

Örnek

(4/8)

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

CİNSİYET için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0,444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5} \right)^2 + \left(\frac{1}{5} \right)^2 \right] = 0,320$$

Örnek

(5/8)

Gini değerleri

$$Gini_{EGITIM} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

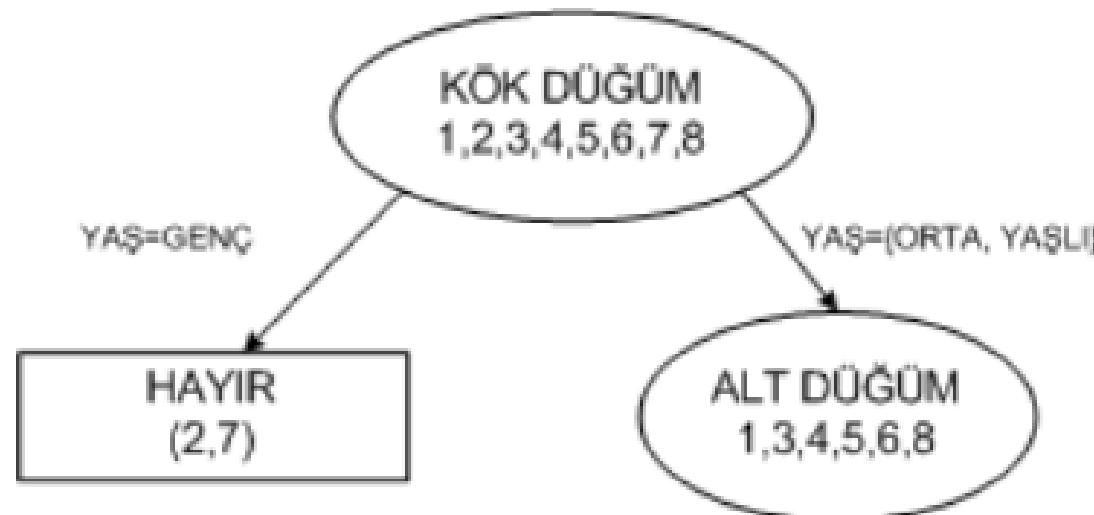
$$Gini_{YAS} = \frac{2(0) + 6(0,278)}{8} = 0,209$$

$$Gini_{CINSIYET} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

İlk bölünme YAŞ niteliğine göre yapılacaktır.

Örnek

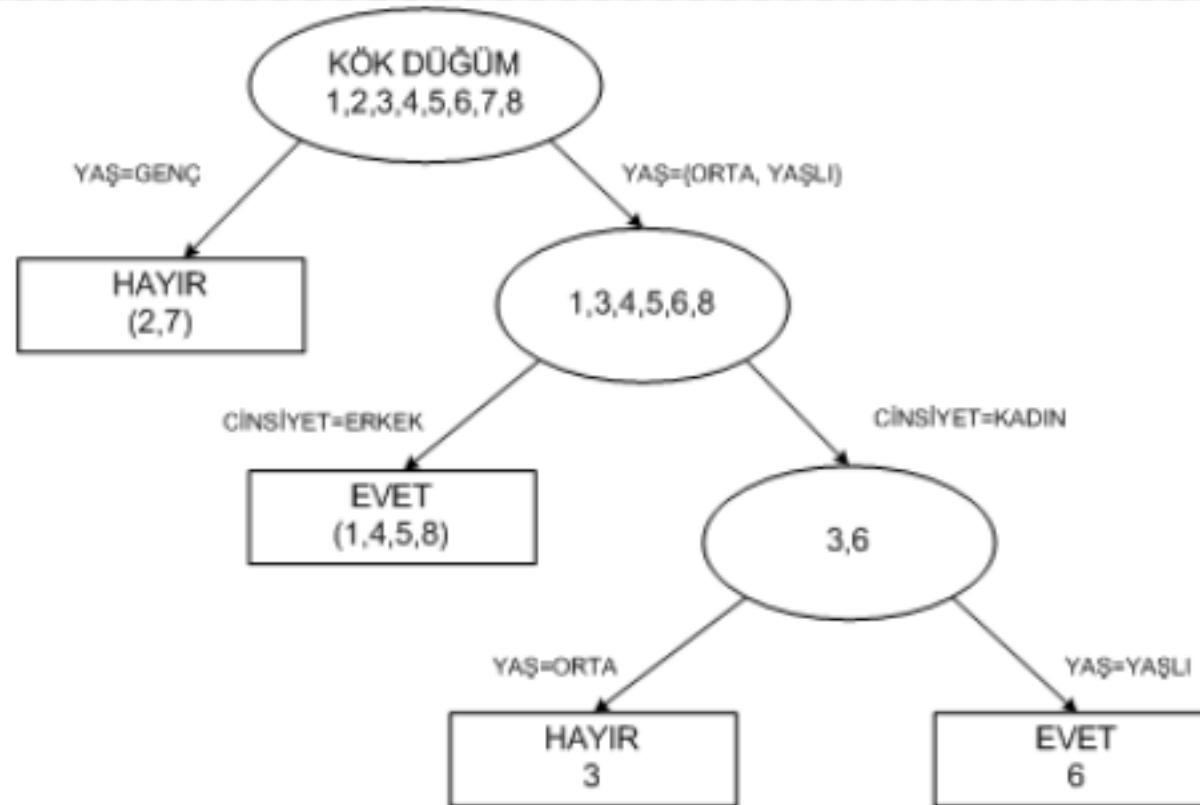
(6/8)



Aynı işlemleri ALT DÜĞÜM için tekrarlanır.

Örnek

(7/8)



■ Karar ağacından elde edilen kurallar

- 1. EĞER (YAŞ = GENÇ) İSE (SONUÇ = HAYIR)
- 2. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = ERKEK) İSE (SONUÇ = EVET)
- 3. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = YAŞLI) İSE (SONUÇ = EVET)
- 4. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = ORTA) İSE (SONUÇ = HAYIR)

Bellek Tabanlı Algoritmalar

- K-en yakın komşu algoritması (K-nearest neighbor algorithm).

K-en yakın komşu algoritması

- Sınıflandırma yöntemlerinden birisi de **K-en yakın komşu algoritmasıdır.**
- Bu yöntem sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarakorneğe katılacak yeni bir gözlemin hangi sınıfı ait olduğunu belirlemek amacıyla kullanılır.
- Bu yöntem örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında i ve j noktaları için örneğin Öklid uzaklık formülü kullanılabilir. (Diğer uzaklıklar veri önişleme kısmında açıklanmıştır)

$$d(i,j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

K-en yakın komşu algoritması

- K-en yakın komşu algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki adımları içerir.
 - a) K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
 - b) Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
 - c) Yukarıda hesaplanan uzaklıklara göre sıralılar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
 - d) Seçilen sıralıların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
 - e) Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

Örnek 1.

- Aşağıda verilen gözlem tablosu X1 ve X2 nitelikleri ve Y sınıfından oluşmaktadır. Bu gözlem değerine bağlı olarak yeni bir gözlem değeri olan $X_1=8$, $X_2=4$ değerlerinin yanı (8,4) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritması ile bulunuz.

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
7	9	İYİ
9	7	KÖTÜ
11	7	KÖTÜ
10	2	KÖTÜ

Örnek 1.

- a) **K'nın belirlenmesi:** $k=4$ kabul edilir.
- b) **Uzaklıkların hesaplanması:** $(8,4)$ noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığuna göre hesaplanır.

$$d(i,j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Birimde birinci gözlem olan $(2,4)$ noktası ile $(8,4)$ noktası arasındaki uzaklık,

$$d(i,j) = \sqrt{(2 - 8)^2 + (4 - 4)^2} = 6.00$$

Benzer şekilde uzaklıklar hesaplandığında tablodaki sonuç ortaya çıkacaktır.

Örnek 1.

- (8,4) noktasının gözlem değerlerine olan uzaklıklarını,

X1	X2	Uzaklık
2	4	6
3	6	5,39
3	4	5
4	10	7,21
5	8	5
6	3	2,24
7	9	5,1
9	7	3,16
11	7	4,24
10	2	2,83

■c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=4 tanesi belirlenir. Bu dört nokta verilen (8,4) noktasına en yakın gözlem değerleridir.

X1	X2	Uzaklık	Sıra
2	4	6	9
3	6	5,39	8
3	4	5	6
4	10	7,21	10
5	8	5	5
6	3	2,24	1
7	9	5,1	7
9	7	3,16	3
11	7	4,24	4
10	2	2,83	2

Örnek 1.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (8,4) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu dört gözlem içinde bir tane **İYİ** 3 tane **KÖTÜ** sınıfı vardır.

X1	X2	Uzaklık	Sıra	k komşunun Y değeri
2	4	6	9	
3	6	5,39	8	
3	4	5	6	
4	10	7,21	10	
5	8	5	5	
6	3	2,24	1	İYİ
7	9	5,1	7	
9	7	3,16	3	KÖTÜ
11	7	4,24	4	KÖTÜ
10	2	2,83	2	KÖTÜ

- e) **Yeni gözlemin sınıfı:** KÖTÜ değerlerinin sayısı İYİ değerlerinin sayılarından fazla olduğu için (8,4) noktasının sınıfı **KÖTÜ** olarak belirlenir.
-

Örnek 2.

- Aşağıda verilen gözlem tablosunda Y sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirleyelim. Gözlemlerin gerçek değerleri değil normalize edilmiş değerleri kullanılacaktır. Gözlem değerlerini (0,1) aralığına çekmek için min-max normalleştirmesi kullanılacaktır.

X1	X2	X3	Y
10	5	19	EVET
8	2	4	HAYIR
18	16	6	HAYIR
12	15	8	EVET
3	15	15	EVET

Örnek 2.

- Min-max normalleştirmesi sonucu dönüştürülen değerler aşağıdadır.
- $X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$ (min-max normalizasyonu

X1	X2	X3	Y
0,47	0,21	1	EVET
0,33	0	0	HAYIR
1	1	0,13	HAYIR
0,6	0,93	0,27	EVET
0	0,93	0,73	EVET

- Aday noktanın normalizasyon değeri (0.27, 0.43, 0.07)

Örnek 2.

- a) **K'nın belirlenmesi:** k=3 kabul edilir.
- b) **Uzaklıkların hesaplanması:** (0,27, 0,43, 0,07) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığuna göre hesaplanır.

$$d(i,j) = \sqrt{(0,47 - 0,27)^2 + (0,21 - 0,43)^2 + (1 - 0,07)^2} = 0,98$$

X1	X2	X3	Uzaklık
0,47	0,21	1	0,98
0,33	0	0	0,44
1	1	0,13	0,93
0,6	0,93	0,27	0,63
0	0,93	0,73	0,87

Örnek 2.

- c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=3 tanesi belirlenir.

X1	X2	X3	Uzaklık	Sıra
0,47	0,21	1	0,98	5
0,33	0	0	0,44	1
1	1	0,13	0,93	4
0,6	0,93	0,27	0,63	2
0	0,93	0,73	0,87	3

Örnek 2.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (0,27, 0,43, 0,07) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu üç gözlem içinde bir tane **HAYIR** 2 tane **EVET** sınıfı vardır.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri
0,47	0,21	1	0,98	5	
0,33	0	0	0,44	1	HAYIR
1	1	0,13	0,93	4	
0,6	0,93	0,27	0,63	2	EVET
0	0,93	0,73	0,87	3	EVET

- e) **Yeni gözlemin sınıfı:** **EVET** değerlerinin sayısı **HAYIR** değerlerinin sayılarından fazla olduğu için (7,8,5) gözleminin sınıfı **EVET** olarak kabul edilir.
-

Ağırlıklı Oylama

- K-en yakın komşu algoritması sınıfı bilinmeyen gözlem değeri için k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanmaktadır. Ancak seçilen bu sınıf sadece k komşunun göz önüne alınması nedeniyle her zaman uygun olmayabilir. Bu son aşamada k komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine **ağırlıklı oylama** (weighted voting) denilen bir yöntem uygulanabilir.
- Söz konusu ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıyla göre ağırlıklı uzaklıkların hesaplanması dayanır.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

- $d(i,j)$ ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf olarak kabul edilir.

Örnek 2. Ağırlıklı Oylama Sonucu

- Ağırlıklı Oylama sonucunda Örnek 2.'deki değerin sınıfının HAYIR olduğu görülür.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri	Ağırlıklı Oylama
0,47	0,21	1	0,98	5		
0,33	0	0	0,44	1	HAYIR	5,17
1	1	0,13	0,93	4		
0,6	0,93	0,27	0,63	2	EVET	2,52
0	0,93	0,73	0,87	3	EVET	3,84

VERİ MADENCİLİĞİ

(Kümeleme)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

İçerik

- Kümeleme İşlemleri
- Kümeleme Tanımı
- Kümeleme Uygulamaları
- Kümeleme Yöntemleri

Kümeleme (Clustering)

- Kümeleme birbirine benzeyen veri parçalarını ayırma işlemidir ve kümeleme yöntemlerinin çoğu veri arasındaki uzaklıklarını kullanır.
- Nesneleri kümelere (gruplara) ayırma
- Küme: birbirine benzeyen nesnelerden oluşan grup
 - Aynı kümedeki nesneler birbirine daha çok benzer
 - Farklı kümedeki nesneler birbirine daha az benzer

Kümeleme

- Danışmansız öğrenme: Hangi nesnenin hangi sınıfı ait olduğu ve sınıf sayısı belli değil
- Uygulamaları:
 - verinin dağılımını anlama
 - başka veri madenciliği uygulamaları için ön hazırlık

Kümeleme Uygulamaları

- Örütü tanıma
- Görüntü işleme
- Ekonomi
- Aykırılıkları belirleme
- WWW
 - Doküman kümeleme
 - Kullanıcı davranışlarını kümeleme
 - Kullanıcıları kümeleme
- Diğer veri madenciliği algoritmaları için bir önişleme adımı
- Veri azaltma – küme içindeki nesnelerin temsil edilmesi için küme merkezlerinin kullanılması

Veri Madenciliğinde Kümeleme

- Ölçeklenebilirlik
- Farklı tipteki niteliklerden oluşan nesneleri kümeleme
- Farklı şekillerdeki kümeleri oluşturabilme
- En az sayıda giriş parametresi gereksinimi
- Hatalı veriler ve aykırılıklardan en az etkilenme
- Model oluşturma sırasında örneklerin sırasından etkilenmemesi
- Çok boyutlu veriler üzerinde çalışma
- Kullanıcıların kısıtlarını göz önünde bulundurma
- Sonucun yorumlanabilir ve anlaşılabilir olması

İyi Kümeleme

- İyi kümeleme yöntemiyle elde edilen kümelerin özellikleri
 - aynı kümeye içindeki nesneler arası benzerlik fazla
 - farklı kümelerde bulunan nesneler arası benzerlik az
- Oluşan kümelerin kalitesi seçilen benzerlik ölçütüne ve bu ölçütün gerçekleşmesine bağlı
 - Uzaklık / Benzerlik nesnelerin nitelik tipine göre değişir
 - Nesneler arası benzerlik: $s(i,j)$
 - Nesneler arası uzaklık: $d(i,j) = 1 - s(i,j)$
- İyi bir kümeleme yöntemi veri içinde gizlenmiş örüntülerini bulabilmeli
- Veriyi gruplama için uygun kümeleme kriteri bulunmalı
 - kümeleme= aynı kümedeki nesneler arası benzerliği en büyütken, farklı kümedeki nesneler arası benzerliği en küçültken fonksiyon
- Kümeleme sonucunun kalitesi seçilen kümelerin şekline ve temsil edilme yöntemine bağlı

Kümeleme Yöntemlerinde Kullanılan Uzaklıklar

- Öklid

$$d(i,j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

- Minkowski

$$d(i,j) = \left[\sum_{k=1}^p (|x_{ik} - x_{jk}|^m) \right]^{\frac{1}{m}}$$

- Manhattan

$$d(i,j) = \sum_{k=1}^p (|x_{ik} - x_{jk}|)$$

Kümeleme Yöntemleri

- Hiyerarsik Kümeleme
 - Birleştirici Hiyerarsik Yöntemler
 - En yakın komşu algoritması
 - En uzak komşu algoritması
- Hiyerarsik Olmayan Kümeleme
 - K-Ortalamalar Yöntemi (K-Means)

En yakın komşu algoritması

- En yakın komşu yöntemine «tek bağlantı kümeleme yöntemi» adı da verilmektedir. Başlangıçta tüm gözlem değerleri birer kümeye olarak değerlendirilir. Adım adım bu kümeler birleştirilerek yeni kümeler elde edilir.
- Bu yöntemde öncelikle gözlemler arasındaki uzaklıklar belirlenir. Öklid uzaklık bağıntısı kullanılabilir.
- Uzaklıklar göz önüne $\text{Min } d(i,j)$ seçilir. Söz konusu uzaklıkla ilgili satırlar birleştirilerek yeni bir kume elde edilir. Bu duruma göre uzaklıkların yeniden hesaplanması gereklidir.
- Tek bir gözlemden oluşan kümeler arasındaki uzaklıkları doğrudan hesaplayabiliyoruz. Ancak birden fazla gözlem değerine sahip olan iki kume arasındaki uzaklığının belirlenmesi gerektiğinde farklı bir yol izlenir. İki kümenin içeriği gözlemler arasında «birbirine en yakın olanların uzaklığı» iki kümenin birbirine olan uzaklığını olarak kabul edilir.

Örnek 1.

- Aşağıdaki tabloda verilen beş gözlem değeri, en yakın komşu algoritması ile kümelenmek isteniyor.

Gözlemler	X_1	X_2
1	4	2
2	6	4
3	5	1
4	10	6
5	11	8

- Adım1. Öncelikle uzaklık tablosu oluşturulur. Her bir gözlemin birbiriyle arasındaki öklid uzaklığı hesaplanır.

Örnek 1.

$$d(1,2) = \sqrt{(4 - 6)^2 + (2 - 4)^2} = 2,83$$

$$d(1,3) = \sqrt{(4 - 5)^2 + (2 - 1)^2} = 1,41$$

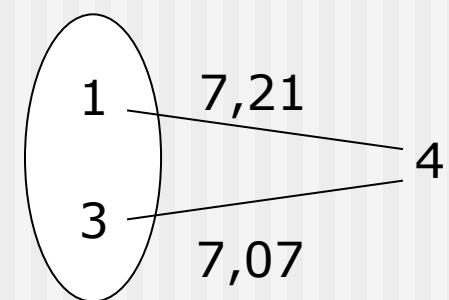
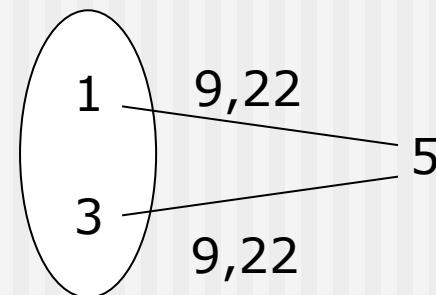
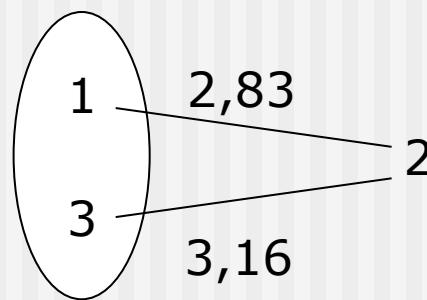
$$d(1,4) = \sqrt{(4 - 10)^2 + (2 - 6)^2} = 7,21$$

...

Gözlemler	1	2	3	4	5
1					
2	2,83				
3	1,41	3,16			
4	7,21	4,47	7,07		
5	9,22	6,4	9,22	2,24	

Örnek 1.

- Adım 2. Uzaklıklar tablosunda Min $d(i,j)$ değerinin 1,41 olduğu görülmektedir. İlgili gözlemler 1 ve 3 gözlemleridir. Bu iki değer birleştirilerek (1,3) kümesi elde edilir. Sonrasında bu kümeye göre uzaklıklar matrisi yeniden incelenir.



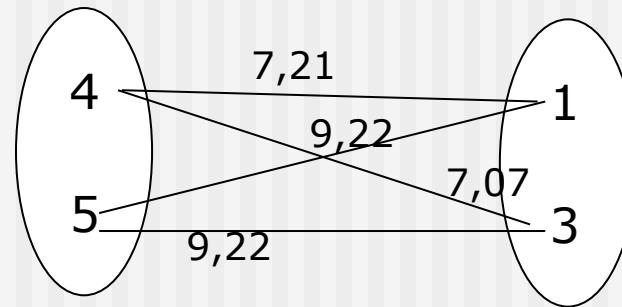
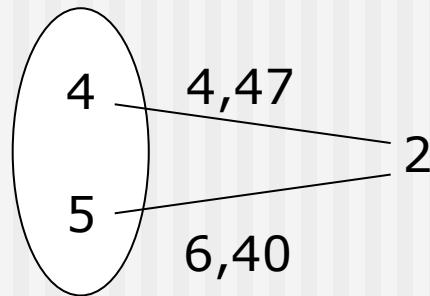
Örnek 1.

- Yeni uzaklık tablosu,

Gözlemler	(1,3)	2	4	5
(1,3)				
2	2,83			
4	7,07	4,47		
5	9,22	6,4	2,24	

- Bu tabloya bakıldığında $\min d(i,j) = 2,24$ olduğu görülür. Bu değerin 4 ve 5 gözlemleri arasındaki uzaklıği görülür. (4,5) yeni bir kümeye oluşturur. Bu durumda (1,3), 2 ve (4,5) kümeleri arasındaki uzaklık tablosu yeniden oluşturulur.
-

Örnek 1.



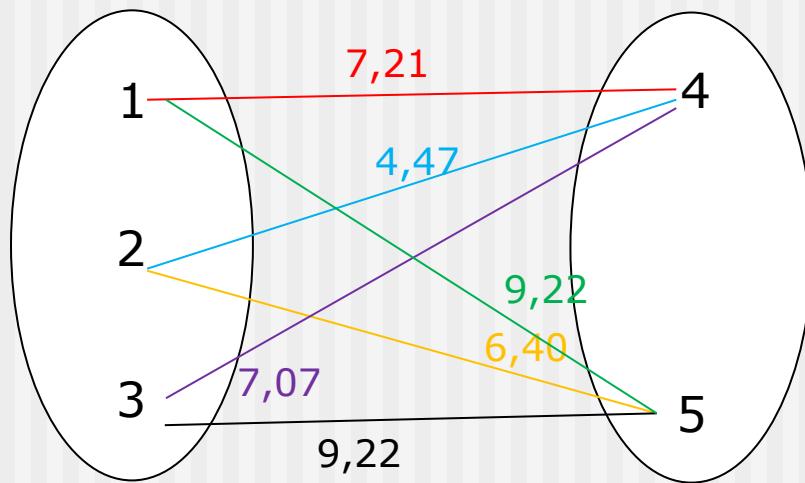
Örnek 1.

- Bu durumdaki uzaklık tablosu,

Gözlemler	(1,3)	2	(4,5)
(1,3)			
2	2,83		
(4,5)	7,07	4,47	

- Adım 4. En son uzaklıklar tablosu incelendiğinde $\text{Min } d(i,j)=2,83$ olduğu görülür. O halde bu uzaklık ile ilgili olan 2 gözlemi ile (1,3) kümesi birleştirilecektir. Elde edilen (1,2,3) kümesi ile (4,5) kümesi arasındaki uzaklığı belirlemek için kümeler içindeki her bir değer eşlenir ve en küçük olan belirlenir. En küçük uzaklık 4,47 olduğuna göre iki küme arasındaki uzaklığın bu değer olduğu kabul edilir.

Örnek 1.



önüne alınarak kümeler şu şekilde belirlenir.

Uzaklık	Kümeler
1,41	(1,3)
2,24	(4,5)
2,83	(1,2,3)
4,47	(1,2,3,4,5)

- Adım 5. Elde edilen iki küme birleştirilerek sonuç küme bulunur. Bu küme $(1,2,3,4,5)$ gözlemlerinden oluşan kümedir. Uzaklık düzeyi göz

En uzak komşu algoritması

- En yakın komşu algoritması ile benzer adımları içerir. Gözlemler arasındaki uzaklıklar hesaplanır ve minimum değerli olan birleştirilir. Sonraki küme uzaklıkları tablosu oluşturulurken en uzak mesafe kullanılır.

K-Ortalamalar Yöntemi (K-Means) (1/2)

- Bu yöntemde daha başlangıçta belli sayıdaki küme için toplam ortalama hatayı minimize etmek amaçlanır.
- N noyutlu uzayda N örnekli kümelerin verildiğini varsayılm. Bu uzay $\{C_1, C_2, \dots, C_k\}$ biçimde K kümeye ayrılın. O zaman $\sum n_k = N$ ($k=1,2,\dots,k$) olmak üzere C_k kümесinin ortalama vektörü M_k şu şekilde hesaplanır.

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

- Burada X_k değeri C_k kümese ait olan $i.$ örnektir. C_k kümesi için kare-hata, her bir C_k örneği ile onun merkezi (centroid) arasındaki Öklid uzaklıklarını toplamıdır. Bu hataya «küme içi değişme» adı da verilir.

K-Ortalamalar Yöntemi (K-Means) (2/2)

- Küme içi değişimeler şu şekilde hesaplanır.

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

- K kümesini içeren bütün kümeler uzayı için kare-hata içindeki değişimelerin toplamıdır. O halde söz konusu kare-hata şu şekilde hesaplanır.

$$E_k^2 = \sum_{k=1}^K e_k^2$$

- Kare-hata kümeleme yönteminin amacı verilen K değeri için E_k^2 değerini minimize eden K kümelerini bulmaktır. O halde k-ortalamalar algoritmasında E_k^2 değerinin bir önceki iterasyona göre azalması beklenir.

K-Means Algoritmasının Adımları

- K-Means algoritmasına başlamadan önce k küme sayısının belirlenmesi gereklidir. Sonra aşağıdaki işlemler gerçekleştirilir.
 1. Her bir kümenin merkezi belirlenir. Bu merkezler M_1, M_2, \dots, M_k biçimindedir.
 2. e_1, e_2, \dots, e_k küme içi değişimeler hesaplanır. Bu değişimelerin toplamı olan E_k^2 değeri bulunur.
 3. M_k merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır. Bir gözlem değeri hangi yakın ise o merkez ile ilgili küme içine dahil edilir.
 4. Yukarıdaki 2. ve 3. adımlar kümelerde değişiklik olmayıncaya kadar devam ettirilir.

K-Means Algoritmasının Özellikleri

- Gerçeklemesi kolay
- Karmaşıklığı diğer kümeleme yöntemlerine göre az
- K-Means algoritması bazı durumlarda iyi sonuç vermeyebilir
 - Veri grupları farklı boyutlarda ise
 - Veri gruplarının yoğunlukları farklı ise
 - Veri gruplarının şekli küresel değilse
 - Veri içinde aykırılıklar varsa

Örnek 2.

- Aşağıdaki gözlem değerleri k-ortalamalar yöntemi ile kümelenmek isteniyor.

Gözlemler	Değişken1	Değişken2
X_1	4	2
X_2	6	4
X_3	5	1
X_4	10	6
X_5	11	8

- Kümelerin sayısı başlangıçta $k=2$ kabul edilir. Rasgele iki küme belirlenir.

$$C_1 = \{X_1, X_2, X_4\}$$
$$C_2 = \{X_3, X_5\}$$

Örnek 2.

Gözlemler	Değişken1	Değişken2	Küme Üyeliği
X ₁	4	2	C ₁
X ₂	6	4	C ₁
X ₃	5	1	C ₂
X ₄	10	6	C ₁
X ₅	11	8	C ₂

- Adım 1. a) Belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{ \frac{4 + 6 + 10}{3}, \frac{2 + 4 + 6}{3} \right\} = \{6.67, 4.0\}$$

$$M_2 = \left\{ \frac{5 + 11}{2}, \frac{1 + 8}{2} \right\} = \{8.0, 4.5\}$$

Örnek 2.

- b) Küme içi değişimeler şu şekilde hesaplanır.

$$\begin{aligned} e_1^2 &= [(4 - 6,67)^2 + (2 - 4,0)^2] + [(6 - 6,67)^2 + (4 - 4,0)^2] \\ &\quad + [(10 - 6,67)^2 + (6 - 4,0)^2] = 26,67 \end{aligned}$$

$$e_2^2 = [(5 - 8)^2 + (1 - 4,5)^2] + [(11 - 8)^2 + (8 - 4,5)^2] = 42,50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 26,67 + 42,50 = 69,17$$

Örnek 2.

- C) M_1 ve M_2 merkezlerinden olan uzaklıkların minimum olması istendiğinden aşağıdaki hesaplamalar yapılır. Öklid uzaklık formülü kullanılarak söz konusu mesafeler hesaplanır. Örneğin (M_1, X_1) noktaları arasındaki uzaklık $M_1=\{6.67, 4.00\}$ ve $X_1=\{4, 2\}$ olduğuna göre şu şekilde hesaplanır.

$$d(M_1, X_1) = \sqrt{(6,67 - 4)^2 + (4 - 2)^2} = 3,33$$

$$d(M_2, X_1) = \sqrt{(8 - 4)^2 + (4,5 - 2)^2} = 4,72$$

- Bu işlemler sonucunda X_1 gözlem değerinin M_1 ve M_2 merkezlerine olan uzaklıkları göz önüne alındığında $d(M_1, X_1) < d(M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

Örnek 2.

Gözlemler	M_1 'den uzaklık	M_2 'den uzaklık	Küme Üyeliği
X_1	$d(M_1, X_1) = 3,33$	$d(M_2, X_1) = 4,72$	C_1
X_2	$d(M_1, X_2) = 0,67$	$d(M_2, X_2) = 2,06$	C_1
X_3	$d(M_1, X_3) = 3,43$	$d(M_2, X_3) = 4,61$	C_1
X_4	$d(M_1, X_4) = 3,89$	$d(M_2, X_4) = 2,50$	C_2
X_5	$d(M_1, X_5) = 5,90$	$d(M_2, X_5) = 4,61$	C_2

Örnek 2.

- Bu durumda yeni kümeler şu şekilde olacaktır.

$$C_1 = \{X_1, X_2, X_3\}$$

$$C_2 = \{X_4, X_5\}$$

- Adım 2. Yukarıda belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{ \frac{4 + 6 + 5}{3}, \frac{2 + 4 + 1}{3} \right\} = \{5, 2.33\}$$

$$M_2 = \left\{ \frac{10 + 11}{2}, \frac{6 + 8}{2} \right\} = \{10.5, 7\}$$

Örnek 2.

- b) Küme içi değişimeler şu şekilde hesaplanır.

$$\begin{aligned} e_1^2 &= [(4 - 5)^2 + (2 - 2,33)^2] + [(6 - 5)^2 + (4 - 2,33)^2] \\ &\quad + [(5 - 5)^2 + (1 - 2,33)^2] = 9,33 \end{aligned}$$

$$e_2^2 = [(10 - 10,5)^2 + (6 - 7)^2] + [(11 - 10,5)^2 + (8 - 7)^2] = 2,50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 9,33 + 2,50 = 11,83$$

- Bu değerin bir önceki iterasyonda elde edilen $E^2 = 69,17$ değerinden daha küçük olduğu anlaşılmaktadır.

Örnek 2.

- M_1 ve M_2 merkezlerinden gözlem değerlerine olan uzaklıklar hesaplanır. Bunun sonucunda $d(M_1, X_1) < d(M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

Gözlemler	M_1 'den uzaklık	M_2 'den uzaklık	Küme Üyeliği
X_1	$d(M_1, X_1) = 1,05$	$d(M_2, X_1) = 8,20$	C_1
X_2	$d(M_1, X_2) = 1,94$	$d(M_2, X_2) = 5,41$	C_1
X_3	$d(M_1, X_3) = 1,33$	$d(M_2, X_3) = 8,14$	C_1
X_4	$d(M_1, X_4) = 6,20$	$d(M_2, X_4) = 1,12$	C_2
X_5	$d(M_1, X_5) = 8,25$	$d(M_2, X_5) = 1,12$	C_2

Örnek 2.

- Bu durumda yeni kümeler şu şekilde oluşacaktır.

$$C_1 = \{X_1, X_2, X_3\}$$

$$C_2 = \{X_4, X_5\}$$

- Kümelerde önceki adıma göre herhangi bir değişme olmadığı için iterasyona son verilir.

VERİ MADENCİLİĞİ

(Birliktelik Kuralları)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

İçerik

- Birliktelik Kurallarının Tanımı
- Destek ve Güven Ölçütleri
- Apriori Algoritması

Birliktelik Kuralları (Association Rules)

- Birliktelik kuralları
 - Veri kümesi içindeki yaygınörüntülerin, nesneleri oluşturan nitelikler arasındaki ilişkilerin bulunması □
- Birliktelik kurallarını kullanma: veri içindeki kuralları belirleme □
 - Hangi ürünler çoğunlukla birlikte satılıyor? □
 - Kişisel bilgisayar satın alan bir kişinin bir sonraki satın alacağı ürün ne olabilir? □
 - Yeni bir ilaca duyarlı olan DNA tipleri hangileridir? □
 - Web dokümanları otomatik olarak sınıflandırılabilir mi?

Destek ve Güven Ölçütleri

- Birliktelik çözümlemelerinin en yaygın uygulaması perakende satışlarda müşterilerin satın alma eğilimlerini belirlemek amacıyla yapılmaktadır. Müşterilerin bir anda satın aldığı tüm ürünleri ele alarak satın alma eğilimini ortaya koyan uygulamalara «**pazar sepet çözümlemesi**» denilir.
- Pazar sepet çözümlemelerinde satılan ürünler arasındaki ilişkileri ortaya koymak için «destek» ve «güven» gibi iki ölçütten yararlanılır. Bu ölçütlerin hesaplanmasıında destek sayısı adı verilen bir değer kullanılır. Kural destek ölçütü tüm alışverişler içinde hangi oranda tekrarlandığını belirler.

Destek ve Güven Ölçütleri

- Kural güven ölçütü A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koyar. A ürün grubunu alanların B ürün grubunu da alma durumu yani birliktelik kuralı $A \rightarrow B$ biçiminde gösterilir. Bu durumda kural destek ölçütü şu şekilde ifade edilir.

$$\text{destek}(A \rightarrow B) = \frac{\text{sayı}(A, B)}{N}$$

- Burada sayı(A,B) destek sayısı A ve B ürün gruplarını birlikte içeren alışveriş sayısını göstermektedir. N ise tüm alışverişlerin sayısını göstermektedir. A ve B ürün gruplarının birlikte satın alınması olasılığını ifade eden kural güven ölçütü şu şekilde hesaplanır.

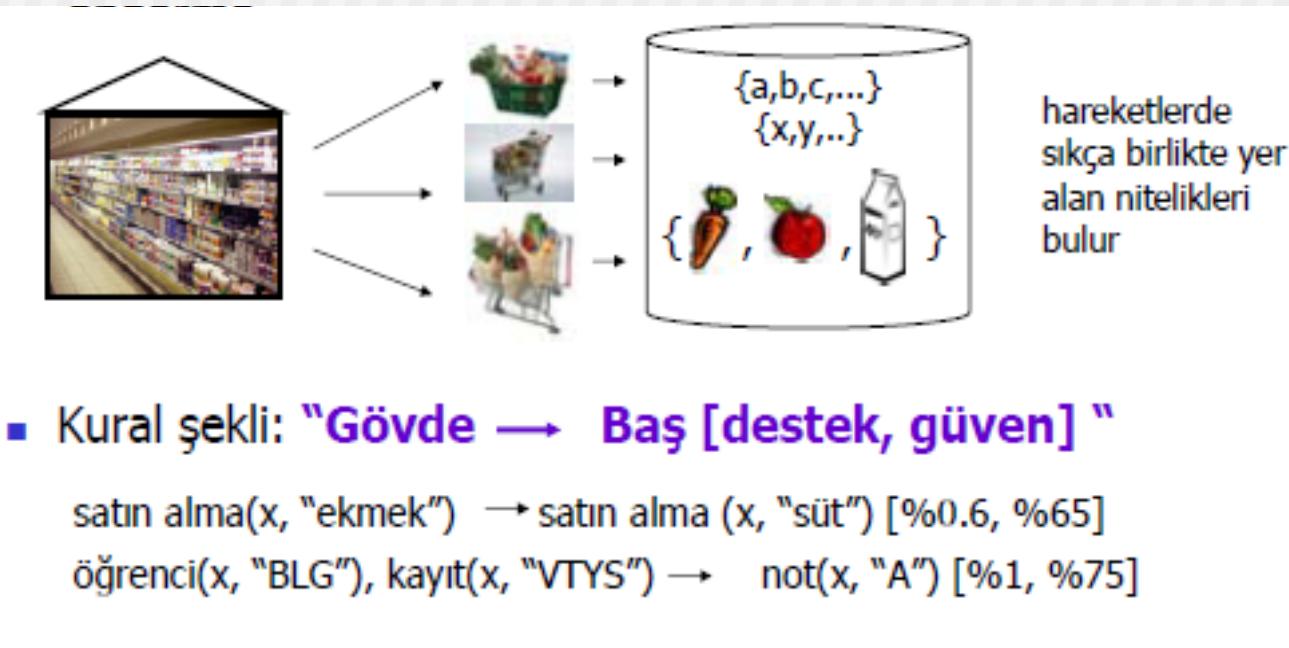
$$\text{güven}(A \rightarrow B) = \frac{\text{sayı}(A, B)}{\text{sayı}(A)}$$

Destek ve Güven Ölçütleri

- Birliktelik kuralları belirlenirken destek ve güven ölçütleri yanı sıra bu değerleri karşılaştırmak üzere eşik değere gereksinim vardır. Hesaplanan destek veya güven ölçütlerinin destek(eşik) ve güven(eşik) değerlerinden büyük olması beklenir. Hesaplanan destek veya güven ölçütleri ne kadar büyük ise birliktelik kurallarının da o derce güclü olduğuna karar verilir.

Birliktelik Kuralları Bulma

- Bir niteliğin (veya nitelikler kümesinin) varlığını harekette bulunan başka niteliklerin varlıklarına dayanarak öngörme



Birliktelik Kuralları Bulma

- Bütün niteliklerden oluşan küme $I=\{i_1, i_2, \dots, i_d\}$
 - $I=\{\text{ekmek, süt, bira, kola, yumurta, bez}\}$
- Hareket $T_j \subseteq I$, $Tj=\{j_1, j_2, \dots, j_k\}$
 - $T1=\{\text{ekmek, süt}\}$
- Hareketlerden oluşan veri kümesi $D=\{T_1, T_2, \dots, T_N\}$



Market Alışveriş verisi

Hareket	Öğeler
T1	Ekmek, Süt
T2	Ekmek, Bez, Bira, Yumurta
T3	Süt, Bez, Bira, Kola
T4	Ekmek, Süt, Bez, Bira
T5	Ekmek, Süt, Bez, Kola

Yaygın nitelikler:

Bez, bira
Süt, ekmek, yumurta, kola
Bira, ekmek, süt

Bulunan İlişkilendirme Kuralları

$\{\text{Bez}\} \rightarrow \{\text{Bira}\}$,
 $\{\text{Süt, Ekmek}\} \rightarrow \{\text{Yumurta, Kola}\}$,
 $\{\text{Bira, Ekmek}\} \rightarrow \{\text{Süt}\}$

Apriori Algoritması

- Birliktelik kurallarının üretilmesi için birçok yöntem kullanılmaktadır. Bunlardan en yaygın kullanılanı Apriori Algoritmasıdır.
- Apriori algoritması, özellikle çok büyük ölçekli veri tabanları üzerindeki veri madenciliği çalışmalarında geliştirilmiştir. Genel anlamda ilişki kuralı (association rule, birliktelik kuralı) çıkarımında kullanılan bir algoritmadır. Algoritmanın amacı, veri tabanında bulunan satırlar arasındaki bağlantıyı ortaya çıkarmaktır.
- Algoritmanın ismi, kendinden önceki çıkarımlara bağlı olduğu için, latince, önce anlamına gelen “prior” kelimesinden gelmektedir.
- Algoritma yapı olarak, aşağıdan yukarıya (bottom-up) yaklaşımı kullanmakta olup her seferinde tek bir elemanı incelemekte ve bu elemanla diğer adayların ilişkisini ortaya çıkarmaya çalışmaktadır.
- Ayrıca algoritmanın her eleman için çalışmasını, bir arama algoritmasına benzetmek mümkündür. Algoritma, bu anlamda genişlik öncelikli arama (breadth first search) yapısında olup adayları birer ağaç (tree) gibi düşünerek bu ağaç üzerinde arıyor kabul edilebilir.

Apriori Algoritmasının Adımları

- 1. Minimum destek sayısı ve minimum güven değerinin belirlenmesi
- 2. Öğe kümeler içerisindeki her bir öğenin destek değerinin bulunması
- 3. Minimum destek değerinden daha düşük desteğe sahip olan öğelerin devre dışı bırakılması
- 4. Elde edilen tekli birliktelikler dikkate alınarak ikili birlikteliklerin oluşturulması
- 5. Minimum destek değerinden düşük olan öğe kümelerin çıkartılması
- 6. Üçlü birlikteliklerin oluşturulması
- 7. Üçlü birlikteliklerden minimum destek değerini geçenlerin dışındakilerin çıkarılması
- 8. Üçlü birlikteliklerden birliktelik kurallarının çıkarılması

Örnek 1.

- Bir mağazada alışveriş yapan müşterilere ilişkin olarak kayıtlar tutulmuş ve beş müşterinin yaptığı alışveriş göz önüne alınmıştır. Müşterilerin bir defada yaptığı alışverişler bir satırda yer almaktadır ve aşağıdaki tabloda verilmiştir. Bu tablodaki veriler kullanılarak müşteri davranışları Apriori Algoritmasıyla ortaya konmak isteniyor.

Müşteri Aldığı Ürünler	
1	Şeker, Çay, Ekmek
2	Ekmek, Peynir, Zeytin, Makarna
3	Şeker, Peynir, Deterjan, Ekmek, Makarna
4	Ekmek, Peynir, Çay, Makarna
5	Peynir, Makarna, Şeker, Bira

Örnek 1.

- a) Çözülemeye bazı varsayımlarla başlanır. Destek ve güven ölçütleri için eşik değerleri belirlenir.

$$\text{destek(eşik)} = \%60$$

$$\text{güven(eşik)} = \%75$$

- Burada destek(eşik)=%60 olduğuna ve tüm müşteri sayısı 5 olduğuna göre **eşik destek sayısının** $(0,60)*5=3$ olduğu anlaşılır.
- b) Beş müşterinin alışveriş yaptığı ürünlerin kümlesi {şeker, çay, ekmek, makarna, peynir, deterjan, bira, zeytin} biçimindedir. Nu ürünlerin her biri için destek değerleri hesaplanır.

$$\text{sayı(Şeker)} = 3$$

$$\text{sayı(Deterjan)} = 1$$

$$\text{sayı(Çay)} = 2$$

$$\text{sayı(Bira)} = 1$$

$$\text{sayı(Ekmek)} = 4$$

$$\text{sayı(Zeytin)} = 1$$

$$\text{sayı(Makarna)} = 4$$

Örnek 1.

- Destek değerlerinin hesaplanması

Ürün	Sayı
Şeker	3
Çay	2
Ekmek	4
Makarna	4
Peynir	4
Deterjan	1
Bira	1
Zeytin	1

Örnek 1.

- c) Bu tablo üzerinde bazı ürünler eşik değere göre çıkarılır. Eşik destek sayısı 3 olduğuna göre bu eşik değerden küçük desteği sahip olan ürünler çözümlemeden çıkarılır. Buna göre oluşan yeni tablo aşağıdadır.

Ürün	Sayı
Şeker	3
Ekmek	4
Makarna	4
Peynir	4

Örnek 1.

- d) Çözülemeye katılacak ürünler bu şekilde belirlendikten sonra ikili gruplar oluşturarak bu grupların destek sayıları hesaplanır.

$\text{sayı}(\text{şeker},\text{ekmek})=2$

$\text{sayı}(\text{şeker},\text{makarna})=2$

$\text{sayı}(\text{şeker},\text{ekmek})=2$

$\text{sayı}(\text{şeker},\text{peynir})=2$

$\text{sayı}(\text{ekmek},\text{makarna})=3$

$\text{sayı}(\text{ekmek},\text{peynir})=3$

$\text{sayı}(\text{makarna},\text{peynir})=4$

Örnek 1.

- İkili ürün gruplarının destek değerleri

Ürün	Sayı
Şeker,Ekmek	2
Şeker,Makarna	2
Şeker,Peynir	2
Ekmek,Makarna	3
Ekmek,Peynir	3
Makarna,Peynir	4

Örnek 1.

- e) Bu tablodan bazı ürünler eşik değerine göre çıkarılır. Buna göre,

Ürün	Sayı
Ekmek,Makarna	3
Ekmek,Peynir	3
Makarna,Peynir	4

Örnek 1.

- f) Çözümlemeye katılacak ürünler bu şekilde belirlendiğine göre bu ürünlerin üçlü gruplar oluşturulur.

$\text{sayı(ekmek,makarna,şeker)}=1$

$\text{sayı(ekmek,makarna,çay)}=1$

$\text{sayı(ekmek,makarna,peynir)}=3$

...

$\text{sayı(ekmek,peynir,şeker)}=1$

$\text{sayı(ekmek,peynir,deterjan)}=1$

...

$\text{sayı(makarna,peynir,şeker)}=2$

$\text{sayı(makarna,peynir,çay)}=1$

...

Örnek 1.

- Eşik destek sayısına göre kalan üçlü ürün grupları aşağıdadır.

Ürün	Sayı
Ekmek, Makarna, Peynir	3

- Bu aşamadan sonra birliktelik kuralları elde edilebilir.

Örnek 1.

$$\text{sayı}(A, B) = \text{sayı}(ekmek, makarna, peynir) = 3$$

$$\begin{aligned} \text{destek}(A \rightarrow B) &= \frac{\text{sayı}(ekmek, makarna, peynir)}{N} \\ &= \frac{3}{5} = 0.6 \end{aligned}$$

biçiminde kural destek ölçütü elde edilir. Bu destek ölçütü koşul olarak verdiğimiz eşik değerden küçük değildir. O halde bu nesne kümesini kullanabileceğimiz anlaşılır. Kural destek sayılarına bağlı olarak birliktelik kuralları türeterek bu kurallar için güven ölçütlerini elde edeceğiz.

Sonuç 1:

$$\begin{aligned} \text{güven}(Ekmek, makarna} \rightarrow \text{peynir}) &= \frac{\text{sayı}(Ekmek, makarna, peynir)}{\text{sayı}(Ekmek, makarna)} \\ &= \frac{3}{3} = \%100 \end{aligned}$$

Örnek 1.

Sonuç 2:
$$\text{güven}(Ekmek \rightarrow peynir, makarna) = \frac{\text{sayı}(Ekmek, makarna, peynir)}{\text{sayı}(Ekmek)}$$
$$= \frac{3}{4} = \%75$$

Sonuç 3:
$$\text{güven}(peynir \rightarrow ekmek, makarna) = \frac{\text{sayı}(Ekmek, makarna, peynir)}{\text{sayı}(peynir)}$$
$$= \frac{3}{4} = \%75$$

Sonuç 4:
$$\text{güven}(makarna \rightarrow ekmek, peynir) = \frac{\text{sayı}(Ekmek, makarna, peynir)}{\text{sayı}(makarna)}$$
$$= \frac{3}{4} = \%75$$

Örnek 1.'e ait Birliktelik Kuralları

Birliktelik kuralı	Anlamı	Güven
Ekmek&Makarna → Peynir	Ekmek ve Makamanın bulunduğu ürün kümesinde Peynirin olma olasılığı	%100
Ekmek → Peynir&Makama	Ekmeğin yer aldığı bir ürün kümesinde peynir ve makarnanın olma olasılığı	%75
Peynir → Ekmek&Makarna	Peynirin yer aldığı bir ürün kümesinde ekmek ve makarnanın olma olasılığı	%75
Makama → Ekmek&Peynir	Makamanın yer aldığı bir ürün kümesinde ekmek ve peynirin olma olasılığı	%75

METİN MADENCİLİĞİ

Metin Madenciliği, işletme dokümanları, müşteri yorumları, web sayfaları ve XML dosyalarını içeren, yapısal olmayan verilerden, önceden bilinmeyen, potansiyel olarak kullanışlı bilgiyi keşfetme sürecidir. Elde edilen bilgiyle, analiz edilecek olan metin kaynaklarında açık olarak görülmeyen ilişkiler hipotezler veya eğilimler olduğu anlaşıılır (MECCA, RAUNICH, & PAPPALARDO, 2007; WITTEN, 2003)

Metin Madenciliği, Metin Veri Madenciliği (Text Data Mining) ve Metin Veri tabanlarından Bilgi Keşfi (Knowledge Discovery from Textual Databases) olarak da adlandırılır (DELEN & CROSSLAND, 2008).

Metin Madenciliği, işletme arşivinde veya internet üzerindeki belgelerde bu belgeye benzer belgelerin olup olmadığı elle bir sınıflandırma gerekmeden benzerliği hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar kelimelerin tekrarı sayesinde yapılır (ALPAYDIN, 2000).

Metin madenciliği, veri madenciliğinin bir parçası olarak düşünülmesine rağmen alışlagelen veri madenciliğinden farklıdır. Ana farklılık, metin madenciliğinde örüntülerin olay tabanlı veri tabanlarından daha çok, *doğal dil metinlerinden* çıkartılmasıdır (DELEN & CROSSLAND, 2008).

Metin madenciliğinin yararları, metinsel verilerin büyük bir çoğunluğunun işletme işlemlerinden elde edildiği alanlarda açıkça görülür. Örneğin, müşteri serbest formundaki şikayet ve memnuniyet metinlerinden gelen anlamlı bilgiler, ürün geliştirme, hata izleme garanti süresi gibi konularda işletmeye girdi oluşturur (DELEN & CROSSLAND, 2008).

Metin Madenciliğinin ne yaptığına bakıldığında en temel seviyede yapısal olmayan metin belgelerini sayısallaştırıp daha sonra veri madenciliği araç ve tekniklerini kullanarak onlardan anlamlı örüntüler çıkarttığı görülür. Başka bir deyişle metin madenciliği, en genel haliyle doğal dilde yazılmış metinler içinden,

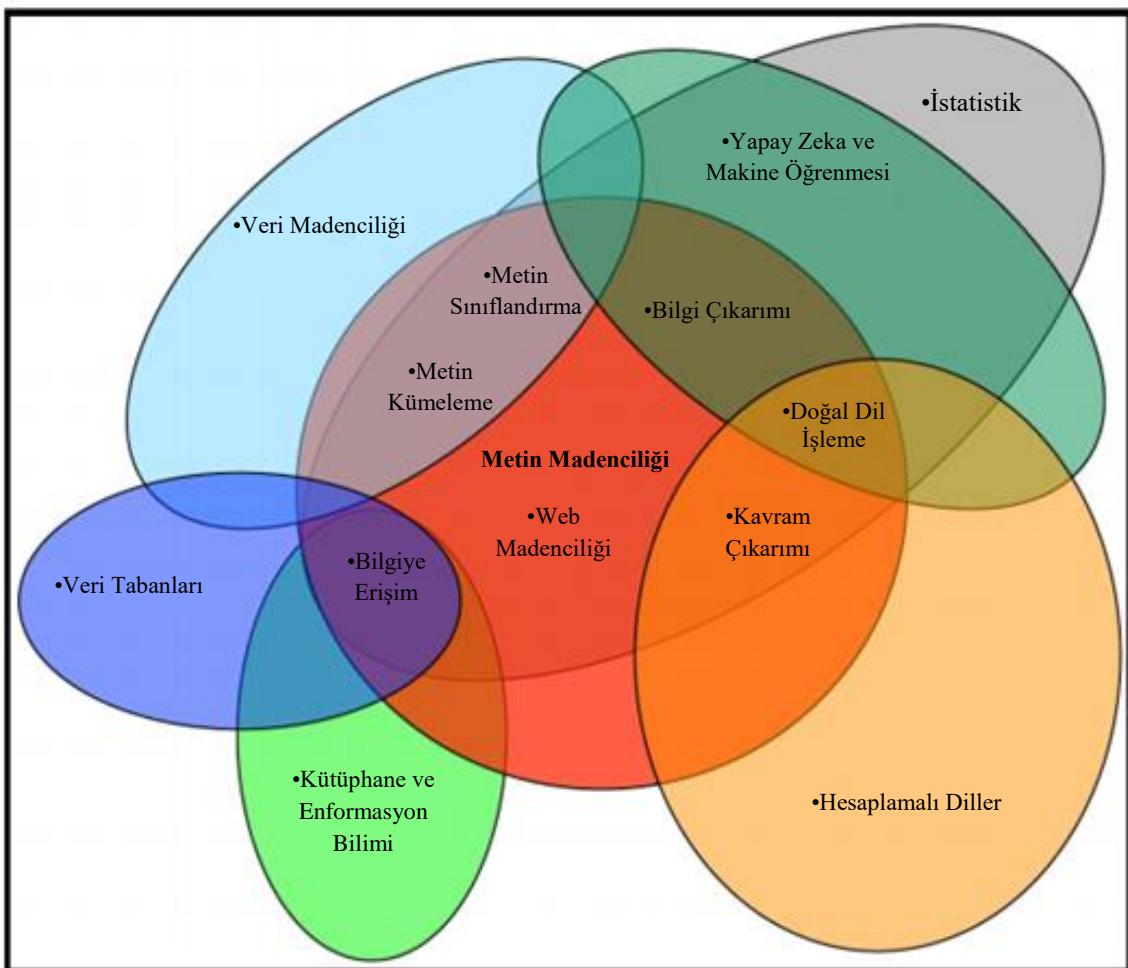
- aynı konudaki belgeleri bulur,
- birbirile ilişkili belgeleri bulur,
- ve bulunan belgeleri sıralar.

Daha ileri düzeyde bakıldığında Metin Madenciliği teknikleri belgeyi özetlemek, bilgi çıkarmak amacıyla da kullanılır.

Metin madenciliğinin yararları, metinsel verilerin büyük bir çoğunluğunun işletme işlemlerinden elde edildiği alanlarda açıkça görülür. Örneğin, müşteri serbest formundaki şikayet ve memnuniyet metinlerinden gelen anlamlı bilgiler, ürün geliştirme, hata izleme garanti süresi gibi konularda işletmeye girdi oluşturur (DELEN & CROSSLAND, 2008; MOHAMMAD, 2007).

Yapısal olmayan metinleri otomatik işlemenin kullanıldığı diğer alan, elektronik iletişim ve e-maillerdir. Metin madenciliği yalnızca sınıflandırmaya ve junk mailleri filtrelemeye yardım etmez aynı zamanda e-maillere otomatik olarak cevap vermekte de kullanılır. Metin madenciliği yargı, sağlık ve diğer endüstrilerde geleneksel olarak zengin belgeler ve sözleşmelerle elde edilen verilere de ulaşmayı sağlar (DELEN & CROSSLAND, 2008) (MOHAMMAD, 2007).

Miner vd.(2012)'e göre Metin Madenciliğinin ilişkili olduğu disiplinler ve yöntemler Şekil 1.'de gösterilmiştir (MINER, DELEN, ELDER, FAST, HILL, & NISBET, 2012)



Şekil 1. Metin Madenciliği ve İlişkili Olduğu Alanlar

Başka bir açıdan bakıldığında giderek artan belge yiğinlarının faydalı bilgiye dönüştürülmesini sağlamak için geliştirilen Metin Madenciliği çalışmaları, Bilgiye Erişim (Information Retrieval) ve Bilgi Çıkarımı (Information Extraction) olmak üzere iki alanda incelenmektedir.

Zohar'a (2002) göre Metin Madenciliği metodları,

- Bilgiye Erişim (Information Retrieval),
- Bilgi Çıkarımı (Information Extraction),
- Web Madenciliği (Web Mining),
- Kümeleme (Clustering),

olmak üzere dört grupta toplanmaktadır (ZOHAR, 2002). Buna göre Tablo 1.'de bu metodların girdi ve çıktıları özetlendiği gibidir.

Tablo 1. Metin Madenciliği Metotlarının Girdi ve Çıktıları (ZOHAR, 2002)

Bilgiye Erişim	Bilgi Çıkarımı	Web madenciliği	Kümeleme
<p><i>Girdi:</i> Metin Belgesi Kaynağı, Kullanıcı sorusu (metin tabanlı)</p> <p><i>Çıktı:</i> Soru ile ilişkili olan sıralanmış belgeler kümesi</p>	<p><i>Girdi:</i> Metinsel belgeler kaynağı</p> <p>İyi tanımlanmış sınırlandırılmış soru</p> <p><i>Çıktı:</i> İlişkili bilgi cümleleri</p> <p>İlişkili bilginin çıkarımı ve ilişkili olmayan bilginin yok sayılması</p> <p>Önceden belirlenmiş formatta çıktı ve ilgili bilgi linki.</p>	<p>Webteki özel bilginin çıkarımı ve metinsel belgelerin erişimi ve indekslenmesi</p>	<p>Benzer metin belgelerinin toplanması</p>

Bilgiye Erişim (Information Retrieval)

Bilgiye Erişim kavramı ilk kez Calvin Mooers tarafından 1948 yılında “Application of Random Codes to the Gathering of Statistical Information” başlığını taşıyan yüksek lisans tezinde *Information Retrieval* terimi altında kullanılmıştır. Vickery, Mooers’ın kavrama İngilizce olarak getirdiği ilk tanımı şu şekilde aktarır. Bilginin bir depodan özelliklerine göre konusal olarak aranarak erişilmesidir (TÜRKEEŞ, 2007).

Bilgiye Erişim (IR), metin madenciliğinde ilk adımdır. IR’ın amacı kullanıcıların bilgi ihtiyaçlarını karşılayacak olan belgeleri bulmasına yardımcı olmaktadır.

IR, birçok konu alanına sahipliği nedeniyle geniş bir alana yayılmaktadır ve kullanıcıların belirli konulardaki belgeleri bulabilmesi gibi büyük bir topluluktan oluşan metni sunması için modeller geliştirmiştir. Problem, kullanıcı şu an ne ile ilgilenmekte ve belirli bir konu kümesi hakkında belgeler nasıl sunulmalı ve tanımlanması gibidir (SEZER, 2006).

Bilgiye Erişim, bilgi ihtiyacını karşılayan yapılandırılmamış materyalleri (genellikle dokümanlar) geniş bir koleksiyonun içerisinde bulmaktadır. Eskiden bilgiye erişim sadece bazı meslek grupları tarafından özel amaçlar için kullanılmaktaydı. Fakat değişen günümüz dünyasında, milyonlarca insan mail ve web aramaları için kullanmaktadır. Böylelikle IR geleneksel veri tabanı arama yöntemlerinin önüne geçmeye başlamıştır.

Bilgiye Erişim sistemlerinde kullanılan standart iki ölçü vardır (CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Recall (Doğruluk): Araştırmacı tarama yaptığı konularda bütün kaynaklara erişmek istemektedir. Bilgi sistemlerinde araştırmacının bu isteğinin karşılanma derecesi Recall ile ifade edilir. Recall, bir bilgi sisteminin soru ile ilgili olarak bulduğu yayınların içindeki gerçekten soru ile ilgili olan yayınların sayısının veritabanında bulunan ilgili yayınların

sayısına oranını gösterir (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

$$Recall = \frac{\text{Veritabanı içinde dönen ilgili belge sayısı}}{\text{ilgili toplam doküman}}$$

Precision (Duyarlılık): Araştırmacı istediği bilgileri çok fazla zaman harcamadan bulmak istemektedir. Zaman söz konusu olunca ilk akla gelen bilgi sisteminin tarama hızıdır. Ancak hızlı bir tarama sistemi araştırmacının amacı açısından yeterli değildir. Araştırmacının bilgi sisteminin kendisine sorgu ile ilgili olarak gösterdiği yaynlarda gerçekten ilgili olanları seçmesi gerekmektedir. Araştırmacının zamanının büyük bir kısmı da bu evrede harcanmaktadır. Araştırılan yaynları bulma süresini doğrudan etkileyen ve tarama sonuç listesinin iyiliğini gösteren bu özellik ise Precision olarak adlandırılır. Precision bir bilgi sisteminin sorgu ile ilgili olarak bulduğu yaynların içindeki kullanıcının istediği yaynların sayısının bulunan yaynların sayısına oranıdır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

$$Precision = \frac{\text{Veritabanı içinde dönen ilgili belge sayısı}}{\text{geri dönen doküman}}$$

Recall ve *Precision* ölçümlerinin her ikisini birden artırmak bilgilerin tasnif edilmesi ile olur. Bu konudaki robotların *Recall* ve *Precision* oranları düşüktür. Kütüphaneler ise yüksektir.

Bilgiye Erişim sistemlerinde ağırlık(w) verme önemli bir rol oynar ve birçok farklı ağırlık verme modeli geliştirilmiştir. En yaygın olarak kullanılan model, yerel(local) ve genel(global) ağırlık verme şemalarının bir arada kullanılmasıdır. Yerel ağırlık vermede terim frekansı (Term Frequency, TF), genel ağırlık vermede ise ters doküman frekansı (Inverse Document Frequency, IDF) kullanılır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Terim Frekansı (TF), bir doküman içerisinde bir terimin tekrar sıklığıdır. Ters Doküman Frekansı (IDF) bir terimin bütün doküman koleksiyonu (D) içindeki önemidir.

Bu modele göre, terimin önemi, belge içerisinde o terimin geçme sayısıyla doğru orantılıyken; bütün belge havuzu içerisinde o terimin geçme sıklığıyla ters orantılıdır. D belgesinde, i teriminin ağırlığı şu şekilde hesaplanır

$$w_i = tf_i \times \log \frac{D}{df_i}$$

Frekansı düşük olan terimler için IDF skoru yüksek, frekansı yüksek olan terimler için IDF skoru düşüktür.

TF-IDF değeri, az miktarda doküman içerisinde terim yüksek miktarda geçiyor ise yüksek değer alır. Eğer terim her dokümanda geçiyorsa TF-IDF değeri en düşük değerini alır (PİLAVCILAR, 2007; SEZER, 2006; CAN, KOÇBERBER, BALÇIK, KAYNAK, ÖÇALAN, & VURSAVAŞ, 2008).

Bilgi Çıkarımı (Information Extraction)

Bilgi çıkarımı en basit şekliyle geniş ölçekli bilgilerden özet çıkarılması olarak adlandırılabilir. Başka bir ifadeyle büyük veri yığınları içerisinde özet bilgiler elde edilmesidir. Anahtar kelimeler veya örnek dokümanlar gibi kullanıcı girişleriyle bağlantılı olan bilgi ya da dokümanların bulunması bilgi çıkarımı örnekleridir. Bu çalışmalar sonucunda web sayfalarından bilgiler karşılaştırılarak bulunabilir, geniş ölçekli metinlerden özet bilgiler çıkarılabilir, sorgulara karşılık gelen ifadeler bulunabilir (SODERLAND, 1999).

Bilgi Çıkarımı, yöntemleri metin içindeki unsurları varlıklarını otomatik olarak çıkarır ve bunlar arasındaki ilişkileri ortaya koyar. Metin içindeki cümleler ve paragraflar içerdikleri önermelerle varlıklara ait bilgiler taşır. Bilgi Çıkarımı teknikleri bu önermelere bağlı olarak belgeyi oluşturan varlıklarını ve bu varlıklar arasındaki ilişkileri çıkarırlar (DAŞ, 2008; KAISER & MIKSCH, 2005).

Bilgi çıkarım işleminin en zor adımlarından birisi de veriyi işlerken belirli bir yapıya oturtmaktadır. Örneğin internet üzerinde yayınlanan verilerin herhangi bir standart yapısı bulunmamakta, veriler dağınık halde istenildiği gibi yayınlanmaktadır.

Bilgiye Erişim yöntemlerine nazaran daha etkin sonuçlar elde edilmesini sağlayan bilgi çıkarma tekniklerinin avantajı belge içindeki içeriğin anlamını ön plana çikan terimlerin ve terimler arası ilişkilerin bulunmasında yatar. Ancak bazen belgelerin incelenmesindeki amaç daha önceden fark edilmemiş gerçeklerin ve ilişkilerin ortaya çıkarılmasıdır. Bu aşamada devreye bilgi keşfi teknikleri girer. Bilgi keşfi için kullanılan yöntemler metnin içeriklerini derler, birbiri ile entegre eder ve başka kaynaklardan elde edilen sonuçlarla birleştirilerek üst seviye bir anlam ve ilişki kümesi oluşturmaya çalışır. Özellikle konuya bağlı olarak terimler ve terimler arası ilişkilerin üzerine de çıkarılır ve konuya özel yapılar ve fonksiyonlara bağlı bir ilişki kümesi oluşturulur. Bu amaçla geliştirilen sistemlerin sadece belgeleri değil veri tabanlarındaki verileri de kullanması gereklidir (KUSHMERICK, 1997).

Bilgi çıkarım işlemi, temelde anahtar kelime ve/veya benzerlik tabanlı çıkarımlara dayanmaktadır (KUSHMERICK, 1997). Anahtar kelime tabanlı bilgi çıkarımında, herhangi bir doküman ya da metinden bilgi çıkarılırken anahtar kelimelerden oluşan bir kümeye oluşturulur.

Benzerlik tabanlı çıkarım sistemleri ortak anahtar kelimeler kümesini temel alarak, benzer dokümanları bulmaktadır. Bu tür bir çıkarımın çıktısı, kelimelere yakınlığı ve birbirleriyle ilişki derecelerini temel almaktadır. Günümüzde internet ve bilgi teknolojilerinin hızla gelişmesi ve insanların hayatında önemli bir yer tutması sebebiyle, bu ortamlardan bilgi çıkarımı önem kazanmıştır. Herhangi bir ürünün satış sitelerinden aranması ve karşılaşmalıdır olarak değerlendirilmesinden, elektronik posta içeriklerinin yorumlanmasına kadar çeşitli uygulamalar internetten bilgi çıkarımı işlemine örnek olarak düşünülebilir.

Bilgi Çıkarım sistemi sonuçlarının değerlendirilmesinde bilgi erişim sistemlerinde de olduğu gibi duyarlık ve doğruluk ölçütleri kullanılmaktadır. Fakat burada belgeler yerine, yapılan tahminler ölçüm değişkenleri olarak kullanılmaktadır. Duyarlık, sistemin doğru yaptığı tahminlerin tüm tahminlere bölümü ile hesaplanmaktadır. Doğruluk ise sistemin yaptığı doğru tahminlerin metinde bulunan bütün varlıkların sayısına bölünmesi ile elde edilmektedir (OFLAZER, 2002; GÜVEN, 2007)

Bilgiye Erişim ve Bilgi Çıkarımının Karşılaştırılması

Bilgi Çıkarımı, bilgi parçalarını çıkarmak için doğal dil işlemeyi temel alan bir teknolojidir. Bu süreç girdi olarak metinleri ele alır ve çıktı olarak belirli bir formatta açık şekilde ifade edilebilecek veriler üretir. Bu veri kullanıcıların görüntü elde etmesi için doğrudan kullanılabilir veya daha sonra analiz etmek için veri tabanında veya elektronik tablolarda saklanabilir veya Google gibi internet arama motorlarında olduğu gibi Bilgiye Erişimi uygulamalarında dizinleme amaçlarını yerine getirmek için kullanılabilir (TURMO, AGENO, & CATALA, 2006).

Bilgi Çıkarımı, Bilgiye Erişimden oldukça farklıdır;

- Bilgiye Erişim sistemi uygun metinleri bulur ve bunları kullanıcıya sunar.
- Bilgi Çıkarımı uygulaması metinleri analiz eder ve sadece kullanıcıların ilgilendikleri metinlerden özel bilgi elde ederler (TURMO, AGENO, & CATALA, 2006).

Örneğin, tarım ürünlerini ilgilendiren ticari grup yapılarından bilgi bekleyen bir Bilgiye Erişim sistemi kullanıcıyı uygun kelime listesini girecektir ve karşılığında olası eşleşmeleri içeren belge kümese (örneğin gazete makaleleri) ulaşacaktır. Daha sonra kullanıcı belgeleri okuyacaktır ve bilgilerin içerisinde bir ayıklama işlemi gerçekleştirilecektir. İşlem uygulandıktan sonra elektronik tablo halinde bilgi girişi yapılabilir ve bunlardan rapor veya sunu çizelgeleri oluşturulabilir. Bunun tersine bir Bilgi Çıkarımı sistemi uygun şirket ve grup adlarını doğrudan ilgilendiren değerleri otomatik olarak elektronik tablo halinde sunacaktır (TURMO, AGENO, & CATALA, 2006).

Metin Madenciliğinin Diğer Uygulamaları

- 1) **Konu izleme:** Kullanıcı profillerini kullanarak ve kullanıcı görüşlerinden oluşturulan belgelere bağlı olarak kullanıcı için ilginç olabilecek diğer belgelerin tahmin edilmesidir (DELEN & CROSSLAND, 2008). Sosyal ağlarda kullanıcı profillerine göre farklı kullanıcılarla farklı reklamların gösterilmesi bu konuya örnek olarak verilebilir.
- 2) **Özetleme (Summarization):** Okuyucuya zaman kazandırmak amacıyla belgenin aslini bozmadan metnin özeti alınması olarak tanımlanabilir. Başka bir deyişle otomatik metin özetteme bir bilgisayar programı aracılığıyla istenilen metinlerin özetinin çıkarılmasıdır. Belge özetlemenin amacı bir belgenin amacını anlatan kısa bir özetinin otomatik olarak oluşturulmasıdır. Etkin bir özetteme sistemi kullanıcıların arama sonucu olarak elde ettikleri belgelerin özetlerine bakarak tüm belgeyi inceleme zorunluluğu olmadan doğru belgeye ulaşıp ulaşmadıklarını belirleyebilmeleridir (DELEN & CROSSLAND, 2008; TÜLEK, 2007).

Bu konudaki ilk çalışma 1959 yılında Luhn adlı bir bilim adamı tarafından yapılmıştır. Luhn kelimelerin kullanım frekansından yararlanmıştır (TÜLEK, 2007).

Metin Özetteme çeşitleri iki grupta toplanabilir.

- Cümle Seçerek Özetteme (Extract Summarization)
- Yorumlayarak Özetteme (Abstract Summarization)

Cümle seçenek özetlemede (Extract Summarization) özetlenecek metin önemli cümleler, istatistiksel metodlarla, sezgisel çıkarımlarla veya bunların ikisinin kombinasyonıyla seçilerek bu cümlelerden oluşan bir özetleme yapılır. Bu özetlemede özeti oluşturan cümleler içeriği akıllıca değişik şekilde anlatan cümleler değil, yazı içinden seçilmiş olan önemli cümlelerdir (TÜLEK, 2007).

Yorumlayarak özetlemede (Abstract Summarization) özetleme, özetlenecek metnin akıllıca yorumlanması ile yapılır. Bu özetlemede orijinal metindeki ifadeler akıllı bir şekilde kısaltılarak tekrar yazılmasına çalışılır (TÜLEK, 2007).

- 3) **Sınıflandırma:** İçinde önceden tanımlanmış konu kategorilerinin yer alacağı şekilde bir belgenin ana temalarının tanımlanmasıdır. İçerik bazlı belge yönetimi işi belgelere ulaşımda esnekliği amaçlamaktadır. Metin sınıflama çalışması bu amaca ulaşmak için kullanılan bir adımdır ve konuşma dili ile yazılmış metinleri anlamlarına göre daha önceden belirlenmiş sınıflara ayırmaya çalışır. Günümüzde metin sınıflama kontrollü bir kelime haznesine bağlı olarak belgeleri indeksleme, belgeleri filtreleme, otomatik olarak metadata oluşturma web sayfalarını otomatik olarak hiyerarşik düzenlemeye tabi tutma gibi pratik olarak uygulanan pek çok alanda görmek mümkündür (DELEN & CROSSLAND, 2008; TÜLEK, 2007).
- 4) **Kümeleme:** Metin madenciliğindeki önemli noktalardan biri de kümeleme metodlarıdır. Kümeleme önceden belirlenmiş bir kategoriler kümese sahip olmaksızın birbirine benzer belgelerin gruplandırılmasıdır. Karar ağaçları, makine öğrenmesi, istatistik gibi çeşitli teknikler bu nokta için kullanılmaktadır. Bunların içinden en önemlileri, karar ağaçları, yapay sinir ağları bulanık mantık, yaklaşımı kümeler ve içerik öğrenmedir. Benzer belgelerin aranması da metin madenciliği uygulamasıdır ve benzer olarak ön işleme ve sınıflandırma kümeleme aşamalarını içerir (AMASYALI, 2008).

Başka bir deyişle kümeleme verilerin kendi aralarındaki benzerlikleri göz önüne alınarak gruplandırılması işlemidir.

Herhangi iki doküman arasındaki benzerlik, dokümanların Vektör Uzay Modeli ile vektör haline getirilmesinden sonra **Kosinüs Benzerliği** ile hesaplanır.

d_1 ve d_2 iki doküman vektörü olduğunda, kosinüs benzerliği şu şekilde hesaplanır.

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \cdot \|d_2\|}$$

$d_1 \bullet d_2$: iki dokümanın vektör çarpımı

$\|d_i\|$: i dokümanın (vektörünün) uzunluğu (normu) (G. ÖĞÜDÜCÜ, 2011)

Kümeleme işlemlerinde değerlendirme için iki ölçüt vardır. Entropi ve F-Ölçütü (F-Measure).

Entropi: Rassal bir değişkenin belirsizlik ölçütü olarak bilinen Entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Entropi sadece nesnelerin baskın sınıfta olup olmadığını ölçmekle kalmaz, kümelerdeki sınıfların dağılımlarını da dikkate alır. Entropi değerinin 0 olması kümeyi tamamen tek bir sınıfından oluştugu gösterirken, 1'e yakın bir entropi değeri kümeyi bütün sınıfların tek düzeye (uniform) dağılıma göre oluşmuş bir karışımını içerdigini gösterir.

$$E = - \sum_{i=1}^n p_i * \log_2(p_i)$$

p_i : i mesajının üretilme olasılığı

F-Ölçütü: F-Ölçütü (F-Measure), yaygın olarak kullanılan diğer bir dışsal kalite ölçüm yöntemidir. Kümeleme Doğruluğu (clustering accuracy) olarak da bilinir. F-Ölçütü için Precision ve Recall değerlerinin hesaplanması gerekmektedir.

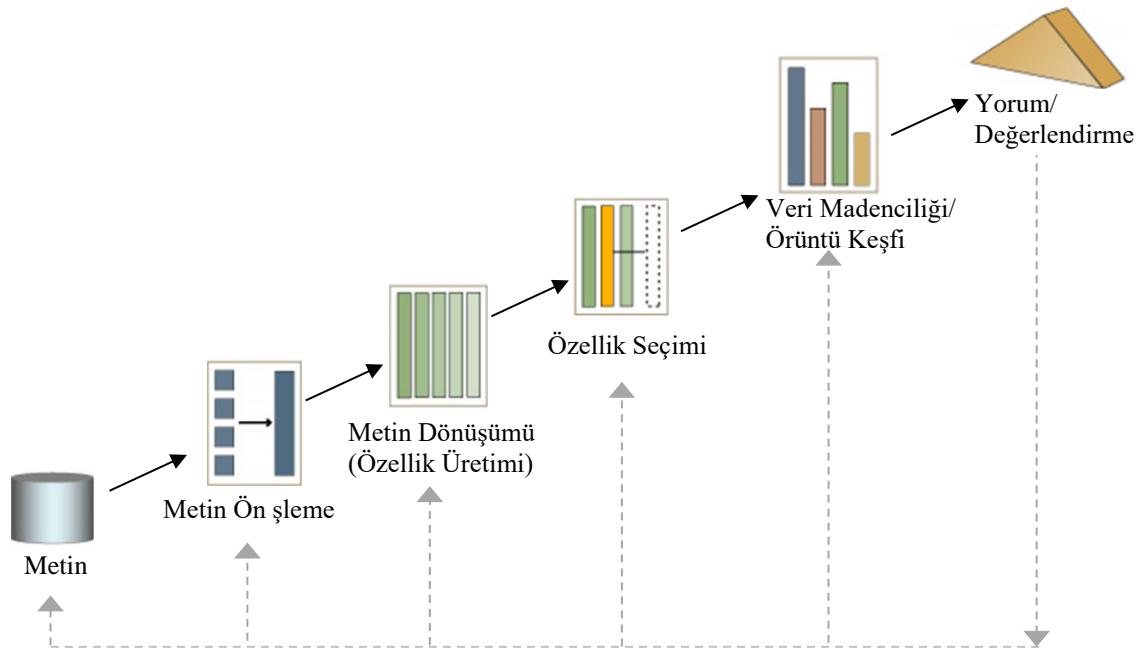
$$F(i,j) = \frac{2 * \text{Recall}(i,j) * \text{Precision}(i,j)}{\text{Recall}(i,j) + \text{Precision}(i,j)}$$

- 5) **Kavram bağlama-ekleme:** Yayın olarak paylaşılan kavramları tanımlayarak ve böyle yaparak da geleneksel arama metodlarını kullanarak aradıklarını bulamayan kullanıcılara yardım ederek ilişkili belgelerin bağlanmasıdır (DELEN & CROSSLAND, 2008).
- 6) **Soru cevaplama:** Bilginin yönlendirdiği örüntü eşleştirmeleriyle verilen bir soru için en iyi cevabin bulunmasıdır (DELEN & CROSSLAND, 2008).

Soru cevaplama sistemleri bilgiye ulaşma ihtiyacı ile ortaya çıkmış olan ve genelde bilgisayar destekli yapılardır. Soru-cevap benzerliklerini karşılaştırarak ya da varolan kaynaklar üzerinde yapay zeka gibi insan türevi teknikler uygulanarak sorulara yeni cevaplar üretmeye çalışan sistemler geliştirilmiştir (AMASYALI, 2008).

Metin Madenciliğinin Adımları

Büyük miktardaki metinsel verilerden potansiyel olarak yararlı ve önceden bilinmeyen belirli bir önemi olan bilginin çıkarılması olarak nitelendirilen Metin Madenciliği şekilde görüldüğü gibi temelde altı adımdan oluşmaktadır. Metin Madenciliği işlemleri, Veri Madenciliğine benzer olarak Şekil 2.'deki gibi özetlenebilir.



Şekil 2. Metin Madenciliğinin Adımları (ZOHAR, 2002)

Bu bilgilerden sonra Metin Madenciliği işlemleri ve içerdikleri yöntemler Tablo 2.'de görüldüğü gibi özetlenebilir. Bunlar hakkında detaylı bilgili izleyen kısımda yer almaktadır.

Tablo 2. Metin Madenciliği İşlemleri (ZOHAR, 2002)

Metin	Metin Ön İşleme	Metin Dönüşümü	Özellik Seçimi	Veri Madenciliği/ Bilgi Keşfi	Yorum/ Değerlendirme
	Söz dizimsel/ Semantik analiz Sözcük türü etiketleme Kelime anlamı belirginleştirme Ayristırma(parsing)	Kelime torbası, Kelimeler Kök bulma, Durdurma kelimeleri	Basit hesaplama İstatistik (boyut azaltma, ilişkisiz ozellikler)	Sınıflandırma (danışmanlı) Kümeleme (Danışmansız)	Analiz Sonuçları

Metin koleksiyonu oluşturma

İlgilenilen konularda bilgiye erişim sistemleri kullanılarak metin koleksiyonu oluşturma sürecidir. Bu süreç, günümüzde genel olarak internet üzerinden, özellikle Google vb. arama motorları kullanılarak gerçekleştirilmektedir. Çevrim içi veri tabanlarının yanı sıra veri tabanlarında ya da kişisel bilgisayarlarda bulunan metin türü veriler ile oluşturulan koleksiyonlar da metin madenciliğinde kullanılmaktadır (PİLAVCILAR, 2007; OĞUZ, 2009).

Metin önişleme

Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.), kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etmek ve var olan hataları düzeltmek gibi metin belgelerin yapıtaşları olan kelimelerle ilgili işlemleri içeren süreçtir.

Metin madenciliğinin en büyük sorunu işleyeceği veri kümесinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir (GÜVEN, 2007).

Belgeler için dizin oluşturmadan önce yapılacak ön işleme işlemleri şöyledir.

- Doküman doğrusallaştırma
 - ✓ *Markup & Format Removal:* Dokümanı oluşturan etiket ve özel formatların çıkarılması)
 - ✓ *Tokenization:* Metin küçük harflere çevrilmesi ve noktalama işaretlerinin çıkarılması)

Metin önişleme çalışmaları aynı zamanda doğal dil işleme çalışmaları kapsamında incelenen bir alandır. Doğal dil işlemenin belge analizi sürecindeki en önemli faydası terimlerin yanı kelimelerin ayırtılması, eklerinden arındırılarak anlamını kaybetmeyen en kısa biçimlerine dönüştürülmesidir. Çünkü aynı anlam için kullanılan kelimeler dilbilgisi kuralları gereği farklı biçimlerde bulunabilir ve bu farklı kullanım biçimleri ortadan kaldırılmadığı takdirde farklı anlam taşıyan terimler gibi işleme alınarak, belgelerin gerçek anlamına ulaşılmasını engelleyebilirler. Doğal dil işleme çalışmaları kapsamında yürütülen girişimler dört ana grup altında toplanabilir (ÖZBİLİCİ, 2006).

- a) Morfolojik (Biçimbirimsel) Analiz
 - b) Sözdizimsel (Sentaktik) Analiz
 - c) Semantik (Anlamsal) Analiz
 - d) Anlam kargasasının giderilmesi
-
- a) **Morfolojik Analiz:** Biçimbirim, sözcüklerin yapısıyla ile ilgilenir. Türkçe için sözcüklerin türetilmesi ve ekler çok önem taşır. Her dilde iki farklı şekilde sözcük oluşturulabilir. Bunlardan biri çekim, diğeri ise türetme yöntemidir. Çekim yoluyla sözcük oluşturulurken bir sözcüğün farklı şekilleri kullanılır. Türetme ise var olan eski sözcüklerde yapım ekleri eklenmesi yoluyla yeni sözcük oluşturma yöntemidir (ÖZBİLİCİ, 2006; NABIYEV, 2010; KESGIN, 2007).
 - b) **Sözdizimsel Analiz:** Bilgisayarla doğal dil modellemelerinde anlamsal analize geçmeden önce, kelimeler yiğininin geçerli bir cümle yapısı oluşturup oluşturmadığı kontrol edilmelidir. Rasgele kelimelerin yan yana gelmesiyle geçerli bir cümle meydana gelmeyecektir. Geçerli bir cümle yapısı oluşturulamadığı zaman, buradan anlam çıkarılmasını beklemek yanlış olacaktır (ÖZBİLİCİ, 2006; NABIYEV, 2010; KESGIN, 2007).

Sözdizimsel analiz, cümlenin yapısal bir tanımını oluşturabilmek için morfolojik analizin sonuçlarını kullanır. Bu işlemi yapmanın amacı, ardı ardına gelen kelime yiğinlarının bu

kelimeler yiğininin ifade ettiği cümle birimlerini tanımlayan bir yapıya dönüştürmektir. Cümle birimleri, kelimeler tamlamalar veya buna benzer cümle parçacıkları olabilir.

- c) **Semantik Analiz:** Bir cümlenin ne demek istedığının anlaşılması, diğer bir deyişle bir cümle ile ifade edilmek istenilen duyu veya düşüncenin ne olduğunu anlaşılması, anlamsal analiz yardımıyla yapılır.
- d) **Anlam kargaşasının giderilmesi:** Anlamsal analiz yapılrken, öncelikli olarak kelimelerin tek tek veritabanından uygun nesnelerle eşleştirilme işleminin yapılması gereklidir. Bu işlem, her zaman birebir eşleme olamayabilir. Diğer bir deyişle, kelimelerin ifade ettikleri anlamlar her zaman bir tane olmayabilir. Ayrik kelimelerin bir cümledeki doğru anlamını bulma işlemine “kelime anlam berraklaştırılması” denir. Bu işlem, cümle içinde geçen bir kelimenin sözlükteki anlamlarının belirlenip bunlardan uygun olanının seçilmesidir. Cümle içinde geçen her bir kelime, diğer kelimelerin doğru anımlarının ortaya çıkarılması için önem taşımaktadır (DELİBAŞ, 2008; ÖZBİLİCİ, 2006; NABIYEV, 2010; SAY, 2003).

Metin dönüşümü

Kelimelerin düzgün bir biçimde hecelerine ve eklerine ayrılmadan sonraki işlem, kelimelerin kökünün tespit edilmesidir. İngilizce için kullanılan Porter Stemmer Yöntemi gibi bir kök bulma algoritması kullanmak hızlı olması açısından önemli olsa da Türkçe gibi sondan ekli bir dilde başarı yüzdesi istenen düzeyde değildir ve özel durumları yakalayamamaktadır.

Snowball: Kök bulmak için tasarlanmış küçük bir karakter işleme dilidir. Snowball kullanılarak birçok dil için kök bulma algoritmaları geliştirilmiştir. Türkçe için snowball kullanılarak geliştirilen kök bulma algoritmaları Evren Kapusuz tarafından yürütülmektedir (AYDIN & KILIÇASLAN, 2010; ORHAN, 2006).

Kelime Türü: Bir kelimenin kökü bulunduktan sonraki adım kelimenin türünün bulunmasıdır. Bu işlemde Pos Tagging denir. Pos Tagging 2 fazdan oluşur. Birincisi eğitim(training) fazıdır. Bu fazda kelimelerin kökleri manuel olarak tanımlanmış algoritmalar kullanılarak machine learning sistemi vasıtıyla işlenir. İkinci faz ise tagging fazıdır. Bu fazda, birinci adımda kullanılan algoritma, öğrenilen parametrelerle göre yeniden işlenir ve kelimeler türlerine ayrılır.

Stopword İşlemi: Tekrar eden ve tek başına anlam taşımayan kelimelere stopword kelimeleri denir. Bilgiye Erişimde bir stopword listesi, belgeleri bir diğerinden ayırt etme durumuna etkisi olmayan sıklıkla kullanılan kelimeleri içerir. Stopword kelimelerini azaltmak sorgu sürecinin verimini arttırır. Bir stopword listesinin yapılandırılması farklı ve bazen rastgele kararları içerir. Bilgiye Erişim literatüründe, verilen özel diller için farklı uzunluklarda stopword listelerini bulmak mümkündür.

Bag of words: Bu aşamada gruplanan tüm dokümanlardaki tüm kelimelerin kullanım sıklıkları hesaplanır ve bir havuzda toplanır. Daha sonrasında ise bu kelimelerin değerleri (Word Weighting) hesaplanır. Kelime değeri, bir kelimenin belirli bir alan (sağlık, spor, politika, ...) ile ilgili bir metnin içinde bulunma sıklığı olarak açıklanabilir. Örneğin 10000 kelimelik spor kategorisindeki bir haberin içinde gol veya hakem kelimelerinin bulunma

sıklıkları, aynı kelimelerin sağlık kategorisindeki bir haber içinde bulunma sıklığına göre kat be kat fazladır (ORHAN, 2006; KARADENİZ, 2007).

Özellik seçme

Metin madenciliği uygulamalarında her zaman gürültülü ve önemsiz bilgi içeren metin koleksiyonlarıyla uğraşma ihtiyacı bulunmaktadır. İlgili verilerin saptanması üzerine odaklanan özellik seçme, büyük miktarlardaki veriler üzerinde işlem yapılırken iş yükünü azaltmada yardımcı olmaktadır. Özellik seçme aşamasında, ön işlemden geçen metinlerdeki önemli kelimeleri (varlıklar) belirleme (isimler, tamlamalar, bileşik kelimeler, kısaltmalar, sayılar, tarihler, para birimleri vb.) ve ilişkili olmayan özelliklerin çıkarılması, sadece birkaç dokümda gözlemlenen özelliklerin çıkarılması, birçok dokumanda gözlemlenen özellikleri azaltma vb. işlemleri yapılmaktadır (CEBİROĞLU, TANTUĞ, ADALI, & ERENLER, 2003; ERYİĞİT, 2006).

Veri Madenciliği/Bilgi Keşfi

Metinsel verilerden bilgi keşfi için veri madenciliğinde geçen Sınıflandırma ve Kümeleme yöntemleri kullanılabilir. Sınıflandırma yöntemleri şu şekilde özetlenebilir.

- Entropiye Dayalı algoritmalar (ID3, C4.5)
- Sınıflandırma ve Karar Ağaçları (Twoing, Gini,)
- Bellek tabanlı sınıflandırma modelleri (En yakın komşu algoritması)
- Optimizasyon tabanlı Sınıflandırma Modelleri (Destek Vektör Makinesi)
- İstatistiksel Sınıflandırma Modelleri (Navie Bayes)

Kümeleme yöntemleri de aşağıda sıralandığı gibi özetlenebilir.

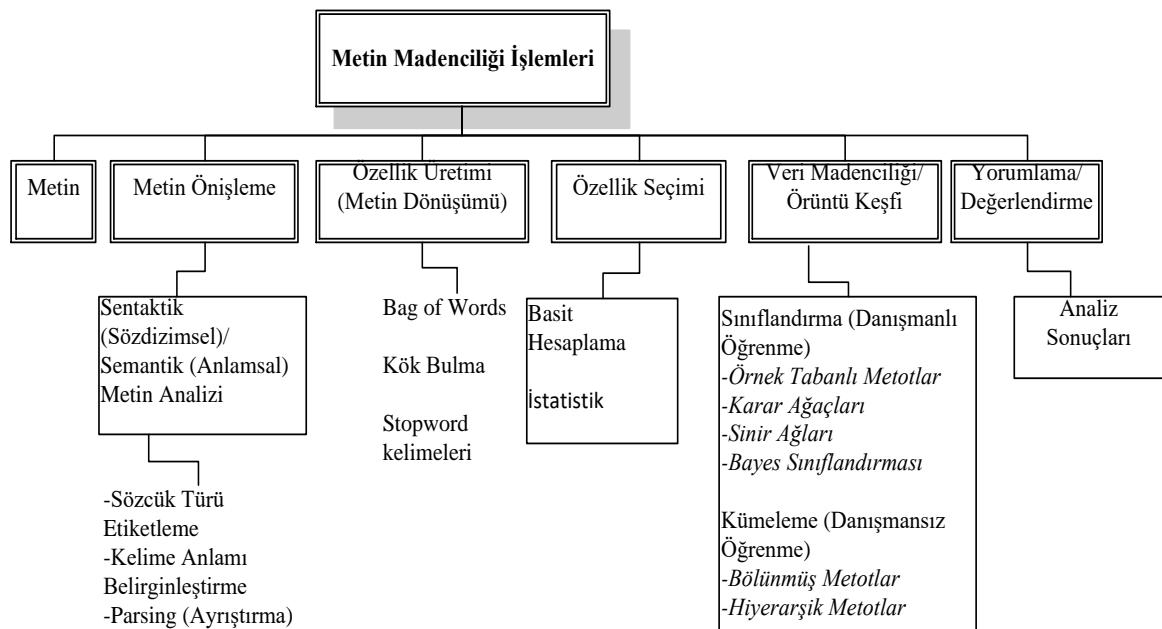
- Hiyerarşik Metotlar (En yakın komşu algoritması, En uzak komşu algoritması,)
- Hiyerarşik olmayan Metotlar (k-ortalamalar)

Veri Madenciliği için belirtilen bu sınıflandırma ve kümeleme yöntemleri ön işleme adımlarından geçirilerek metinsel verilere uygulanmaktadır (ÖZKAN, 2013).

Değerlendirme ve yorumlama

Veri madenciliği yöntemleri ile verilerin analizinden elde edilen sonuçların değerlendirilip kullanıcıya uygun ve anlaşılır bir şekilde sunulması işlemidir.

Metin Madenciliği işlemleri ve içerdikleri yaklaşımlar Şekil 3.'teki gibi özetlenebilir.



Şekil 3. Metin Madenciliği Adımları ve İçerdikleri Yaklaşımlar (ZOHAR, 2002)

KAYNAKLAR

- AKAT, Ö., TAŞKIN, Ç., & ÖZDEMİR, A. (2006). Uluslararası Alışveriş Merkezi Tüketicilerinin Satın Alma Davranışı: Bursa İlinde Bir Uygulama. *Uludağ Üniversitesi Sosyal Bilimler Dergisi*, 2006(2).
- ALPAYDIN, E. (2000). *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*. Bilişim 2000 Eğitim Semineri.
- AMASYALI, M. F. (2008). *Yeni Makine Öğrenmesi Metotları ve İlaç Tasarımına Uygulamaları*. İstanbul: Yıldız Teknik Üniversitesi FBE, Doktora Tezi.
- AYDIN, Ö., & KILIÇASLAN, Y. (2010). Tümevarımlı Mantık Programlama ile Türkçe için Kelime Anlamı Belirginleştirme Uygulaması. *Akademik Bilişim*. Muğla.
- CAN, F., KOÇBERBER, S., BALÇIK, E., KAYNAK, C., ÖÇALAN, H., & VURSAVAŞ, O. (2008). Information Retrieval on Turkish Texts. *Journal of the American Society for Information Science and Technology*, 407-421.
- CEBİROĞLU, G., TANTUĞ, A., ADALI, E., & ERENLER, Y. (2003). Sentetik Türkçe Sözcük Kökleri Üretimi. Çanakkale: International XII. Turkish Symposium on Artificial Intelligence and Neural Networks - TAINN.
- ÇİÇEKLİ, İ. (2010). *Otomatik Özetteleme ve Anahtar Kelime Bulma*. Ankara: TÜBİTAK.
- DAŞ, R. (2008). *Web Kullanıcı Erişim Küttüklerinden Bilgi Çıkarımı*. Elazığ: Fırat Üniversitesi FBE, Doktora Tezi.
- DELEN, D., & CROSSLAND, M. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 1707-1720.
- DELİBAŞ, A. (2008). *Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi*. İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- ERYİĞİT, G. (2006). *Türkçe'nin Bağlılık Ayırıştırması*. İstanbul: İstanbul Teknik Üniversitesi FBE, Doktora Tezi.
- G. ÖĞÜDÜCÜ, Ş. (2011). *İTÜ Veri Madenciliği Ders Notları*. İstanbul.
- GÜVEN, A. (2007). *Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği*. İstanbul: Yıldız Teknik Üniversitesi FBE, Doktora Tezi.
- KAISER, K., & MIKSCH, S. (2005). *Information Extraction A Survey*. Vienna, Avusturya: Vienna University of Technology Institut of Software Technology & Interactive Systems.
- KARADENİZ, İ. (2007). *Türkçe İçin Biçimbirimsel Belirsizlik Giderici*. İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- KESGIN, F. (2007). *Türkçe Metinler için Konu Belirleme Sistemi*. İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- KUSHMERICK, N. (1997). *Wrapper Introduction for Information Extraction*. University of Washington, Ph.D.
- MECCA, G., RAUNICH, S., & PAPPALARDO, A. (2007). A new algorithm for clustering search results search results. *Data & Knowledge Engineering*, 504-522.

- MINER, G., DELEN, D., ELDER, J., FAST, A., HILL, T., & NISBET, R. (2012). *Practical Text Mining and Statistical analysis for Non-Structured Text Data Applications*. Waltham, USA: Elsevier.
- MOHAMMAD, M. (2007). *Text Mining: A Burgeoning Quality Improvement Tool*. Ankara: Msc. Thesis, METU.
- NABİYEV, V. (2010). *Yapay Zeka: İnsan-Bilgisayar Etkileşimi*. Ankara: Seçkin Yayıncılık.
- OFLAZER, K. (2002). *Türkçe İçin Bir Sonlu Durumlu "Hafif" Doğal Dil Çözümleyicisi ve Bilgi Çıkarımı Uygulamasının Gerçekleştirilmesi*. TÜBİTAK PROJESİ, PROJE NO:199E027.
- OĞUZ, B. (2009). *Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi*. Antalya: Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi.
- ORHAN, Z. (2006). *Türkçe Metinlerdeki anlam Belirsizliği Olan Sözcüklerin Bilgisayar Algoritmaları ile Anlam Belirginleştirilmesi*. İstanbul: İstanbul Üniversitesi FBE, Doktora Tezi.
- ÖZBİLİMİCİ, A. (2006). *Türkçe Doğal Dili Anlamada İlişkisel Ayrık Bilgiler Modeli ve Uygulaması*. Sakarya: Sakarya Üniversitesi FBE, Yüksek Lisans Tezi.
- ÖZKAN, Y. (2013). *Veri Madenciliği Yöntemleri*. İstanbul: Papatya Yayıncılık.
- PİLAVCILAR, İ. (2007). *Metin Madenciliği İle Metin Sınıflandırma*. İstanbul: Yıldız Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- SARAÇOĞLU, R., TÜTÜNCÜ, K., & ALLAHVERDİ, N. (2008). A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications*, 2545-2554.
- SAY, B. (2003). *Türkçe İçin Biçimbirimsel ve Söz dizimsel Olarak İşaretlenmiş Ağacı Yapılı Bir Derlem Oluşturma*. TÜBİTAK EEEAG Projesi.
- SEZER, E. (2006). *Web Sayfaları İçin Anlamsal Erişim Sistemi*. Ankara: Hacettepe Üniversitesi FBE, Doktora Tezi.
- SODERLAND, S. (1999). Learning Information Extraction Rules for Semi-structured and Free Text. *Machine Learning*, 233-272.
- TURMO, J., AGENO, A., & CATALA, N. (2006). Adaptive Information Extraction. *ACM Computing Surveys*.
- TÜLEK, M. (2007). *Türkçe İçin Metin Öztleme*. İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- TÜRKEŞ, M. (2007). *Bilgi Erişiminde Tamlama Temelli Dizinleme*. İstanbul: İstanbul Teknik Üniversitesi FBE, Yüksek Lisans Tezi.
- WITTEN, I. (2003). Text Mining. Computer Science, University of Waikato.
- ZOHAR, E. (2002). Introduction to Text Mining. *Supercomputing*. Automated Learning Group National Center for Supercomputing Applications, University of Illinois.

VERİ MADENCİLİĞİ

(Web Madenciliği)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

İçerik

- Internet
- World Wide Web
- Web'in Oluşumu
- Web Tarayıcılar
- Web Arama Motorları
- Web Madenciliği
 - Web yapı madenciliği (**Web structure mining**)
 - Web içerik madenciliği (**Web content mining**)
 - Web kullanım madenciliği (**Web usage mining**)

İnternet

- Günümüzde World Wide Web (Kısaca Web) hayatımızın her alanında giderek yaygın bir şekilde kullanılmaktadır.
- **Web, en büyük ve yaygın kullanılan bilgi kaynağı olup arama ve bilgiye erişim hızlı ve kolay bir şekilde yapılabilmektedir.**
- Web üzerinde milyarlarca doküman (Web sayfası) bulunmakta ve milyonlarca kişi sürekli yeni dokümanlar eklemektedir.
- Web, veriye erişimi ve hızlı aramayı sağlamakla birlikte diğer kişilerle bilgi paylaşımını da sağlamaktadır.
- İnternet **diğer kişilerle sesli ve görüntülü görüşme için de kullanılmaktadır.** Bu açıdan **İnternet'in sanal bir topluluk olduğu söylenebilir.**

İnternet

- İnternet günümüzde alışveriş şeklini de değiştirmiştir.
- Mağazaya giderek alışveriş yapmak yerine bilgisayar başında ürünleri almakta ve ödemelerini yapmaktadır.
- Bankacılık, rezervasyon, ödeme başta olmak üzere tüm işlemler elektronik olarak yapılabilmektedir.
- Bu hem maliyet hem de konfor yönünden daha çok tercih edilmektedir.
- **İnternet yaşam kalitesini ve iş yapış şeklinizi de değiştirmiştir.**

World Wide Web

- **Web, kullanıcıların bir bilgisayardan diğer bilgisayarda bulunan veriye ulaşmasını sağlayan Internet tabanlı bilgisayar ağıdır.**
- Web standart istemci-sunucu (client-server) modelini kullanmaktadır.
- Bu modelde kullanıcılar kendi bilgisayarlarındaki program ile uzaktaki bilgisayar bağlanırlar.
- Web üzerinde gezinti için tarayıcı (browser) denilen programlar kullanılır.
- Browser'lar uzaktaki bilgisayardan istekte bulunurlar ve HTML (HyperText Markup Language) biçiminde gelen bilgiyi yorumlayarak istemci taraftaki kullanıcının ekranında görüntüülerler.
- Web üzerinde gezinti yapılırken dokümanlar arasındaki bağlantılar (hyperlink) kullanılır.
- Bu şekilde oluşturulan dokümanlar hypertext olarak adlandırılırlar.

Web'in Oluşumu

- **Web 1989 yılında Tim Berners-Lee tarafından bulunmuştur.** World Wide Web terimini ilk kullanan ve ilk istemci programını yazan kendisidir.
 - **Tim Berners-Lee "Information Management: A Proposal" adlı bir öneriyi çalışmakta olduğu CERN laboratuарında 1989 yılında sunmuştur.**
 - Bu önerisinde hiyerarşik doküman yapısının avantajlarını ve dezavantajlarını ortaya koymuştur.
 - Önerilen doküman yapısıyla bağlantılar (hypertext) aracılığıyla dokümanlar arasında geçiş yapılmaktadır.
 - **Bu öneri dağıtık hypertext sistem olarak adlandırılmıştır ve günümüz Web mimarisinin temelini oluşturmaktadır.**
-

Web'in Oluşumu

- Başlangıçta destek bulamamış olsa da 1990 yılında Tim-Berners Lee tarafından tekrar önerilmiştir.
- Aynı yıl desteklenen proje ile günümüz Web mimarisi geliştirilmeye başlanmıştır.
- İstemci ve sunucu arasında geliştirilen protokol ile iletişim sağlanmıştır.
- Bu çalışmayla **HyperText Trasfer Protocol (HTTP)**, **HyperText Markup Language (HTML)** ve **Universal Resource Locator (URL)** tanımlanmıştır.

Web Tarayıcıları

Mosaic ve Netscape Browser'lar

- **Web'in önemli gelişmelerinden birisi de 1993 yılında mosaic tarayıcının geliştirilmesidir.**
- Mosaic grafik arayüze sahiptir ve Unix işletim sistemi için geliştirilmiştir. Kısa süre sonra mosaic tarayıcının Windows ve Macintosh versiyonları geliştirilmiştir.
- **1994 yılının ortalarında Netscape tarayıcı geliştirilmiştir.**
- Microsoft tarafından geliştirilen **Internet Explorer tarayıcı 1995 yılında geliştirilmiştir.**
- **Web'in popüler ve başarılı olmasında en önemli aşamalardan birisi Mosaic tarayıcının geliştirilmesidir.**

Web Arama Motorları

Internet

- **Internet, Web'in iletişim ağını sağlar.**
- Internet'e ilişkin çalışmalar ARPA (Advanced Research Projects Agency) tarafından desteklenmiştir.
- **İlk ARPANET bağlantısı 4 node ile 1969 yılında yapılmıştır.** 1972 yılında ise 40 node ile bağlantı yapılmıştır.
- **1973 yılında Vinton Cerf ve Bob Kahn** tarafından **TCP/IP (Transmission Control Protocol / Internet Protocol)** protokolünün ilk versiyonu geliştirilmiştir.
- Geliştirilen TCP/IP protokol yığını ile birbirinden uzakta farklı ağlar içinde yer alan bilgisayarlar birbirine bağlanmıştır.
- **1982 yılında TCP/IP protokolünü kullanan Internet doğmuştur.**

Web Arama Motorları

Search Engines

- Bilginin Dünya üzerinde dağıtık ve çok büyük boyutlarda bulunmasından dolayı bilgiyi bulmak ve erişmek daha önemli hale gelmeye başladı.
- **Çok büyük bir alanda ve dağıtık bulunan bilginin bulunması için arama motorları geliştirilmeye başlanmıştır.**
- **Excite arama motoru 1993 yılında** 6 Stanford Üniversitesi öğrencisi tarafından **geliştirilmiştir.**
- 1994 yılında EINET Galaxy geliştirilmiştir ve **1994 yılında Yahoo! geliştirilmiştir.**
- Yahoo! diğer alternatiflerine göre favoriler listesi ve öneriler dizini sunmaktadır.
- Ardından Lycos, Infoseek, Alta Vista, Inktomi, Ask Jeeves, Northernlight gibi arama motorları geliştirilmiştir.

Web Madenciliği

- **Son on yılda Web'in gelişimi sonucunda Dünya'nın en büyük veri kaynağı ortaya çıkmıştır.**
 - **Web kendine özgü çok sayıda karakteristik özelliğe sahiptir ve çok büyük veri üzerinde veri madenciliği önemli ve zor bir iş haline gelmiştir.**
 - Web üzerindeki veri miktarı çok büyütür ve gün geçtikçe hızla artmaktadır. Aranan her türlü bilgi Web üzerinde bulunabilmektedir.
 - **Web üzerinde yapılandırılmış tablolar, yapılandırılmış Web sayfaları, düz metinler ve multimedia dosyaları gibi çok farklı dosyalar bulunmaktadır.**
 - **Web üzerindeki veri heterojendir.**
-

Web Madenciliği

- **Aynı bilgiye sahip Web sayfaları çok farklı biçimlerde ve içeriğe sahip** şekilde Web üzerinde bulunabilmektedir.
- **Bu farklılık Web sayfalarındaki bilgilerin entegrasyonunu çok zor hale getirmektedir.**
- **Web üzerindeki bilginin çok önemli bir kısmı bağlantılar sahiptir.**
- **Hyperlink'ler aynı site üzerindeki Web sayfaları arasında veya çok farklı sitelerdeki Web sayfaları arasında olabilmektedir.**
- **Hyperlink'ler Web sayfaları için çok önemlidir.**
- **Çok sayıda** Web sayfası tarafından **link verilen sayfalar otorite sayfalar**

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi gürültüye sahiptir. Gürültü iki farklı kaynaktan dolayı oluşmaktadır.**
- **Bunlardan birincisi, Web sayfası gezinti linkleri, reklamlar, copyright bilgileri, privacy bilgileri, v.b. gibi çok farklı türde veriye sahiptir.**
- **İyi bir Webbilgisi analizi için gürültüleri ortadan kaldırmak gereklidir.**
- **İkincisi, Web üzerindeki bilginin kalite kontrolü bulunmamaktadır** ve herhangi birisi istediği bilgiyi bir link üzerindeki Web sayfasına yazabilir.
- **Web üzerindeki verinin büyük bir kısmı düşük kalitede, hatalı ve eksiktir.**
- Web üzerinde ticari uygulamalar bulunmaktadır ve insanlar çok sayıda

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi dinamiktir ve sürekli değişmektedir.**
- **Değişiklikleri anlık izlemek bazı uygulamalar için çok önemlidir.**
- **Web sanal bir topluluktur.** Web sadece insanlar arasında veri iletişimini değil **insanlar arasındaki etkileşimi de sağlamaktadır.**
- Yukarıdaki özelliklerin hepsi Web üzerindeki bilginin elde edilmesi için kullanılacak yöntemler için hem fırsatları hem de zorlukları beraberinde getirmektedir.
- **Web madenciliği, veri madenciliğinde kullanılan tüm tekniklerin uygulanmasını içermez.**
- Çok zengin ve farklı özelliklere sahip veriyi bulundurmasından dolayı **Web madenciliği kendine özgү algoritmala sahiptir.**

Web Madenciliği

- **Web madenciliği kullanılabilir bilgiyi Web bağlantılarından, sayfa içeriklerinden ve kullanılan veriden** elde eder.
- Web madenciliği çok sayıda veri madenciliği tekniğini kullanır ancak **sahip olduğu verinin heterojen olması, yarı yapılandırılmış veya yapılandırılmamış** olmasından dolayı **sadece veri madenciliği uygulaması olarak görmek doğru değildir.**
- Çok sayıda veri madenciliği yöntemi son on yılda geliştirilmiştir.
- Web mining üç türde ele alınmaktadır. Bunlar;
 - **Web yapı madenciliği**
 - **Web içerik madenciliği**
 - **Web kullanım madenciliği**yöntemleridir.

Web yapı madenciliği

- Web yapısı madenciliği **faydalı ve kullanılır bilgiyi** Web sayfalarında bulunan **bağlantılardan çıkarır.**
- **Bağlantılar kullanılarak hangi sayfanın daha önemli olduğu gibi bilgiler elde edilebilir.**
- **Ayrıca aynı ortak ilgilere sahip olan benzer kullanıcıları belirleyebiliriz.**
- **Klasik veri madenciliğinde bu tür bilgiler bulunmaz.**

Web içerik madenciliği

- **Web içerik madenciliğinde faydalı ve kullanılır bilgiler Web sayfalarının içeriğinden elde edilir.**
- Örneğin **Web sayfaları içeriklerine göre sınıflandırılabilir.**
- **Bu özellikler klasik veri madenciliğinde de kullanılmaktadır.**
- Web sayfalarında **kullanıcıların forum bilgilerine müşteri görüşlerine dayanarak çıkarımlar yapılmaktadır.**

Web kullanım madenciliği

- **Web kullanım madenciliği**, kullanıcıların **Web sayfalarına erişim bilgilerini kullanır.**
- **Kullanıcıların tıklama bilgileri, sayfalarda gezinme bilgileri, sayfalar üzerindeki etkileşim bilgileri** gibi veriler kullanılır.
- Yukarıdaki işlerin yanı sıra Web üzerindeki verilerin zengin ve çok çeşitli oluşu Web madenciliğinde çok farklı uygulama alanları oluşturmaktadır.
- **Web madenciliği süreci ile veri madenciliği süreci birbirine benzemektedir.** Sadece **veri toplama aşaması farklıdır.**
- **Klasik veri madenciliğinde veriler bir veri ambarında tutulur.**
- **Web madenciliğinde ise veriler dağıtık bulunan Web üzerinde bulunur ve toplanması çok önemli ve zor bir iştir.**
- Veriler elde edildikten sonra **ön işleme, Web madenciliği ve post-processing** işlemleri gerçekleştirilir.

Web madenciliğinin kullanım alanları

- Web madenciliğinin günümüzde birçok alanda kullanılmasının en önemli sebebi, kişilerin web sayfalarında göstermiş oldukları davranışlarını, hareketlerin ve yapmış oldukları işlem bilgilerinin var olan iş süreçlerine entegrasyonunu sağlayarak müşterinin en iyi şekilde anlaşılmasını sağlayan müşteri odaklı bir sistem oluşturmasıdır.
- Web madenciliği kullanım alanları aşağıdaki gibidir.
- Web üzerinden ürün satışı gerçekleştiren şirketler web verilerini analiz ederek müşteri profili ve kümeleri oluşturmaktadır.
- Google vd. arama motorları web içerik madenciliği uygulayarak aranan anahtar kelimeyi içeren web sitelerini belirlemektedirler.
- Web madenciliği uygulanarak web sitelerinin iyileştirilmesi ve güncel kalması sağlanmaktadır.

VERİ MADENCİLİĞİ VE TÜRKİYE'DEKİ UYGULAMA ÖRNEKLERİ

Serkan SAVAŞ¹, Nurettin TOPALOĞLU², Mithat YILMAZ³

Geliş: 11.01.2012 Kabul: 22.03.2012 (Tarama Makalesi)

ÖZET

Günümüz teknolojisi hızla ilerlemekte ve her geçen gün gücü de artmaktadır. Bilgisayarların bilgi saklama kapasitelerinin artmasıyla birlikte bilgi kaydı yapılan alanların sayısı da artmaktadır. Bundan dolayı eldeki verilerin analizi ve sonucu bu verilerden kestirme yöntemlerinin önemi karar vericiler için gittikçe artmaktadır. Bilgisayar sistemleri ile üretilen veriler tek başlarına degersizdir, çünkü çıplak gözle bakıldığından bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade etmeye başlar. Bu yüzden büyük mikardaki verileri işleyebilen teknikleri kullanabilmek büyük önem kazanmaktadır. Bu ham veriyi bilgiye veya anamlı hale dönüştürme işlemleri veri madenciliği ile yapılabilmektedir. Bu çalışmada veri madenciliğinin günümüz disiplinleri arasında geldiği noktaya deinceilmiş ve Türkiye'de veri madenciliği üzerine yapılan çalışmalar ve gerçekleştirilen uygulamalar incelenmiştir.

Anahtar Kelimeler: Veri madenciliği, Türkiye'deki uygulamaları.

DATA MINING AND APPLICATION EXAMPLES IN TURKEY

ABSTRACT

Today's technology is advancing rapidly and its power is increasing everyday. The number of the fields which are storing information is increasing with the increasing of computers' information storage capacity. Therefore the importance of analyzing data and prediction results from these data is increasing for decision-makers. Data which are produced by computers are worthless alone because they are meaningless when you look with eyes. These data become meaningful when they are processed for an aim. Because of this, using the techniques which can process large amount of data is becoming important. Changing this raw data to information and to significant state is possible with data mining. In this study data mining's position between today's disciplines is mentioned and data mining application examples in Turkey are examined.

Keywords: Data mining, applications in Turkey.

¹ Teknik Öğretmen, Kızılcahamam Teknik ve Endüstri Meslek Lisesi.

² Gazi Üniversitesi Teknoloji Fakültesi.

³ Gazi Üniversitesi Teknik Eğitim Fakültesi.

1.GİRİŞ

Bilgisayar sistemleri ile üretilen veriler tek başına degersizdir, çünkü çiplak gözle bakıldığından bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade etmeye başlar (Kalikov, 2006). Bilgi bir amaca yönelik işlenmiş veridir. “Ham veri” veya yalnız geçmişte ne olduğunun bir görüntülemesi olan “enformasyon”a dayalı karar almak mümkün değildir. Geçmişte yaşanan kötü bir tecrübe kaynaklanan kaybin engellenmesi de mümkün değildir. Önemli olan geçmişe ait olaylara dair gizli bilgilerin keşfedilmesi, ileriye yönelik durumsal öngörüler veren modeller ile önceden tedbir almamızı sağlayacak bir yönetim anlayışına geçmek ve olası kayıpları öngörebilmektir (İnan, 2003). Bu yüzden büyük miktardaki verileri işleyebilen teknikleri kullanabilmek büyük önem kazanmaktadır. Bu ham veriyi bilgiye veya anamlı hale dönüştürme işlemleri veri madenciliği ile yapılmaktadır (Kalikov, 2006). Veri madenciliği, bu gibi durumlarda kullanılan büyük miktardaki veri setlerinde saklı durumda bulunan örüntü ve eğilimleri keşfetme işlemidir (Thuarisingham, 2003).

Günümüzde veri madenciliği işletmeler için çok önemli hale gelmiştir. Çok büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içinde değerli verileri bulunduran bir veri madeni gibi düşünülebilir. Bu büyülükteki verilerin analizi, bu analiz sonucunda daha anamlı bilgi elde etme ve elde edilen bilgiyi yorumlama işi, insan yeteneği ve ilişkisel veri tabanlarının yapabileceklerini aşmaktadır. Bilhassa dijital veri miktarında artış patlaması ve buna karşılık, bu veriler üzerinde araştırma ve uygulama yapan kişilerin sayısının değişmemesi, çalışmaları veri madenciliğine doğru zorlamıştır. Bu ihtiyaçların sonucunda otomatik ve akıllı veri tabanı analizi için yeni kuşak teknikler doğmuştur. Bu teknikler öyle olmalıdır ki, veriyi akıllı ve otomatikleşmiş şekilde işe yarar bilgiye dönüştürebilsin. Tüm bunların sonucunda veri madenciliği cevap olarak sunulmuş ve giderek önemini artıran bir araştırma alanı haline gelmiştir. Bu çalışmada veri madenciliğinin günümüz disiplinleri arasında geldiği nokta, Türkiye'de veri madenciliği üzerine yapılan çalışmalar ve gerçekleştirilen uygulamalar incelenmiştir.

2. VERİ MADENCİLİĞİ (DATA MINING)

Veri madenciliği, büyük miktarlardaki verinin içinden geleceği tahmin edilmesinde yardımcı olacak anamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. Ayrıca veri madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan ve veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesini sağlayan veri analizi tekniğidir (Kalikov, 2006). Bu işlemlerin uygulama alanı oldukça genişir. Bu alanlar içerisinde Şekil 1.1'de gösterildiği gibi, veri tabanı sistemleri, Veri Görselliği, Yapay Sinir Ağları, İstatistik, Yapay Öğrenme, vb. gibi disiplinler bulunmaktadır.



Şekil 1.1. Veri madenciliği ve disiplinler

Veri madenciliği araçları kullanılarak, işletmelerin daha etkin kararlar almasına yönelik karar destek sistemlerinde gerekli olan eğilimlerin ve davranış kalıplarının ortaya çıkarılması mümkün olmaktadır. Geçmişteki klasik karar destek sistemlerinin kullanıldığı araçlardan farklı olarak, veri madenciliğinde çok daha kapsamlı ve otomatize edilmiş analizler yapmaya yönelik, birçok farklı özellik bulunmaktadır (İnan, 2003).

Veri madenciliğinin işletmelere sunduğu en önemli özellik, veri grupları arasındaki benzer eğilimlerin ve davranış kalıplarının belirlenmesidir. Bu süreç aynı zamanda otomatize edilmiş bir biçimde hayatı geçirilebilmektedir. Bu fonksiyon özellikle hedef pazarlara yönelik pazarlama faaliyetlerinde yoğun olarak kullanılmaktadır (İnan, 2003). Başka bir özelliği ise daha önceden bilinmeyen, veri ambarları içerisinde bulunan ancak ilk etapta görülemeyen bilgilerin ortaya çıkarılabilmesidir. Örneğin bir firma sattığı ürünleri analiz ederek, ilerideki kampanyalarını şekillendirebilir ya da sattığı ürünler arasındaki bağlantıları keşfetebilir. Burada amaç daha önceden fark edilmeyen veri kümelerinin bulunabilmesidir.

Günümüzün ekonomik koşulları ve yaşanan hızlı değişim ortamlarında, iş deneyimi ve önsezilere dayanarak alınan kararlarda yanlış karar alma riski çok yüksektir. Riski azaltmanın tek yolu bilgiye dayalı yönetimi öngören karar destek çözümleridir. Veri madenciliği teknikleri gerçek anlamda bir karar destek sistemi oluşturmada olmazsa olmaz araçlardır. Bu noktada bilgi teknolojilerinden yararlanmak kaçınılmaz olmuştur.

2.1. Veri Madenciliğinin Tanımı

Bu güne kadar farklı kaynaklarda veri madenciliğinin pek çok tanımıyla karşılaşılmıştır. Bu kaynaklardan bazlarına göre veri madenciliğinin tanımı şöyledir:

- Jacobs (1999), veri madenciliğini, ham datanın tek başına sunamadığı bilgiyi çıkaran, veri analizi süreci olarak tanımlamıştır (Jacobs, 1999).
- Veri madenciliği, büyük veri yiğinları arasından gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağlantıların, bilgisayar programı kullanarak aranması işidir (Doğan ve Türkoğlu, 2007).
- Hand (1998), veri madenciliğini istatistik, veritabanı teknolojisi, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veritabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizi olarak tanımlamıştır (Hand, 1998).
- Kitler ve Wang (1998), veri madenciliğini oldukça tahminci anahtar değişkenlerin binlerce potansiyel değişkenden izole edilmesini sağlama yeteneği olarak tanımlamışlardır (Kitler ve Wang, 1998).

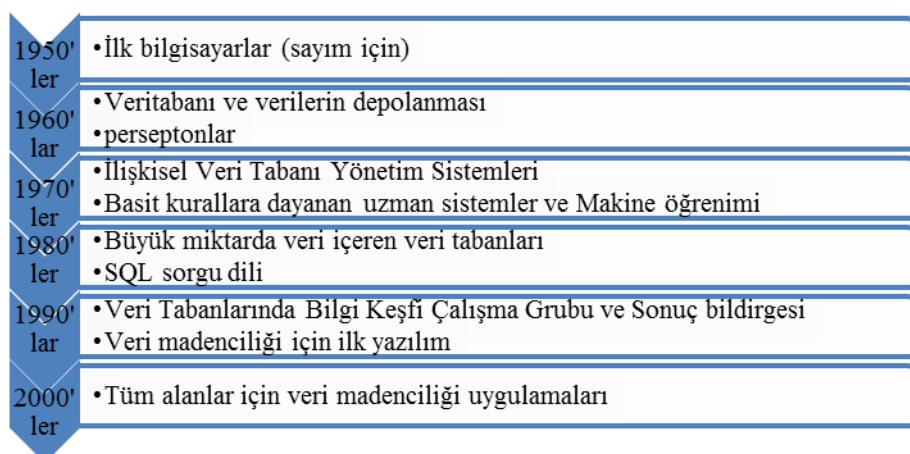
Bu tanımlardan yola çıkarak söyle bir tanım yapmak mümkündür: Veri madenciliği, çok büyük miktarda bilginin depolandığı veri tabanlarından, amacımız doğrultusunda, gelecek ile ilgili tahminler yapmamızı sağlayacak, anlamlı olan veriye ulaşma ve veriyi kullanma işidir.

2.2. Veri Madenciliğinin Tarihi

Günümüzde neredeyse her eve bilgisayar girmiştir ve internet erişimi hemen hemen her yerden sağlanmaktadır. Disk kapasitelerinin artması, her yerden bilgiye ulaşma olasılığı, bilgisayarların çok büyük miktarlarda veri saklamasına ve daha kısa sürede işlem yapmasına olanak sağlamıştır. Geçmişten günümüze veriler her zaman yorumlanmış, bilgi elde etmek istenmiştir ve bunun için de donanımlar oluşturulmuştur. Bu sayede bilgi, geçmişten günümüze taşınır hale gelmiştir.

1950'li yıllarda ilk bilgisayarlar sayımlar için kullanılmaya başlamıştır. 1960'larda ise veri tabanı ve verilerin depolanması kavramı teknoloji dünyasında yerini almıştır. 1960'ların sonunda bilim adamları basit öğrenmeli bilgisayarlar geliştirebilmişlerdir. Minsky ve Papert, günümüzde sınır ağları olarak bilinen perceptron'ların sadece çok basit olan kuralları öğrenebileceğini göstermişlerdir (Adriaans ve Zantinge, 1997). 1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmıştır. Bilgisayar uzmanları bununla beraber basit kurallara dayanan uzman sistemler geliştirmiştir ve basit anlamda makine öğrenimini sağlamışlardır. 1980'lerde veri tabanı yönetim sistemleri yaygınlaşmış ve bilimsel alanlarda, mühendisliklerde vb. alanlarda uygulanmaya başlanmıştır. Bu yıllarda şirketler, müşterileri, rakipleri ve ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanlarının içerisinde çok büyük miktarlarda

veri bulunmaktadır ve bunlara SQL veri tabanı sorgulama dili ya da benzeri diller kullanarak ulaşılabilir. 1990'larda artık içindeki veri miktarı katlanarak artan veri tabanlarından, faydalı bilgilerin nasıl bulunabileceği düşünülmeye başlanmıştır. Bunun üzerine çalışmalar ve yaynlara başlanmıştır. 1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilenek "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin KDD (Knowledge Discovery and Data Mining) ile ilgili temel tanım ve kavramları ortaya koyması ile süreç daha da hızlanmış ve nihayet 1992 yılında veri madenciliği için ilk yazılım gerçekleştirilmiştir. 2000'li yıllarda veri madenciliği sürekli gelişmiş ve hemen hemen tüm alanlara uygulanmaya başlanmıştır. Alınan sonuçların faydalari görüldükçe, bu alana ilgi artmıştır. Veri madenciliğinin tarihsel gelişim süreci, Şekil 1.2'de gösterilmiştir.



Şekil 1.2. Veri madenciliğinin tarihsel süreci

2.3. Veri Madenciliğinin Kullanıldığı Alanlar

Büyük hacimde veri bulunan her yerde veri madenciliği kullanmak mümkündür. Günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır. Örneğin pazarlama, biyoloji, bankacılık, sigortacılık, borsa, perakendecilik, telekomünikasyon, genetik, sağlık, bilim ve mühendislik, kriminoloji, sağlık, endüstri, istihbarat vb. birçok dalda başarılı uygulamaları görülmektedir (İnan, 2003; Albayrak, 2008; Akgöbek ve Çakır, 2009).

Son 20 yıldır Amerika Birleşik Devletleri’nde çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir. Kaynaklar incelendiğinde veri madenciliğinin en çok kullanıldığı alan olarak tıp, biyoloji ve genetik görülmektedir.

2.4. Veri Madenciliğini Etkileyen Etmenler

Veri madenciliği temel olarak 5 ana faktörden etkilenir (Akpınar, 2000):

1. **Veri:** Veri madenciliğinin bu kadar gelişmesindeki en önemli faktördür.
2. **Donanım:** Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.
3. **Bilgisayar ağları:** Yeni nesil internet, çok yüksek hızları kullanmayı sağlamaktadır. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır.
4. **Bilimsel hesaplamalar:** Günümüz bilim adamları ve mühendisleri, simülasyonu, bilimin üçüncü yolu olarak görmekteler. Veri madenciliği ve bilgi keşfi, teori, deney ve simülasyonu birbirine bağlamada önemli bir rol almaktadır.
5. **Ticari eğilimler:** Günümüzde, işletmeler rekabet ortamında varlıklarını koruyabilmek için daha hızlı hareket etmeli, daha yüksek kalitede hizmet sunmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır.

2.5. Veri Madenciliğinde Karşılaşılan Problemler

Büyük hacimli verilerin bulunduğu veri ortamlarında büyük sorunlar ortaya çıkabilir. Bu nedenle küçük veri kümelerinde, benzetim ortamlarında hazırlanmış veri madenciliği sistemleri, büyük hacimli, eksik, gürültülü, boş, atık, aykırı veya belirsiz veri kümelerinin bulunduğu ortamlarda yanlış çalışabilir. Bu nedenle veri madenciliği sistemleri hazırlanırken bu sorunların çözülmesi gerekmektedir.

Veri madenciliği uygulamalarında karşılaşabilecek sorunlar şunlardır:

Artık veri: Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz niteliklerdir. Bu durum pek çok işlem sırasında karşımıza çıkabilir.

Belirsizlik: Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir.

Boş veri: Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dâhil olmak üzere hiçbir değere eşit olmayan değerdir.

Dinamik veri: Kurumsal çevrim içi veri tabanları dinamiktir ve içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metotları için önemli sakincalar doğurmaktadır.

Eksik veri: Veri kümесinin büyüklüğünden ya da doğasından kaynaklanmaktadır. Eksik veriler olduğunda yapılması gerekenler şunlardır:

- Eksik veri içeren kayıt veya kayıtlar çıkarılabilir.
- Değişkenin ortalaması eksik verilerin yerine kullanılabilir.
- Var olan verilere dayalı olarak en uygun değer kullanılabilir.

Eksik veriler, yapılacak olan istatistiksel analizlerde önemli problemler yaratmaktadır. Çünkü istatistiksel analizler ve bu analizlerin yapılmasına olanak veren ilgili paket programlar, verilerin tümünün var olduğu durumlar için geliştirilmiştir (Albayrak, 2008).

Farklı tipteki verileri ele alma: Gerçek hayatı uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir.

Gürültülü ve kayıp değerler: Veri girişi veya veri toplanması esnasında oluşan sistem dışı hatalara gürültü denir. Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Veri toplanması esnasında oluşan hatalara ölçümden kaynaklanan hatalar da dahil olmaktadır. Bu hataların sonucu olarak birçok çok niteliğin değeri yanlış olabilir ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir.

Sınırlı bilgi: Veri tabanları genel olarak basit öğrenme işlerini sağlayan özellik veya nitelikleri sunmak gibi veri madenciliği dışındaki amaçlar için hazırlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir.

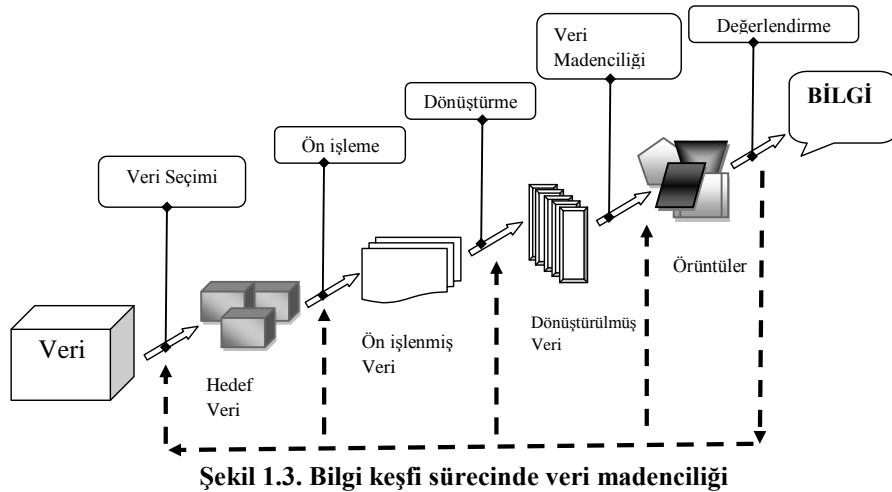
Veri tabanı boyutu: Veri tabanı boyutları büyük bir hızla artmaktadır. Veri tabanı algoritması çok sayıda küçük örneklemi ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklemelerde kullanılabilmesi için çok dikkat gerekmektedir.

2.6. Veri Madenciliği Süreci

Veri madenciliği, aynı zamanda bir süreçtir. Veri yiğinları arasında, soyut kazılar yaparak veriyi ortaya çıkarmayanın yanı sıra, bilgi keşfi sürecinde örüntülerin ayırtırarak süzmek ve bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır. Bu süreç Şekil 1.3'de gösterilmiştir. Üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının fayda sağlama mümkün değildir. Bu sebeple, veri madenciliği sürecine girilmeden önce, başarının ilk şartı, iş ve veri özelliklerinin detaylı analiz edilmesidir.

Veri madenciliği sürecinde izlenen adımlar genellikle aşağıdaki şekildedir (Shearer, 2000):

1. Problemin tanımlanması,
2. Verilerin hazırlanması,
3. Modelin kurulması ve değerlendirilmesi,
4. Modelin kullanılması,
5. Modelin izlenmesi.



Problemin tanımlanması: Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, projenin hangi işletme amaci için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanmasıdır.

Verilerin hazırlanması: Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanın %50 - %85'ini harcamasına neden olmaktadır (Piramuthu, 1998). Verilerin hazırlanması, "toplama", "değer biçme", "birleştirme ve temizleme", "örneklem seçimi" ve "dönüştürme" aşamalarından oluşmaktadır.

Modelin kurulması ve değerlendirilmesi: Tanimlanan problem için en uygun modelin bulunabilmesi, olabildigince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılincaya kadar yinelelen bir süreçtir.

Modelin kullanılması: Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir.

Modelin izlenmesi: Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla üretilikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve yeniden düzenlenmesini gerektirecektir.

2.7. Veri Madenciliği Metotları

Veri madenciliği ile ilgili kullanılan pek çok yöntemin yanına hemen her geçen gün yeni yöntem ve algoritmalar eklenmektedir. Bunlardan bir kısmı onlarca yıldır kullanılan klasik teknikler diyeBILECEĞİMİZ, ağırlıklı olarak istatistiksel yöntemlerdir.

Diğer yöntemler de genellikle istatistiği temel alan ama daha çok makine öğrenimi ve yapay zekâ destekli yeni nesil yöntemlerdir.

Veri madenciliği modelleri, gördükleri işlevlere göre temel olarak 3 grupta toplanır. Bunlar:

1. Sınıflama (*Classification*) ve Regresyon (*Regression*),
2. Kümeleme (*Clustering*),
3. Birliktelik Kuralları (*Association Rules*),

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme ve birliktelik kuralları modelleri tanımlayıcı modellerdir (Özkes, 2003).

3. TÜRKİYE'DEKİ VERİ MADENCİLİĞİ ÇALIŞMALARI ve UYGULAMALARI

Pek çok alanda etkili bir şekilde kullanılmaya başlayan veri madenciliği, günümüzün en çok uygulanan disiplinlerinden birisi olmuştur. Her geçen sene kendisine daha da yaygın bir kullanım alanı bulmakla birlikte, kolay uygulanabilirliği ve etkili sonuçlar ortaya çıkarması sayesinde, kurum ve kuruluş yöneticileri tarafından en çok başvurulan yöntemlerden bir tanesidir. Literatür taramasıyla ulaşılan veri madenciliği ile gerçekleştirilmiş uygulamaları, eğitim, ticaret, mühendislik, bankacılık ve borsa, tip ve telekomünikasyon başlıklarını arasında sınıflandırarak şu şekilde özetleyebiliriz.

3.1. Mühendislik Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Kıyas Kayaalp tarafından 2007 yılında yapılan bir yüksek lisans çalışmasında, veri madenciliği teknigi ile üç fazlı asenkron motordaki sargı spırıları arasında oluşabilecek kısa devre veya yalıtım bozuklukları ve motor milinde oluşabilecek mekanik dengesizlik hatalarının tespiti gerçekleştirılmıştır (Kayaalp, 2007).

Ali İnan tarafından 2006 yılında yapılan bir çalışmada şu bulgulara ulaşılmıştır: Kişilerin konum bilgilerinin toplanması, kullanımı ve dağıtılması ile ilgili gizlilik kaygıları zaman-mekân bilgisi içeren verilerde veri madenciliği teknikleri uygulanmasının önündeki tek engeldir. Kimlik belirteçlerinin veriden temizlenmesi kişisel gizliliğin sağlanmasında tek başına yeterli olamıyor çünkü umuma açık ev ve işyeri adresleri kullanılarak kişilerin hareket yörüngeleri ile kimliklerinin eşlenmesi mümkündür. Var olan gizliliği koruyan veri madenciliği teknikleri de yeterli olmuyor çünkü bu tekniklerin zaman-mekân bilgisi içeren verilere uygulanabilmesi için ardisık konum gözlemlerinin kişilerin birbirinden bağımsız nitelikleri olduğunu varsayılmak gerekmektedir. Ancak bu varsayılm hatalı olacaktır. Bu nedenle konum-zaman veri tabanlarında veri madenciliğini mümkün kilmak, bu tip veriler için özel olarak tasarlanmış algoritmalar gerektir. Bu çalışmada zaman-mekân nitelikleri olan veriler için bir gizliliği koruyan veri madenciliği teknigi ve iki ön-isleme teknigi önerilmiştir: (1) Dağıtık kümeleme, (2) Merkezi anonimleştirme ve (3)

Dağıtık anonimleştirme. Önerilen tekniklerin güvenlik ve performans analizleri de yapılmış ve sonuçta mantıklı varsayımlar altında minimum mahrem bilgi kaybıyla veri madenciliğinin mümkün olduğu gözlemlenmiştir (Inan, 2006). Gökhan Yavaş tarafından 2003 yılında gerçekleştirilen başka bir çalışmada ise mobil kullanıcıların hareket modellerinin veri madenciliği kullanılarak çıkarılması ve bu modeller kullanılarak mobil kullanıcıların daha sonraki hareketlerinin tahmin edilmesi için yeni bir algoritma geliştirilmiştir. Üç aşamadan oluşan bu algoritmanın ilk aşamasında kullanıcı hareket modelleri, kullanıcıların önceden kaydedilmiş mobil yörüngelerinden veri madenciliği kullanılarak çıkarılmaktadır. İkinci aşamada bulunan hareket modellerinden hareket kuralları üretilmekte, son aşamada ise bu hareket kuralları kullanımının bir sonraki hücreler arası hareketinin tahmini için kullanılmaktadır. Sunulan algoritmanın performansı simülasyonlar yardımıyla iki farklı tahmin yöntemiyle karşılaştırılmıştır. Performans sonuçları algoritmanın diğer metodlardan daha doğru tahminler yapabildiğini göstermiştir (Yavaş, 2003).

Sibel Kırmızıgül Çalışkan ve İbrahim Soğukpinar 2008 yılında, veri madenciliği yöntemlerinden “K-means” ve “K en yakın komşu” yöntemlerinin iyileştirilmesi amacıyla; nüfuz tespiti için kümelemeyi ve sınıflandırmayı, denetimli ve denetimsiz öğrenimi, k-means ve k en yakın komşu yöntemlerini bir arada kullanan hibrit bir yapı geliştirmiştir. Farklı boyutlardaki veri gruplarında düşük performans gösterebilen, fakat gerçeklemesi kolay ve zaman karmaşası az olan “K-means” ile tek ve geniş bir küme için belirlenen k ve eşik değeri, küme içindeki farklı özelliklere sahip normal davranış ve saldırısı verileri için zorunlu kılan ve zaman karmaşası çok olan, fakat k komşu ortalaması aldığı için gürültülü verilerden az etkilenen “k en yakın komşu” yöntemleri bir arada kullanılmıştır. Geliştirilen uygulamada en hızlı sonucu veren k-means uygulaması ile test kümesi daha küçük alt kümelere ayrılarak k en yakın komşu yönteminin zaman karmaşası ve bellek gereksinimi azaltılmıştır (Çalışkan ve Soğukpinar, 2008).

N. Duru ve M. Canbay 2007 yılında veri madenciliği ile deprem verilerinin analizi üzerine bir çalışma gerçekleştirmiştir. Bu çalışma deprem verileri kullanılarak seçilen bir bölgeye ait sismik tehlikeni diğer deyişle gerçekleşme olasılığının veri madenciliği yönünden ele alınarak incelenmesini kapsamaktadır. Çalışma sonuçları jeofizik sonuçlar ile korele edilerek doğruluk payı da araştırılmıştır. Her gelecek 10 yıl için % sismik tehlike değeri artış göstererek devam etmiş, örneğin 6 magnitüdündeki bir depremin olma olasılığı 10 yıl içinde %27 iken, 30 yıl içinde %60 ve 60 yıl için de %80'leri bulmaktadır. Bu değerler daha önce çalışma bölgesinde yapılmış çalışmalarla uyum göstermektedir. Ancak burada unutulmaması gereken bu çalışmanın deprem tahmini için kullanılan tekniklerden sadece birisi olduğu ve bu çalışmanın konusu itibariyle çalışma bölgelerinin tektonik özelliklerini hiç irdelemeden dahi olsa olumlu sonuçlara varılabilmesinin mümkün olduğunu gösterilebilmesidir. Ayrıca yapılan çalışmanın sonuçlarının büyük bölgelere göre küçük bölgelerde daha iyi sonuç verdığının görülmemesidir. Uygulama, dünya ölçüğündeki her noktanın analizini yapacak şekilde geliştirilmiş olup, ihtiyaç halinde programa eklemeler yapmak suretiyle, başka bu tür çalışmalar yapacak şekilde tasarlanmıştır (Duru ve Canbay, 2007).

Yaşar Doğan tarafından 2004 yılında Deniz Harp Okulu'nda, su altı taktik duyarga ağlarında veri madenciliği tabanlı hedef sınıflandırması çalışması hazırlanmıştır. Bu çalışmada, açık, sıçık ve çok sıçık sularda denizaltı, küçük sualtı taşıma araçları, sualtı mayınları ve dalgaçları sınıflandırmada maliyeti çok az olan mikroduyargalar kullanılmıştır. Algoritma, yüzeydeki şamandıralara bağlı ve ayarlanabilir derinliklere indirilebilen duyargalarдан oluşan taktik su altı duyarga ağları için tasarlanmıştır. Sınıflandırma veri madenciliği teknigi olarak karar ağacı algoritmaları kullanılmıştır (Doğan, 2004).

Eyüp Sıramkaya'nın 2005 yılında hazırladığı bir uygulamada internet üzerinden ulaşılabilen basın-yayın kaynaklarında yer alan görsel ve metinsel verilerin hızlı ve etkin bir şekilde erişimi ve bu kaynaklardan anlamlı ve önemli bilgilerin çıkarılması hedeflenmiştir. Çalışmalar istihbarat açısından önem taşıyan kişi ve örgütlerle ilgili haberler üzerinde yoğunlaşmıştır. Sunucu bilgisayarda internet üzerinde yer alan haber kaynaklarından toplanmış ve işlenmiş metinsel belgelerden oluşan veri-tabanı ile bu bilgileri içleyen uygulama yazılımları bulunmaktadır. Bir arayüz ile kullanıcının bu bilgileri sorgulaması sağlanmıştır. Çalışma, Birliktelik Kural Madenciliği teknigi ile uygulanmıştır. Bu teknik uygulanırken Apriori Algoritması kullanılmıştır. Yapılan veri madenciliği çalışmasında Bulanık Mantık çalışması, kişi-kİŞİ ilişkilerini bulmakta uygulanmıştır. Bu uygulamadaki amaç kullanıcıların arama yapmak istedikleri kişilerin isimlerini yazarken yapabilecekleri yazım hatalarını elemeiktir. İsimlerdeki harflerin konumlarının birbirlerine göre uzaklıklarını temel alarak bulanık mantık kurallarının uygulandığı bir algoritma kullanılmıştır (Sıramkaya, 2005).

Yomi Castro 2006 yılında, bir yazılımın yeni sürümlerindeki hata oranını eski sürümlerine göre olan değişikliklerini temel alarak tahmin eden bir model ortaya koyma amaçlı bir uygulama gerçekleştirmiştir. Bu uygulamada bahsedilen değişiklikler yazılımdaki bir yenilik, bir algoritma değişikliği ve hatta bir hata ayıklama değişikliği olabilir. Bu tür değişikliklerin türünü formel ve nesnel bir bakış açısıyla analiz ederek ve buna yazılımın hacimsel değişikliğini de katarak, yeni sürümündeki hata oranını doğru bir şekilde tahmin edebilme amaçlanmıştır. Bu araştırmada önerilen modeli kullanarak, yazılım hayat döngüsündeki test sürecini kısaltabilmek ve harcanan gücü azaltabilmek mümkün olmuştur. Buna ek olarak, yeni bir yazılım sürümünün sağlamlığını saptamak bu model sayesinde mümkündür. Bu model, aynı zamanda bir yazılım ürününe katılan yeniliklerin, hata ayıklama değişiklikleri gibi değişiklik türlerinin, hata oluşturma ihtimallerine olan katkısını ayrı ayrı anlamaya yardımcı olmaktadır (Castro, 2006).

Seda Dağlar Toprak tarafından yeni bir melez çok ilişkili veri madenciliği teknigi 2005 yılında gerçekleştirilmiştir. Bu çalışmada kavram öğrenme, kavram ile kavramı gerçekleme öenkoşulları arasındaki eşleştirme olarak tanımlanmış ve ilişkisel kural madenciliği alanında bulusal yöntem olarak kullanılan Apriori kuralı örüntü uzayını küçültmek amacıyla kullanılmıştır. Önerilen sistem, kavram örneklerinden ters çözümürlük operatörü kullanılarak genel kavram tanımlarını oluşturan ve bu genel örüntülerin Apriori kuralını temel alan bir operatör yardımı ile

özelleştirek güçlü kavram tanımlamaları elde eden melez bir öğrenme sistemi olarak tanımlanmıştır. Sistemin iki farklı sürümü, üç popüler veri madenciliği problemi için test edilmiş ve sonuçlar önerilen sistemin, en gelişkin ilişkisel veri madenciliği sistemleri ile karşılaştırılabilir durumda olduğunu göstermiştir (Toprak, 2005).

Coşku Erdem, 2006 yılında, matematiksel morfoloji kullanarak yoğunluk temelli kümeleme adında bir uygulama gerçekleştirmiştir. Bu uygulamadaki algoritma veri depolarının imgelere benzerliğinden yola çıkarak bir imge işleme tekniği olan gri tonlu morfolojinin çok boyutlu veri üzerine uygulanması temeline dayanmaktadır. Önerilen bu algoritmanın gerek sentetik gerekse doğal veri üzerindeki başarımı değerlendirilmiş ve uygun parametrelerle çalıştırıldığında başarılı ve yorumlanabilir sonuçlar üretelebildiği görülmüştür. Ek olarak, algoritmanın işlemsel karmaşıklığının düşük boyutlu veri için veri noktası sayısı ile doğrusal, yüksek boyutlu veri içinse temelde morfoloji işlemlerine bağlı olarak boyut sayısı ile üstel olarak artığı hesaplanmıştır (Erdem, 2006).

T. Tugay Bilgin ve A. Yılmaz Çamurcu, veri madenciliğinde güncel araştırma alanlarından biri olan çok boyutlu veri tabanları ve bunların görselleştirilmesinde kullanılan görselleştirme tekniklerini incelemiştir ve bu alanda çalışmalar gerçekleştiren araştırma grupları ve bunların geliştirdikleri yeni yöntemler ve teknikleri irdelemiştir. Ayrıca başka bir çalışmada T. Tugay Bilgin, veri akış diyagramları ve veri akışı tabanlı veri madenciliği süreçleri görselleştirilmesini açıklamıştır. Üç farklı tür veri akış tabanlı yazılımı incelemiştir ve detaylı özelliklerini karşılaştırmıştır (Bilgin ve Çamurcu, 2008; Bilgin, 2009).

2004 yılında Serkan Toprak tarafından, ilişkisel veri tabanları üzerinde çoklu ilişkisel yapıdaki ortak kuralları bulmayı sağlayan bir uygulama geliştirilmiştir. Uygulama altyapısı olarak ilişkisel veri tabanlarındaki desenleri tanımlayabilen, bu desenleri eklerle geliştirebilen ve bu desenlerin çeşitli ölçmeleri için gerekli sayımları veri tabanından temel yetilerle alan bir yapı kullanılmıştır. Bu altyapı, veri tabanının tanımında yer alan bilgileri kullanarak arama alanının daraltılmasını sağlamıştır. Bu çalışma, Apriori algoritmasını arama alanını daha da küçültmek için kullanarak ve altyapı tarafından desteklenmeyen özyinelemeli desenlerin bulunmasını sağlayarak altyapıya yenilikler getirmiştir. Apriori algoritması her tablo üzerinde sık karşılaşılan desenleri bulmak için kullanılmış ve bu algoritmanın gerekli destek değerini bulma yöntemi değiştirilmiştir. Veri tabanındaki özyinelemeli ilişkileri belirlemek için bir yöntem sunulmuş ve uygulama bu durumlar için tablo kısaltmalarının kullanıldığı bir çözüm sağlamıştır. Veri tabanı alanlarında saklanan sürekli değerleri bölümleyebilmek için eşit derinlik yöntemi kullanılmıştır. Uygulama bir veri madenciliği yarışması olan KDD Cup 2001'den alınan örnek genlerde yer tahmini problemi ile test edilmiş ve ortaya çıkan sonuçlar yarışmayı kazanan yaklaşımın sonuçlarıyla karşılaştırılmıştır (Toprak, 2004).

Ulaş Baran Baloğlu tarafından 2006 yılında gerçekleştirilen uygulamada, DNA veri kümelerinde bulunan biyolojik sıralar üzerinde veri madenciliği yapılarak tekrarlı

örüntüler ve potansiyel motifler çıkartılmıştır. Önerilen yöntem yukarıdan-aşağı veri madenciliği ve genetik algoritma tabanlı hibrit bir çözümdür. Bu yöntemdeki yaklaşım iki temel adımda ele alınabilir. Birinci adım, genetik algoritma kullanılarak aday motiflerin bir popülasyonunun oluşturulmasıdır. Bunu diğer nesillerin genetik operatörler ve uygunluk fonksiyonu kullanılarak oluşturulması takip eder. İkinci adımda, veri madenciliği yöntemi yukarıdan-aşağı haliyle kullanılarak aday motiflerin uygunluğunun değerlendirilmesi yapılır. *E. coli* bakterilerinden alınmış DNA sıralarında önerilen yöntem denenerek uygulanabilirliği ve üstün yanları gösterilmiştir (Baloğlu, 2006).

Baş Yıldız 2010 yılında, sık kümelerin bulunması için gizliliği koruyan bir yaklaşım önermiştir. Ayrıca bu çalışmada Matrix Apriori algoritması üzerinde değişiklikler yapılmış ve sık küme gizleme çerçevesi de geliştirilmiştir (Yıldız, 2010).

Yasemin Kılınç 2009 yılında hazırladığı bir çalışmada, birliktelik kuralları için bir yöntem sunmuştur. Apriori algoritmasının ürettiği kurallar elenerek bir elektronik firmasında üretim ve mal giriş kalite verileri üzerinde uygulanmıştır. Ortaya çıkarılan kurallar test verileri ile doğrulanmış ve sonuçlar analiz edilmiştir (Kılınç, 2009).

3.2. Tıp Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Baş Aksoy tarafından 2009 yılında Dekompresyon Analizinin Cluster Analizi üzerine bir veri madenciliği uygulaması gerçekleştirilmiştir. Bu çalışmada, farklı clustering algoritmaları (k-ortalama, COBWEB, EM) ile Divers Alert Network (Dalgıçların Acil Durum Ağı)'nın dalış yaralanmaları bildirim formlarından elde edilen belirti ve bulgu listeleri kullanılarak dekompresyon hastalığı sınıflandırılmış ve sonuçlar klasik sınıflandırma yöntemleri, yeni yapılan istatistiksel sınıflandırma yöntemleri ve tedavi sonuçları ile karşılaştırılmıştır. Ayrıca teşhiste yardımcı olabilecek birliktelik kuralları (association rules) elde edilmiştir. Sonuç olarak, clustering yöntemleriyle elde edilen sınıfların yeni yapılan istatistiksel sınıflandırmalarla ve klasik sınıflandırmalarla uyumlu olduğu ve hafiften şiddetli vakalara giden hiyerarşik yapıda olduğu gözlemlenmiştir (Aksoy, 2009).

Pınar Yıldırım, Mahmut Uludağ ve Abdulkadir Görür tarafından 2008 yılında yapılan çalışmada, hastane bilgi sistemlerindeki veri madenciliği uygulamalarına değinilmiştir (Yıldırım vd., 2008) .

Şengül Doğan ve İbrahim Türkoğlu tarafından 2008 yılında gerçekleştirilen bir çalışmada, kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde, hekime yardımcı olacak ve kolaylık sağlayabilecek bir karar destek sistemi oluşturulmuştur. Öruntu tanıma süreci esas alınmış olup, sistemin işleyışı veri madenciliği tekniklerinden olan karar ağaçları yapısı ile sağlanmaktadır. Sisteme giriş olarak, biyokimya parametrelerinden demir eksikliği anemisi hastalığı için temel belirleyiciler olan Serum demiri, Serum demir bağlama kapasitesi (SDBK) ve

Ferritin enzimleri kullanılarak, çıkış olarak da Anemi(+) ve Anemi(-) değerlendirmelerinde bulunulmuştur. Tasarlanan sisteme 96 hasta verisi değerlendirilmiştir. Karar destek sisteminin sonuçları, doktorun verdiği kararlarla tamamen örtüşmüştür (Doğan ve Türkoğlu, 2008).

Mustafa Danacı, Mete Çelik ve A. Erhan Akkaya tarafından 2010 yılında gerçekleştirilen çalışmada kanser çeşitlerinden biri olan ve kadınlar arasında en sık görülen meme kanseri hakkında kısa bilgi verilmiştir. Daha sonra Xcyt örüntü tanıma programı yardımı ile doku hakkında genel veriler elde edilmiş, Weka programı kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi yapılmıştır (Danacı vd., 2010).

3.3. Bankacılık ve Borsa Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Nihal Ata, Ercengül Özkok ve Uğur Karabey tarafından 2007 yılında gerçekleştirilen bu çalışmada, yaşam çözümlemesi yöntemlerini veri madenciliği konusu çerçevesinde ele aldıktan sonra kredi kartı sahiplerine ait bir veri kümesi için yaşam olasılıkları, hazard olasılıkları ve regresyon modelleri incelemiştir. Buna göre çalışmada yaş, gelir ve medeni durumun, müşterilerin kredi kartı kullanmayı bırakmalarını etkileyen önemli risk faktörleri olduğu görülmüştür (Ata vd., 2008).

Ali Sait Albayrak ve Şebnem Koltan Yılmaz tarafından 2009 yılında gerçekleştirilen bir çalışmada, İMKB 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerinden yararlanarak veri madenciliği tekniklerinden birisi olan karar ağaçları tekniği uygulanmıştır. Seçilen finansal göstergelere göre sanayi ve hizmet sektörlerinde faaliyet gösteren firmaları ayıran en önemli değişkenler saptanmıştır. Ayrıca Ali Sait Albayrak tarafından gerçekleştirilen başka bir çalışmada, yerli ve yabancı olarak önceden grup üyeliği belirlenmiş bankaların sınıflandırmasında yaygın olarak kullanılan veri madenciliği tekniklerinden, diskriminant, lojistik regresyon ve karar ağaç modelleri karşılaştırılmıştır. Üç sınıflandırma tekniği, bankalarla ilgili seçilmiş likidite, gelir-gider, karlılık ve faaliyet oranları kullanılarak karşılaştırılmaktadır. Araştırmanın sonuçları, bankaların sınıflandırmasında karar ağaç modelinin geleneksel diskriminant ve lojistik regresyon modellerine üstünlük sağlayarak alternatif etkili bir sınıflandırma tekniği olarak kullanılabileceğini göstermiştir (Albayrak ve Yılmaz; Albayrak, 2009).

H. Ali Ata ve İbrahim H. Seyrek tarafından 2009 yılında gerçekleştirilen bu çalışmada, denetçiler tarafından yaygın olarak bilinmeyen bazı veri madenciliği teknikleri, finansal tablolardaki hileleri tespit etmeye yardımcı olmak üzere kullanılmıştır. Çalışma İMKB'de işlem gören ve imalat sektöründe faaliyet gösteren 100 firmanın bilgilerine dayalı olarak gerçekleştirilmiştir. Araştırma sonucunda kaldırıcı oranı ve aktif karlılık oranının finansal tablo hilesini tespit etmede önemli finansal oranlar olduğu belirlenmiştir (Ata ve Seyrek, 2009).

İpek Savaşçı ve Rezan Tatlıdil tarafından 2006 yılında müşteri ilişkileri yönetimi üzerine bir çalışma gerçekleştirilmiştir. Bu çalışmada bireysel bankacılık alanında uygulanan müşteri ilişkileri yönetim süreci incelenmiş ve müşteri sadakatinin yaratılmasını sağlayan kredi kartlarında uygulanan CRM stratejileri değerlendirilmiştir (Savaşçı ve Tatlıdil, 2006).

3.4. Eğitim Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Konya Selçuk Üniversitesi’nde Onur İnan(2003) tarafından, hazırlık sınıfı, birinci sınıf ve mezun durumunda olan öğrenciler üzerinde, üniversite veri tabanındaki veriler kullanılarak; öğrencilerin başarılarını etkileyen etmenler, başarı düzeyleri, üniversitede kazanan öğrenci portföyleri ve mezun olamayan öğrencilerin okulu bitirmelerini etkileyen etmenler üzerinde çalışmalar gerçekleştirilmiş ve sonuçları yorumlanmıştır (İnan, 2003).

Serdar Çiftci(2006) tarafından gerçekleştirilen çalışmada, uzaktan eğitime katılan öğrencilerin ders çalışma etkinliklerinin değerlendirilmesi için yapılan anketler ve log dosyaları karşılaştırılarak, sonuçların farklı olup olmadıkları incelenmiştir (Çiftçi,2006). Bu çalışmaya benzer bir çalışma olarak Serdar Savaş ve Nursal Arıcı tarafından 2009 yılında gerçekleştirilen bir çalışmada, web tabanlı uzaktan eğitim için video destekli ve animasyon destekli öğretim modeline uygun iki farklı öğretim materyali, bu materyallerin öğrenci başarısı üzerindeki etkilerinin incelenmesi için hazırlanmıştır. Analiz sonucunda video destekli öğretim materyallerinin animasyon destekli öğretim materyallerine göre öğrenci başarısını daha olumlu etkilediği belirlenmiştir (Savaş ve Arıcı, 2009).

Y. Ziya Ayık, Abdulkadir Özdemir ve Uğur Yavuz tarafından yapılan çalışmada, Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanılarak incelenmiştir (Ayık vd., 2007).

Ahmet Selman Bozkır, Ebru Sezer ve Bilge Gök (2009) tarafından gerçekleştirilen bir çalışmada, ÖSYM tarafından 2008 ÖSS adayları için resmi internet sitesi üzerinden yapılan anket verileri üzerinde veri madenciliği yöntemleri kullanılarak, öğrencilerin başarılarını etkileyen faktörler araştırılmıştır. Bu araştırmada, veri madenciliği yöntemlerinden karar ağaçları ve kümeleme kullanılmıştır (Bozkır vd., 2009). Buna benzer bir çalışma olarak Şenol Zafer Erdoğan ve Mehpare Timor tarafından 2005 yılında gerçekleştirilen bir çalışmada, öğrencilerin üniversite giriş sınavı sonuçları ve öğrencilerin başarıları arasındaki ilişki, kümeleme analizi ve k means algoritması teknikleri uygulanarak incelenmiştir (Erdoğan ve Timor, 2005). Bu çalışmanın KPSS’ye uygulanmış bir modeline benzeyen çalışmayı Hüseyin Özçınar 2006 yılında gerçekleştirmiştir. Frekans analizi ve regresyon analizi yöntemleri kullanılarak derslere ve yillara göre verinin özellikleri incelenmiştir. Oluşturulan regresyon modeli ile KPSS sonuçlarının değişimi üzerinde anlamlı katkısı olan değişkenler incelenmiş ve oluşturulan modellerin tahmin doğrulukları,

ortalama mutlak hata ve ortalama hata kareler kökü değerleri kullanılarak karşılaştırılmıştır (Özçınar, 2006).

Ahmet Selman Bozkır ve Ebru Sezer tarafından 2009 yılında gerçekleştirilen başka bir çalışmada ise Hacettepe Üniversitesi Beytepe Kampüsü'ndeki öğrenci ve çalışanların, gıda tüketim desenleri incelenmiştir. Çalışmada, karar ağaçları ve birlilik kuralları uygulanmıştır ve çalışma sonunda %80 başarıyla, gıda tüketim deseninin ortaya çıkarıldığı görülmüştür (Bozkır ve Sezer, 2009).

Hidayet Takçı ve İbrahim Soğukpinar tarafından 2002'de gerçekleştirilen bir çalışmada kütüphane sitesi web günlüklerine dayalı olarak kütüphane kullanıcılarının erişim örüntüleri bulunmaya çalışılmıştır. Bu çalışma yapılrken istatistiksel yöntemler kullanılmıştır (Takçı ve Soğukpinar, 2002).

Murat Kayri tarafından 2008 yılında gerçekleştirilen bir çalışmada, öğrencilerin performans göstergelerinin sürekli izlenebilmesi ve ürünler arasındaki örüntünün bilgisayar sistemleri tarafından oldukça kolay yapılabildiği e-portfolyo değerlendirmeleri için veri madenciliğinde kullanılan yöntemlerin alternatif bir ölçme yaklaşımı olarak kullanımı önerilmektedir (Kayri, 2008).

3.5. Ticari Alanda Gerçekleştirilen Veri Madenciliği Uygulamaları

Anarberk Kalikov(2006) tarafından, bir yayinevi firmasının internet sitesindeki veriler dikkate alınarak, veri madenciliği birlilik kuralları tekniği ile sepet ve sipariş tabloları incelenmiştir. Hangi ürünlerin kategorisinin değiştirilmesi gereği, kullanıcıların meslek ve ilgi alanı dağılımları, müşteri ilgi alanlarına göre satış grafikleri ve kullanıcıların ödeme seçenekleri ile ilgili bir veri madenciliği uygulaması gerçekleştirilmiştir (Kalikov, 2006).

Sinem Akbulut(2006) tarafından yapılan çalışma, bir kozmetik markasının müşteri gruplarını ve ayrılma eğilimi gösteren müşteri kesitini belirleyerek; bu müşterilere özel pazarlama stratejileri geliştirilmesini hedeflemektedir. Bölümleme için kümeleme teknikleri, ayrılacak müşteri kesitini belirlemek için sınıflama teknikleri kullanılmıştır (Akbulut, 2006).

Feridun Cemal Özçakır ve A. Yılmaz Çamurcu (2007) tarafından gerçekleştirilen bir çalışmada, bir firmanın pastane satış verileri üzerinde veri madenciliği uygulamak için birlilik kuralları ile bir yazılım tasarlanmıştır. Genelde aynı ürün grubuna ait ürünlerin, en sık birlikte satın alınan ürünler olduğu görülmüştür (Özçakır ve Çamurcu, 2007).

Feyza Gürbüz, Lale Özbakır ve Hüseyin Yapıcı(2008) tarafından gerçekleştirilen başka bir çalışmada, Türkiye'de bir hava yolu işletmesinin parça söküm raporları üzerinde veri madenciliği çalışması gerçekleştirilmiştir. Çalışmanın amacı, uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kural

geliştirmektir. Sonuç olarak parçaların ikaz seviyelerini temsil edecek anlamlı bir kural elde edilmiş ve bulunan kurallar doğrulukları ve güvenilirlikleri bakımından test edilmiştir (Gürbüz vd., 2009).

Mehmet Aydin Ulaş tarafından 2001 yılında yapılan bir yüksek lisans çalışmasında, sepet analizi gerçekleştirilmiştir. Büyük süpermarket zinciri olan Gima Türk A.Ş.'nin verileri üzerine Apriori algoritması uygulanmış ve ortaya çıkan sonuçlar incelenmiştir. Ayrıca mal satışları arasındaki ilişkileri bulmak amacıyla da, bileşen analizi ve k-ortalama öbeklemesi metotları kullanılmıştır (Ulaş, 2001).

Çağatan Taşkin ve Gül Gökay Emel tarafından 2010 yılında veri madenciliğinde kümeleme yaklaşımları ve Kohonen ağları ile perakendecilik sektöründe bir uygulama gerçekleştirilmiştir. Bu uygulamada; bir perakende işletmenin müşterilerinin Kohonen ağları ile kümelenmesi ele alınmıştır. Kümeleme analizinin amacı; ele alınan işletmeye, pazar böülümlendirmesi ve hedef pazar seçimi gibi stratejik pazarlama kararlarında yardımcı olması için önceden bilinmeyen kritik müşteri özellikleri ve önem derecelerini de ortaya çıkararak gerekli öngörüyü sağlamaktır (Taşkin ve Emel, 2010).

Fatma Güntürkün 2007 yılında işletmelerin kalite iyileştirmelerini araştıran bir yüksek lisans çalışması hazırlamıştır. Ayrıca bu çalışmada, sürücü koltuğu kalitesi için müşteri memnuniyeti verisi analiz edilmiştir. Müşterinin sürücü koltuğundan memnuniyetini etkileyen en önemli değişkenlerin belirlenmesi için karar ağaçları yaklaşımı uygulanmıştır. Bu uygulamadan elde edilen sonuçlar diğer bir çalışmada aynı veri kümese uygulanmış ve lojistik regresyon analizinden elde edilen sonuçlarla karşılaştırılmıştır (Güntürkün, 2007).

3.6. Telekomünikasyon Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Umman Tuğba Şimşek Gürsoy tarafından 2010 yılında Türkiye'de telekomünikasyon sektöründe faaliyet gösteren büyük bir firmannın, ayrılma eğilimi gösteren müşterileri belirlenerek; bu müşterilere özel pazarlama stratejileri geliştirilmesi hedeflenmiştir. Ayrılacak müşteri profilini belirlemek için Lojistik Regresyon Analizi ve sınıflandırma tekniklerinden Karar Ağaçları kullanılmış ve uygulamanın sonuçları sunulmuştur (Gürsoy, 2010).

Selman Bozkır, S. Güzin Mazman ve Ebru Akçapınar Sezer tarafından 2010 yılında sosyal ağ kullanımına yönelik bir çalışma gerçekleştirilmiştir. Bu çalışmada güncel sosyal paylaşım sitesi facebook üzerinde kullanıcı şablonları incelenmiştir. Facebook kullanım süresi ve erişim sıklığı 570 facebook kullanıcısı üzerinde incelenerek sonuçları ortaya konmuştur (Bozkır vd., 2010).

4. SONUÇLAR

Tüm dünyada olduğu gibi ülkemizde de veri madenciliğine verilen önem ve gösterilen ilgi her geçen yıl artmaktadır. Veri madenciliğinin kullanım alanları genişleyerek yayılmaktadır. Bu çalışmada Türkiye'de yapılan veri madenciliği uygulamaları incelenmiş ve geçmişten günümüze kadar gerçekleştirilen veri madenciliği çalışmaları anlatılmıştır.

Türkiye'de gerçekleştirilen veri madenciliği çalışmaları, eğitim, ticari, mühendislik, bankacılık, borsa, tıp ve telekom olmak üzere, kullanım alanlarına ayrılmış ve her alanda gerçekleştirilen uygulamalar kendi içinde değerlendirilmiştir. Buna göre eğitim alanında gerçekleştirilen analizlerin çoğu öğrenci başarısı üzerine analizler gerçekleştirmek için yapılmıştır. Bu alanda gerçekleştirilen analiz uygulamalarının, sonraki nesiller için öngörü oluşturmak adına kullanılması, eğitim faaliyetlerine çok yararlı olduğu ve olacağının düşünülmektedir. Ticari alanda gerçekleştirilen uygulamaların tamamına yakını müşteri analizi ve pazar analizi ile ilgili olmuştur. Gerçekleştirilen bu çalışmalar sayesinde firmaların yeni pazarlar elde etmesi, mevcut pazarları koruması ve geliştirmesi, müşteri memnuniyeti, yeni müşteri kazanma ve var olan müşteriyi koruma gibi bilgileri sağlama amaçlanmaktadır. Mühendislik alanında gerçekleştirilen veri madenciliği uygulamalarının çoğu, yeni algoritmalar ortaya çalışma veya var olan algoritma ve teknikleri geliştirme yönünde olmuştur. Bunun sonucu olarak da kullanılan alanla ilgili daha uygun algoritmalar ve var olan algoritmaların türevleri ortaya çıkmıştır. Ancak geliştirilen algoritmaların sadece belirli algoritmalar üzerinde olması bu alanda bir eksik olarak görülmüştür. Bankacılık ve borsa alanında gerçekleştirilen çalışmalar daha çok tahmin gerçekleştirmek amacıyla yapılmıştır. Mevcut müşteri portföyünden, gelecekte karşılaşılacak kredi vb. mali konularda risk analizleri gerçekleştirilmiştir. Ayrıca şirket profillerinin incelenmesi ve hisse senetleri üzerine de araştırmalar gerçekleştirilmiştir. Veri madenciliğinin tahmin yönünün en etkili kullanıldığı alanlardan bir tanesi olarak bankacılık ve borsa alanını göstermek mümkündür. Tıp alanında gerçekleştirilen uygulamaların hastalık belirtileri ve var olan belirtilerden şablon ortaya çalışma amacıyla yapıldığı görülmüştür. Ülkemizde veri madenciliği çalışmalarının tıp alanında çok daha çeşitli ve etkili kullanılması gereği, bu konuda eksikliğin olduğu görülmüştür. Son olarak telekom alanında yapılan çalışmaların müşteri ve kullanıcı profili analizleri için gerçekleştirildiği ortaya çıkmıştır.

Dünyadaki teknolojik gelişmelere paralel olarak ülkemizde de veri madenciliği uygulamaları gittikçe artmaktadır. Ancak incelenen çalışmalar da göstermektedir ki kurum ve kuruluşların çoğu müşteri/kullanıcı analizlerine yönelmiştir. Bunun yanı sıra kurum ve kuruluşların kendi bünyelerinde veri madenciliğini kullanarak, gelişmelerini buna dayalı gerçekleştirmeleri faydalara olacaktır. Ayrıca veri madenciliğinin kullanıldığı alanların çeşitlendirilmesi de, gerek ülkemiz kurum ve kuruluşlarına, gerekse bu kurum ve kuruluşlardan hizmet ve ürün alan ülkemiz insanlarına büyük faydalara sağlayacaktır.

KAYNAKLAR (REFERENCES)

Adriaans, P. ve Zantinge, D., (1997), Data Mining, , Boston, MA, USA Addison Wesley Longman Publishing.

Akbulut, S., (2006) Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Segmentasyonu, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

Akgöbek, Ö. ve Çakır, F., (2009), “Veri Madenciliğinde Bir Uzman Sistem Tasarımı”, Akademik Bilişim 09, 11-13 Şubat Harran Üniversitesi, Şanlıurfa, 801-806.

Akpınar, H., (2000), “Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği”. İ.U. İşletme Fakültesi Dergisi, Cilt 29, S 1, 1-22.

Aksøy, B., (2009), Cluster Analysis Of Decompression Illness, Galatasaray University, Institute of Science and Engineering.

Albayrak, A.S. ve Yılmaz, Ş.K., (2009), “Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama”, S.D.Ü. İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14, No 1, 31-52.

Albayrak, A.S., (2009), “Classification of Domestic and Foreign Commercial Banks in Turkey Based On Financial Efficiency: A Comparison of Decision Tree, Logistic Regression and Discriminant Analysis Models”, S.D.Ü. İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14, No 2, 113-139.

Albayrak, M., (2008), EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti, Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

Ata, A.H. ve Seyrek, İ.H., (2009), “The Use of Data Mining Techniques in Detecting Fraudulent Financial Statements: An Application on Manufacturing Firms”, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14, No 2, 157-170.

Ata, N., Özök, E. ve Karabey, U., (2008), “Survival Data Mining: An Application To Credit Card Holders”, Sigma Mühendislik ve Fen Bilimleri Dergisi, Cilt 26, No 1, 33-42.

Ayık, Y.Z., Özdemir, A. ve Yavuz, U., (2007), “Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi”, Sosyal Bilimler Enstitüsü Dergisi, Cilt 10, No 2.

Baloğlu, U.B., (2006), DNA Sıralarındaki Tekrarlı Örüntülerin ve Potansiyel Motiflerin Veri Madenciliği Yöntemiyle Çıkarılması, Fırat Üniversitesi, Fen Bilimleri Enstitüsü.

Bilgin, T.T. ve Çamurcu, A.Y., (2008), "Çok Boyutlu Veri Görselleştirme Teknikleri", Akademik Bilişim 2008, 30 Ocak - 01 Şubat. Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 107-112.

Bilgin, T.T., (2009), "Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları ve Yazılım Geliştirme Ortamları", Akademik Bilişim 09, 11-13 Şubat, Harran Üniversitesi, Şanlıurfa, 807-814.

Bozkır, A.S. ve Sezer, E., (2009), "Usage of Data Mining Techniques in Discovering The Food Consumption Patterns of Students and Employees of University", Balkan-Kafkas ve Türk Devletleri Uluslararası Mühendislik Sempozyumu, 22-24 October, Isparta, 104-109.

Bozkır, A.S., Mazman, S.G., ve Sezer, E.A., (2010), "Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case", 2nd International Symposium on Information Management in a Changing World", 22-24 September, Hacettepe University, Ankara, 145-152.

Bozkır, A.S., Sezer, E. ve Gök, B., (2009), "Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarısını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti", 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 37-43.

Çalışkan, S.K. ve Soğukpinar, İ., (2008), "KxKNN: K-Means ve K En Yakın Komşu Yöntemleri ile Ağlarda Nüfuz Tespiti", 2. Ağ ve Bilgi Güvenliği Sempozyumu, 16-18 Mayıs, Girne, 120-124.

Çiftci, S., (2006), Uzaktan Eğitimde Öğrencilerin Ders Çalışma Etkinliklerinin Log Verilerinin Analiz Edilerek İncelenmesi, Yüksek Lisans Tezi, Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü.

Danacı, M., Çelik, M. ve Akkaya, A.E., (2010), "Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhis", Akıllı Sistemlerde Yenilikler ve Uygulama Sempozyumu, 21-24 Haz. 2010, Kayseri, 9-12.

Doğan, Ş. ve Türkoğlu, İ., (2008), "Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees", International Journal of Science & Technology, Cilt 3, No 1, 85-92.

Doğan, Ş., ve Türkoğlu, İ., (2007), " Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees", Doğu Anadolu Bölgesi Araştırmaları Dergisi, Cilt 5, No 2, 163-169.

Doğan, Y., (2004), A Data Mining Based Classification Algorithm for Tactical Underwater Sensor Networks, Yüksek Lisans Tezi, Turkish Naval Academy, Computer Engineering.

Duru, N. ve Canbay, M., (2007), “Veri Madenciliği ile Deprem Verilerinin Analizi”, International Earthquake Symposium, Kocaeli, 556-560.

Erdem, C., (2006), Density Based Clustering Using Mathematical Morphology, Yüksek Lisans Tezi, Middle East Technical University, Information Systems.

Erdoğan, Ş.Z. ve Timor, M., (2005), “A Data Mining Application In A Student Database”, Journal Of Aeronautics and Space Technologies, Cilt 2, No 2, 53-57.

Güntürkün, F., (2007), A Comprehensive Review Of Data Mining Applications In Quality Improvement And A Case Study, Yüksek Lisans Tezi, Middle East Technical University, Statistics.

Gürbüz, F., Özbakır, L. ve Yapıçı, H., (2009), “Türkiye’de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madenciliği Uygulaması”, Gazi Üniversitesi Mimarlık Mühendislik Fakültesi Dergisi, Cilt 24, No 1, 73-78.

Gürsoy, U.T.Ş., (2010), “Customer Churn Analysis in Telecommunication Sector”, İstanbul University Journal of the School of Business Administration, Cilt 39, No 1, 35-49.

Hand, D.J., (1998), “Data Mining: Statistics and More?”, The American Statistician, Cilt 52, 112-118.

İnan, A., Privacy Preserving Distributed Spatio-Temporal Data Mining, Yüksek Lisans Tezi, Sabancı University, Computer Science and Engineering, 2006.

İnan, O., (2003), Veri Madenciliği, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.

Jacobs, P., (1999), “Data Mining: What General Managers Need to Know”, Harvard Management Update, Cilt 4, No 10, 8.

Kalikov, A., (2006), Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

Kastro, Y., (2006), A Defect Prediction Method For Software Versioning, Yüksek Lisans Tezi, Boğaziçi University, Computer Engineering.

Kayaalp, K., (2007), Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü.

- Kayıri, M.**, (2008), “Elektronik Portfolyo Değerlendirmeleri İçin Veri Madenciliği Yaklaşımı”, Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, Cilt 5, No 1, 98-110.
- Kılıç, Y.**, (2009), Mining Association Rules For Quality Related Data In An Electronics Company, Middle East Technical University, Industrial Engineering.
- Kitler R. ve Wang W.**, (1998), “The Emerging Role of Data Mining”, Solid State Technology, Cilt 42, No 11, 45.
- Özçakır, F.C. ve Çamurcu, A.Y.**, (2007), “Birlikte Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması”. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl 6, No 12, 21-37.
- Özçınar, H.**, (2006), KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü.
- Özekes, S.**, (2003), "Data Mining Models and Application Areas", İstanbul Commerce University Journal of Science, No.3, 65-82.
- Piramuthu, S.**, (1998), “Evaluating Feature Selection Methods For Learning in Data Mining Applications”, Thirty-First Annual Hawaii International Conference on System Sciences, IEEE Computer Society, 6-9 January, Kohala Coast Hawaii USA, 294.
- Savaş, S. ve Arıcı, N.**, (2009), Web Tabanlı Uzaktan Eğitimde İki Farklı Öğretim Modelinin Öğrenci Başarısı Üzerindeki Etkilerinin İncelenmesi, 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 1229.
- Savaşçı, İ. ve Tatlıdil, R.**, (2006), “Bankaların Kredi Kartı Pazarında Uyguladıkları CRM (Müşteri İlişkiler Yönetimi) Stratejisinin Müşteri Sadakatine Etkisi”, Ege Akademik Bakış Dergisi, Cilt 6, No 1, 62-73.
- Shearer, C.**, (2000), “The Crisp-DM Model: The New Blueprint for Data Mining” Journal of Data Warehousing, Cilt 5 No 4, 13-23.
- Sıramkaya, E.**, (2005), Veri Madenciliğinde Bulanık Mantık Uygulaması, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.
- Takçı, H. ve Soğukpinar, R.**, (2002), İ., "Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi", Bilgi Dünyası, Cilt 3, Sayı 1, 12-26.
- Taşkin, 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, Cilt 15, No 3, 395-409.

Thuarisingham, B.M., (2003), Web Data Mining and Applications in Business Intelligence and Counter Terrorism, CRC Press LLC, Boca Raton, FL,USA.

Toprak, S., (2004), Data Mining For Rule Discovery in Relational Databases, Middle East Technical University, Computer Engineering.

Toprak, S.D., (2005), A New Hybrid Multi-Relational Data Mining Technique, Yüksek Lisans Tezi, Middle East Technical University, Computer Engineering.

Ulaş, M.A., (2001), Market Basket Analysis For Data Mining, Yüksek Lisans Tezi, Boğaziçi University, Computer Engineering.

Yavaş, G., (2003), Using A Data Mining Approach For The Prediction of User Movements in Mobile Environments, Yüksek Lisans Tezi, Bilkent University, Institute of Engineering and Science.

Yıldırım, P., (2008), Uludağ, M. ve Görür, A., "Hastane Bilgi Sistemlerinde Veri Madenciliği", Akademik Bilişim 2008, 30 Ocak - 01 Şubat, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 429-434.

Yıldız, B., (2010), Impacts Of Frequent Itemset Hiding Algorithms On Privacy Preserving Data Mining, İzmir Institute of Technology, Computer Engineering.