Veri Madenciliği

Eda Coşlu
Mehmet Akif Ersoy Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, BURDUR
edacoslu@hotmail.com

Büyük miktardaki veriler içerisinden önemli olanlarını bulup çıkarmaya Veri Madenciliği denir. Veriler üzerinde çözümlemeler yapmak amacıyla ve veriyi çözümleyip bilgiye ulaşabilmek için veri madenciliği yöntemi ortaya çıkmıştır. Veri madenciliği bir sorgulama işlemi veya istatistik programlarıyla yapılmış bir çalışma değildir. Veri madenciliği milyarlarca veri ve çok fazla değişken ile ilgilenir. Teknolojik gelişmeler dünyada gerçekleşen birçok işlemin elektronik olarak kayıt altına alınmasını, bu kayıtların kolayca saklanabilmesini ve gerektiğinde erişilebilmesini hem kolaylaştırıyor, hem de bu işlemlerin her geçen gün daha ucuza mal edilmesini sağlıyor. Ancak, ilişkisel veri tabanlarında saklanan birçok veriden kararlar için anlamlı çıkarımlar yapabilmek bu verilerin bilinçli uzmanlarca analiz edilmesini gerektiriyor. Veri sayısı çok olduğu için bazı özel analiz algoritmaları geliştirilmiştir.

Veri madenciliği uygulamalarında alt yapı gereksinimi veri ambarı sayesinde sağlanır.

Veri madenciliği, özel ve kamu sektörü kuruluşlarında birçok şekilde kullanılabilmektedir. Bunlardan bazıları aşağıdaki gibi sıralanabilir:

Bir süpermarket müşterilerinin satın alma eğilimlerini irdeleyerek, promosyonlarını belli müşterilere yönlendirmesi, aynı kaynakla daha çok satışgerçekleştirmesine yardımcı olabilir.

Büyük bir süpermarketin en basit fatura kayıtları incelendiğinde, tıraş bıçağı alan müşterilerin %56'sının kalem pilde aldığı ortaya çıkmıştır, buna dayanarak firma, tıraş bıçağı ve kalem pil reyonlarını bir araya getirmek suretiyle kalem pil satışlarını %14 arttırmıştır. Ürünler ve satışları arasındaki bu ilişkilerin belirlenmesiyle satış stratejileri değiştirilip kazancın arttırılması mümkündür.

- Bankalar kredi kararlarında kredi isteyenlerin özelliklerini ve davranışlarını irdeleyerek batık kredi oranını azaltabilir.
- Havayolları sürekli müşterilerinin davranış biçimlerini irdeleyerek dahaetkin fiyatlandırma ile kârlılıklarını artırabilirler.

Bir telefon şirketi müşteri davranışlarından öğrendikleri ile yeni hizmetlergeliştirerek, müşteri bağlılığını ve kârlılığını artırabilir.
Maliye Bakanlığı Gelir İdaresi, şirketleriçin risk modelleri kurarak vergi incelemelerini daha etkin yönlendirip,vergi kaçaklarını azaltabilir.
Hastaların teşhis ve tedavi maliyetleriirdelenerek hastalık riskinin ilk aşamada tespiti, kontrolü ve kaynakplanlama açısından faydalı olur.

A.Kusiak ve arkadaşları tarafından akciğer deki tümörün iyi huylu olup olmadığına dair, karar destek amaçlı bir çalışma yapılmıştır. İstatistiklere göre Amerika da 160.000 den fazla akciğer kanseri vakasının olduğu ve bunların %90'ının öldüğü belirlenmiştir. Bu bağlamda bu tümörün erken ve doğru olarak teşhisi önem kazanmaktadır. Noninvaziv testler ile elde edilen bilgi sayesinde %40-60 oranında doğru teşhis konabilmektedir. İnsanlar kanser olup olmadıklarından emin olmak için biyopsi yaptırmayı tercih etmektedirler. Biyopsi gibi invaziv testler hem maliyeti yüksek hem çeşitli riskler taşımaktadır. Faklı yerlerde ve farklı zamanlarda kliniklerde toplanan invaziv test verileri arasında yapılan veri madenciliği çalışmaları teşhiste %100 oranında doğruluk sağlamıştır.(A.Kusiak, K.H. Kernstine, J.A.Kern, K.A.McLaughlinandT.L.Tseng:
MedicalandEngineering Case Studies May, 2000)

Veri Madenciliği Süreci

1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Veri madenciliği algoritmasını uygulama
6. Sonuçları sunum ve değerlendirme

Veri temizleme: Veri tabanında yer alan tutarsız ve hatalı verilere gürültü denir. Verilerdeki gürültüyü temizlemek için; eksik değer içeren kayıtlar atılabilir, kayıp değerlerin yerine sabit bir değer atanabilir, diğer verilerin ortalaması hesaplanarak kayıp veriler yerine bu değer yazılabilir, verilere uygun bir tahmin (karar

ağacı, regresyon) yapılarak eksik veri yerine kullanılabilir.

Veri bütünleştirme: Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi işlemidir. Bunun en yaygın örneği cinsiyette görülmektedir. Çok fazla tipte tutulabilen bir veri olup, bir veri tabanında 0/1 olarak tutulurken diğer veri tabanında E/K veya Erkek/Kadın şeklinde tutulabilir. Bilginin keşfinde başarı verinin uyumuna da bağlı olmaktadır.

Veri indirgeme: Veri madenciliği uygulamalarında çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir.
Veri indirgeme yöntemleri; veri sıkıştırma, örnekleme, genelleme, birleştirme veya veri küpü, boyut indirgeme.

Veri Dönüştürme: Verinin kullanılacak modele göre içeriğini koruyarak şeklinin dönüştürülmesi işlemidir. Dönüştürme işlemi kullanılacak modele uygun biçimde yapılmalıdır. Çünkü verinin gösterilmesinde kullanılacak model ve algoritma önemli bir rol oynamaktadır.
Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Bu yüzden veri üzerinde normalizasyon işlemi yapılmalıdır.

Veri madenciliği algoritmasını uygulama: Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır.

Sonuçları sunum ve değerlendirme: Algoritmalar uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur. Örneğin hiyerarşik kümeleme yöntemi uygulanmış ise sonuçlar dendrogram grafiği sunulur.

Veri Madenciliği Yöntemleri

1. Sınıflandırma
2. Kümeleme
3. Birliktelik Kuralı

1.Sınıflandırma: Sınıflandırma veri madenciliğinin en çok kullanıldığı alandır. Var olan veri tabanının bir kısmı eğitim olarak kullanılarak sınıflandırma kuralları oluşturulur. Bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Veri madenciliğinin sınıflandırma grubu içerisinde en sık kullandığı teknik karar ağaçlarıdır. Aynı zamanda lojistik regresyon, diskriminant analizi, sinir ağları ve fuzzy setleri de kullanılmaktadır. İnsanlar verileri daima sınıflandırdıkları, kategorize ettikleri ve derecelendirdikleri için sınıflandırma, hem veri madenciliğinin temeli olarak hem de veri hazırlama aracı olarak da kullanılabilmektedir.

Sınıflandırma Süreci:

Verilerin sınıflandırılma süreci iki adımdan oluşur.

1-Veri kümelerine uygun bir model ortaya konur. Söz konusu model veri tabanındaki alan isimleri kullanılarak gerçekleştirilir. Sınıflandırma modelinin elde edilmesi için veritabanından bir kısım eğitim verileri olarak kullanılır. Bu veriler veritabanından rastgele seçilir.

Eğitim Verisi

Müşteri	Borç	Gelir	Risk
Ali	Yüksek	Yüksek	Kötü
Ayşe	Yüksek	Yüksek	Kötü
Fatma	Yüksek	Düşük	Kötü
Fuat	Düşük	Yüksek	İyi
Ece	Düşük	Düşük	Kötü
Ayla	Düşük	Yüksek	İyi

⇩

Sınıflandırma algoritması

⇩

Sınıflayıcı Model
EĞER Borç=YÜKSEK ise Risk=Kötü; EĞER Borç=DÜŞÜK Ve Gelir=DÜŞÜK ise RİSK=KÖTÜ; EĞER Borç=DÜŞÜK Ve Gelir=Yüksek ise RİSK=İYİ;

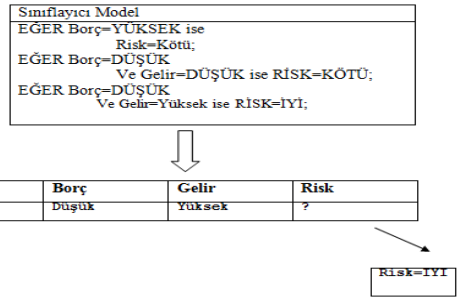
Sınıflandırma model kurma süreci

2-Test verileri üzerinde sınıflandırma kuralları belirlenir. Ardından söz konusu kurallar bu kez test verilerine dayanarak sınanır. Örneğin Ali adlı yeni bir banka müşterisinin kredi talebinde bulunduğunu varsayalım. Bu müşterinin risk durumunu belirlemek için örnek verilerden elde edilen karar kuralı doğrudan uygulanır. Bu müşteri için Borç=Düşük, Gelir=Yüksek olduğu biliniyorsa risk durumunun Risk=İYİ olduğu hemen anlaşılır.

Yukarıdaki test sonucunda elde edilen modelin doğru olduğu kabul edilecek olursa, bu model diğer veriler üzerinde de uygulanır. Elde edilen sonuç model mevcut ya da olası müşterilerin gelecekteki kredi talep risklerini belirlemede kullanılır.

Test Verisi

Müşteri	Borç	Gelir	Risk
Cüneyt	Yüksek	Düşük	Kötü
Fatih	Düşük	Yüksek	İyi
Gökhan	Düşük	Düşük	Kötü
Tarık	Yüksek	Yüksek	Kötü



Müşteri	Borç	Gelir	Risk
Ali	Düşük	Yüksek	?

Risk=İYİ

Örnek:

Borç	Gelir	Statü	Risk
Yüksek	Yüksek	İşveren	Kötü
Yüksek	Yüksek	Ücretli	Kötü
Yüksek	Düşük	Ücretli	Kötü
Düşük	Düşük	Ücretli	İyi
Düşük	Düşük	İşveren	Kötü
Düşük	Yüksek	İşveren	İyi
Düşük	Yüksek	Ücretli	İyi
Düşük	Düşük	Ücretli	İyi
Düşük	Düşük	İşveren	Kötü
Düşük	Yüksek	İşveren	İyi

Tablodan yararlanılarak karar ağacı oluşturulur. Karar ağacı oluşturulduktan sonra karar kuralları oluşturulur.

Kurallar:

Kural.1:Borç Yüksek ise Risk Kötü
Kural.2:Borç Düşük ve Gelir=Yüksek ise Risk=İyi
Kural.3:Borç Düşük ve Gelir=Düşük ve Statü=İşveren ise Risk=Kötü
Kural.4:Borç Düşük ve Gelir=Düşük ve Statü=İşveren ise Risk=Kötü

2.Kümeleme: Verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir ve kümeleme yöntemlerinin çoğu veri arasındaki uzaklıkları kullanır.Hiyerarşik Kümeleme yöntemleri en yakın komşu algoritması ve en uzak komşu algoritmasıdır.Hiyerarşik olmayan kümeleme yöntemleri arasında k-ortalamalar yöntemi sayılabilir. Uygulamada çok sayıda kümeleme yöntemi kullanılmaktadır. Bu yöntemler, değişkenler arasındaki benzerliklerden ya da farklılıklardan yararlanarak bir kümeyi alt kümelere ayırmakta kullanılmaktadır.

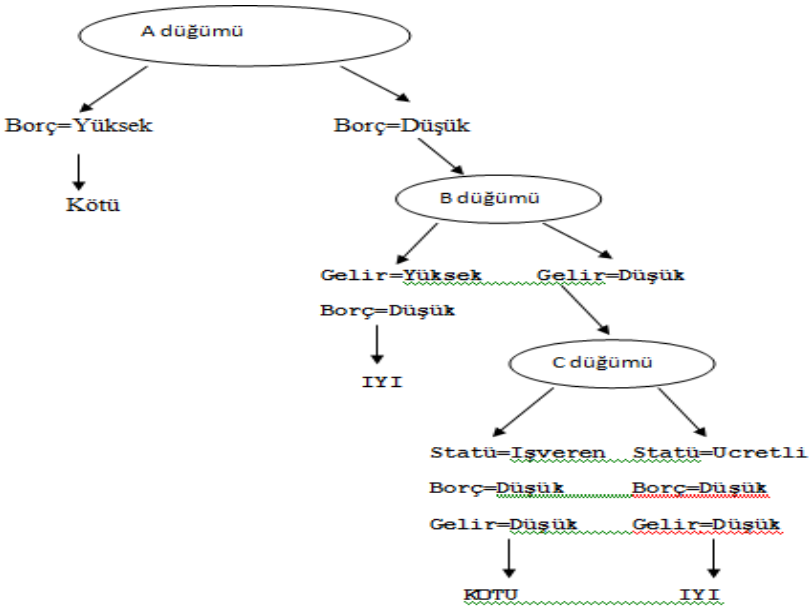
Hangi tekniğin kullanılacağı küme sayısına bağlı olmakla birlikte her iki tekniğin beraber kullanılması çok daha yararlıdır. Böylece hem sonuçları hem de iki tekniğin hangisinin daha uygun sonuçlar verdiğini karsılaştırmak mümkün olmaktadır.

Kümeleme analizinin amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya özetleyici bilgiler elde etmede yardımcı olmaktır. Kümeleme analizinin uygulanabilmesi için verilerin normal dağılımlı olması varsayımı olmakla birlikte, bu varsayım teoride kalmakta ve uygulamalarda göz ardı edilmektedir. Sadece uzaklık değerlerinin normal dağılıma uygunluğu ile yetinilmektedir. Bu varsayımın sağlanması durumunda kümeleme analizinde Kovaryans matrisi için farklı bir varsayım gerekmemektedir.

Karar Ağaçları ile Sınıflandırma

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 yapı ise “dal” olarak adlandırılır. (Quinlan,1993). Karar ağaçları sınıflama algoritmasını uygulayabilmek için uygun bir alt yapı sağlamaktadır. Karar ağacı oluşturmak için birçok yöntem geliştirilmiştir.Bunlar temel olarak:

1. Entropiye dayalı algoritmalar
2. Sınıflandırma ve Resresyon araçları
3. Bellek tabanlı sınıflandırma modelleri



‘Küme, birbirine yakın (benzer) nesnelerin çok boyutlu uzayda oluşturdukları bulutlar benzetmesi’ seklinde tanımlanabilir.(Hüseyin Tatlıdil, Uygulamalı Çok Değişkenli_istatistiksel Analiz, Ankara: Ziraat Matbaacılık, , 2002,s.330.) Kümeleme analizi ise; bu kümeleri oluşturma işlemidir.

Örnek:

Gözlem	X1	X2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Bu tabloya en yakın komşu algoritması uygulandığında;

Kümeler

Küme 1=1,2

Küme 2=4,5

Küme 3=3,4,5

Küme 4=1,2,3,4,5

1984 yılında Londra’da kolera salgını baş göstermiş. Çok ciddi ölümler kaydedilmiş(10675 kişi) John Snow bir harita üzerinde ölen kişilerin yerlerini işaretlediğinde kayıpların bazı bölgelerde yoğunlaştığını fark ediyor. O bölgede su pompalarına bakılıp atık su tesisindeki problem tespit edilerek koleradan meydana gelen ölümler engellenmiş.Ana sokaklardan birindeki su pompasının sapını çıkarmak kolera salgınının sonlanması için yeterli olmuştur.(JacquezGM,Grimson R, Waller LA. Theanalysis of discaseclusters, Part II: Indructiontotechriques.InfectControlHospEpid 1996; 17:385-97)Bu veri madenciliğinde kümeleme yönteminin ilk kez yapıldığı kağıt kalemle analizdir. Veri miktarı az olduğu için kağıt kalemle yapmakta bir sıkıntı yok ama günümüzde bu pek mümkün değil.

3-Birliktelik Kuralları: Veri tabanı içinde yer alan kayıtların birbiriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan veri madenciliği yöntemleridir. Özellikle pazarlama alanında uygulanmaktadır (Pazar sepet analizleri). Bu

yöntemler birlikte olma kurallarını belirli olasılıklarla ortaya koyar.

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” denilmektedir.

Örneğin; bir mağazadan parfüm alan müşterilerin %60’ının aynı alışverişte parfüm satın aldıklarını söylemek, bu birlikte gerçekleşen olaylara örnek olarak verilebilir.

Apriori Algoritması:

Birliktelik kurallarının üretilmesi için kullanılan en yaygın yöntemdir. Aşamaları:

- Destek ve güven ölçütlerini karşılaştırmak üzere eşik değerler belirlenir. Uygulamadan elde edilen sonuçların bu eşik değere eşit ya da büyük olması beklenir.
- Destek sayıları hesaplanır. Bu destek sayıları eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip satırlar çözümlemeden çıkarılır ve koşula uygun kayırlar göz önüne alınır.
- Bu şeçilen ürünler bu kez ikişerli gruplandırılarak bu grupların tekrar sayıları elde edilir. Bu sayılar eşik destek sayıları ile karşılaştırılır. Eşik değerden küçük değerlere sahip satırlar çözümlemeden çıkarılır.
- Bu kez üçerli, dörderli vb. gruplandırmalar yapılarak bu grupların destek sayıları elde edilir ve eşik değeri ile karılaştırılır, eşik değere uygun olduğu sürece işlemlere devam edilir.
- Ürün grubu belirlendikten sonra kural destek ölçütüne bakılarak birliktelik kuralları türetilir ve bu kuralların her birisiyle ilgili olarak güven ölçütleri hesaplanır.

Örnek: Müşteri alışverişleri

Müşteri	Aldığı ürünler
1	Makarna, yağ, meyve suyu, peynir
2	Makarna, ketçap
3	Ketçap, yağ, meyve suyu, bira
4	Makarna, ketçap, yağ, meyve suyu
5	Makarna, ketçap, yağ, bira

Apriori algoritması uygulandığında şu sonuçlar elde edilir:

- {ketçap, meyve suyu}-{yağ} (s=0,4 c=1.0)
- {ketçap, yağ}-{meyve suyu} (s=0,4 c=0.67)
- {yağ, meyve suyu}-{ketçap} (s=0,4 c=0.67)
- {meyve suyu}-{ketçap, yağ} (s=0,4 c=0.67)
- {yağ}-{ketçap, meyve suyu} (s=0,4 c=0.5)
- {ketçap}-{yağ, meyve suyu} (s=0,4 c=0.5)

Sonuç:

Veri Madenciliği istatistik biliminin teknolojiyle bütünleşmesi sonucu oluşmuş bir yöntemler serisidir. Bilgi teknolojilerinin gelişmesi ve konu ile ilgili yeni programların üretilmesi çalışmaları kolaylaştırmaktadır. Ancak veri madenciliği sadece program kullanmak değildir. Veri madenciliği için iş deneyimine, sorunları tanımlama becerisine ve temel istatistik bilgisine ihtiyaç vardır.

Veri madenciliği veriden bilgi üreterek ortalama kararlar yerine veriye dayalı özgün kararlar verilmesini destekleyen, satışları, kârlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır. Veriye dayalı kararların kalitesi ve güvenilirliği artar; bu veriye dayalı kararlarla çalışan kurumların kaynak kullanım etkinliği ve değer yaratma potansiyeli de gelişir.

Kaynaklar

[1] Veri tabanı yönetimi veri ambarı ders notu Yrd. Doç. Dr. Altan MESUTTrakya Üniversitesi Bilgisayar Mühendisliği

[2]Data MiningwithSql Server 2005

[3]Murray J.Mackinnon ve NedGlick, ‘Data Miningand Knowledge Discovery in Databases-AnOverview’, J.Statists., Vol.41, No.3, (1999), s.260.

[4] Veri madenciliğinde kümeleme algoritmaları ve kümeleme analizi Yasemin Koldere Akın

[5]Veri madenciliği yöntemleri Bilgisayar bilimleri ve mühendisliği 2.basım Dr.Yalçın Özkan