T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME YÜKSEK LİSANS PROGRAMI

MARKET VERİ TABANINDA VERİ MADENCİLİĞİ UYGULAMASI

Yüksek Lisans Tezi

Nuri Ender KARAGÖZ

0650Y38222

Danışman: Yrd. Doç. Dr. Dicle TAŞPINAR CENGİZ

İSTANBUL, 2007

T.C.

İSTANBUL TİCARET ÜNİVERSİTESİ

SOSYAL BİLİMLERİ ENSTİTÜSÜ

ONAY SAYFASI

Yüksek Lisans Öğrencisi Nuri Ender KARAGÖZ'ün "Market Veri Tabanında Veri Madenciliği Uygulaması" konunulu tez çalışması jurimiz tarafından İşletme Yüksek Lisans tezi olarak oybirliği/oyçokluğu ile başarılı bulunmuştur.

İmza

Tez Danışman	**: Yrd.Doç.Dr. Dicle Cengiz**	.....
Jüri Üyesi	**: Yrd.Doç.Dr. Ünal H. Özden**	.....
Jüri Üyesi	**: Prof.Dr. Münevver Turanlı**	.....

ONAYLI

Yukarıdaki juri kararı Enstitü Yönetim Kurulunun ...../...../2007 tarih ve ..... kararı ile onaylanmıştır.

Prof. Dr. Kerem ALKİN
Müdür

İstanbul Ticaret Üniversitesi
Sosyal Bilimleri Enstitüsü Müdürlüğüne
Eminönü Yerleşkesi- İstanbul

Halen Enstitünüzün İşletme Yönetimi yüksek lisans programı öğrencisiyim. Hazırlamakta olduğum tez tümüyle özgün bir çalışma olup YÖK ve İTİCU Lisansüstü Yönetmeliklerine uygun olarak hazırlanmıştır. Ayrıca bu çalışmayı yaparken bilimsel etik ve kurallarına tamamıyla uyduğumu; yararlandığım tüm kaynakları gösterdiğimi ve hiçbir kaynaktan yaptığım ayrıntılı alıntı olmadığını beyan ederim.

ÖZET

Bilgi sistemleri ve teknolojinin gelişmesi sonucunda büyük marketler, işletmeler ve diğer kuruluşlarda veritabanlarında kuruluşun amacına ve yapısına bağlı olarak çeşitli türlerde veri toplanmaktadır. Uygun yazılımların gelişimi ve firmaların topladığı veriyi kullanılabilir bilgiye çevirme isteği toplanan bu veriyi işleyerek, verinin içerisindeki kullanılabilir ve ilginç ilişkilerin, birlikteliklerin ve örüntülerin ortaya çıkarılmasını gerekli hale getirmiştir. Veri madenciliği bu gereklilikleri karşılayacak bir disiplin olarak ortaya çıkmıştır. Veri madenciliğinin temelini örüntü tanıma ve sınıflama problemleri üzerinde yoğunlaşan yapay zeka ve istatistik disiplinlerindeki gelişmeler oluşturmaktadır.

Bu çalışmada veri ambarı mimarisi, veri ambarlarında bilgi keşif süreci ile veri madenciliği kavramları ilişkisine değinilmiş veri madenciliği tanımı, diğer disiplinlerle ilişkileri, uygulama alanları, veri madenciliğinde karşılaşılan sorunlar, veri madenciliği modelleri ve teknikleri belirtilerek veri madenciliği süreci üzerinde durulmuştur.

Bu çalışmanın uygulama bölümünde mevcut market veri tabanı üzerinde veri madenciliği birliktelik analizi SPSS Clementine 10.1 programı Apriori ve Weka programı Aprori, Predictive Apriori ve Tertius algoritmaları kullanılarak belirli güven ve destek değerlerine göre yapılmış, elde edilen kurallar değerlendirilerek sonuçlara ulaşılmış ve kıyaslamalar yapılmıştır.

ABSTRACT

As a result of improvements in information systems and technology, different types of data have been collected in databases of big markets, companies and other organizational units in accordance with the aims and architectures of organizational units avalable software improvements and company requirements to transform collected data into usable and practical form have made essential to reveal interesting relations, association rules, patterns in data by processing the data. Data mining seemed a satisfactory dicipline to meet such kind of requirements. Pattern recognition, artificial intelligence which concentrates on classification problems and improvements in statistics constitutes data mining concept.

In this study, architecture of data warehouse, relations between data mining and data prediction process of database, data mining and other diciplines, application concepts, problems and techniques of data mining and process of data mining titles were held.

In the implementation phase of that study, association analysis on current market database has been assesed by using SPSS Clementine and Weka Softwares. Definite support and confidence values have been set and analyzed on Apriori Algorithms for both software and Pedictive and Tertius Algorithms for only Weka software. Finally, all results were evaluated and compared to each other for verification.

İÇİNDEKİLER

Sayfa No.

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No

KISALTMALAR

a.g.e **:** Adı Geçen Eser

C. **:** Cilt

s. **:** Sayfa

S. **:** Sayı

VTBK **:** Veri Tabanında Bilgi Keşfi

WEKA : Waikato Environment for Knowledge Analysis

KDD : Knowledge Discovery in Databases

VM :Veri madenciliği

VTYS :Veri Tabanı Yönetim Sistemleri

BK : Bilgi Keşfi

OLAP : Online Analytical Processing

GPRS : General Packet Radio Service

WAP : Wireless Application Protocol

GİRİŞ

Günümüzde alışverişlerden, bankacılık işlemlerinden, telefon kayıtlarından, uzaktan algılayıcılardan, uydulardan toplanan, devlet ve işletme yönetiminde yapılan işlemler sonucunda saklanan veriler gün geçtikçe çok büyük boyutlarda artmaktadır.

Bilgi sistemleri ve teknolojinin gelişmesi sonucunda; büyük marketler, işletmeler ve diğer kuruluşlarda veritabanlarında kuruluşun amacına ve yapısına bağlı olarak çeşitli türlerde veri toplanmaktadır. Uygun yazılımların gelişimi ve firmaların topladığı veriyi kullanılabilir bilgiye çevirme isteği toplanan bu veriyi işleyerek, verinin içerisindeki kullanılabilir ve ilginç ilişkilerin, birlikteliklerin ve örüntülerin ortaya çıkarılmasını gerekli hale getirmiştir. Veri madenciliği bu gereklilikleri karşılayacak bir disiplin olarak ortaya çıkmıştır. Veri madenciliğinin temelini örüntü tanıma ve sınıflama problemleri üzerinde yoğunlaşan yapay zeka ve istatistik disiplinlerindeki gelişmeler oluşturmaktadır. Ayrıca veri madenciliği, yapay zeka çalışmalarının uzantısı olan makine öğrenimi ve uzman sistemlerin yanı sıra, veri tabanları, optimizasyon, görselleştirme, yüksek performanslı paralel işlemciler gibi çeşitli disiplin ve teknolojilerdeki gelişmelerden de etkilenmektedir. Veri madenciliğinin veri tabanları üzerine uygulanmasıyla Veri Tabanlarında Bilgi Keşfi (VTBK) ortaya çıkmıştır. VTBK süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen veri madenciliği en önemli kesimi oluşturmaktadır. VTBK ile veri madenciliği terimlerinin eş anlamlı olarak da kullanılmasına neden olmaktadır.

VTBK genelde çok büyük hacimli verileri ele almakta kullanılmaktadır. VTBK, veri seçimi, veri temizleme, veri ön işleme, veri indirgeme, veri madenciliği algoritmasının uygulanması ve sonuçların değerlendirmesi basamaklarından oluşur. Büyük ölçekli veri tabanlarından anlamlı ve gizli örüntülerin çıkarılması olarak anılan veri madenciliği (VM) VTBK'nın bir adımı olarak nitelendirilebilir.

VM aracılığıyla, büyük veri kümelerinden oluşan veritabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi, istatistik, matematik disiplinleri, modelleme teknikleri, veritabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır.

Veri madenciliği özellikle elektronik ticaret, bilim, iş ve eğitim alanlarındaki uygulamalarda yeni ve temel bir araştırma sahası olarak ortaya çıkmaktadır. Veri madenciliği eldeki yapının veriden anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini formüle analiz etmeye ve uygulamaya yönelik çalışmaların bütününü içerir. Geniş veri kümelerinden desenleri, değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılmasının yanı sıra elektronik alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edilmesinde önemli rol oynamaktadır. İşletmelerin var olan yoğun rekabet ortamında doğru kararı en hızlı şekilde verebilmeleri için işletmeyle ilgili taraf ve işletme süreçleri hakkında detaylı bilgiye sahip olmaları gerekmektedir. Elektronik ortamda mevcut müşteri hakkında tutulan veri sayısının fazla olması doğru karara en hızlı şekilde ulaşmayı kolaylaştırmaktadır. Fakat doğru karara en hızlı şekilde ulaşmak verileri toplamakla mümkün değildir. Bu amaçla toplanan bu veri yığınlarını analiz edip, yorumlayarak anlamlı raporlar haline dönüştürerek yapılabilir.

VM, istatistik alanındaki birçok metodu kullanmasına rağmen, nesnelerin niteliklere değerlerine bağlı çıkarımlar yapmada bilinen istatistiksel metotlardan ayrılmaktadır. Örneğin ki-kare veya t testi gibi istatistiksel test yöntemleri, birden fazla nitelik arasında ilişki derecesini belirli bir güvenilirlik düzeyinde verebilirken, belirli nitelik değerleri arasındaki ilişkinin derecesini açığa çıkaramazlar. İstatistiksel yöntemler karar verme mekanizmasında VM disiplini ortaya çıkmadan önce çok sık kullanılırdı. Ancak bu yöntemlerin kullanım zorluğu (uzman kişileri tutma/ uzman kişilere başvurma), VM algoritmalarının uygulama zorluğu yanında çok fazladır. Büyük boyutlu yapısal veriyi saklama ve bu verilere etkin bir şekilde erişim sağlamakla yükümlü olan veri tabanı yönetim sistemlerinde (VTYS) veri düzenlemesi, ilgili organizasyonun işletimsel veri ihtiyacı doğrultusunda gerçekleştirilir. Bu işlem her zaman bilgi keşfi (BK) perspektifi ile birebir örtüşmez. Bu açıdan VM algoritması uygulanmadan önce veri ön işleme basamakları gerçekleştirilir. VT' deki veriler üzerinde gerçekleştirilen bu basamaklar, temizleme, boyut indirgeme, tür dönüşümleri, transfer, vb işlemlerdir.

Veri madenciliği, özet olarak çok büyük veri tabanlarından, önceden bilinemeyen, geçerli ve kullanılabilir bilginin çıkarılma işlemi olarak ifade edilebilir. VTBK (Knowledge Discovery in Databases) uygulamaları ile birlikte faaliyet alanına

yönelik karar destek mekanizmaları için gerekli ön bilgileri temin etmek için kullanılır. Veri madenciliğindeki amaç, toplanmış olan bilgilerin, bir takım istatistiksel yöntemlerle incelenip ilgili kurum ve yönetim destek sistemlerinde kullanılmak üzere değerlendirilmesidir. Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur. Yapılacak analizlerden sonra elde edilen verilerin bir istatistikçi gözü ile incelenip daha önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir veri madenciliği süreci olarak kabul edilmektedir.

Tezin ikinci bölümünde, veri ambarları tanımı, veri ambarı mimarisine değinilmiş veri ambarlarındaki verinin analizi ve kullanılacak tekniklerin belirlenmesi ve sistem tasarımı yapılırken üzerinde durulması gereken noktalar belirtilmiştir. Ayrıca veri ambarı ve veri tabanı arasındaki farklar da ortaya konmuştur.

Tezin üçüncü bölümünde, veri ambarlarında bilgi keşif süreci ile veri madenciliği kavramları ilişkisine değinilmiş, veri madenciliği tanımı, diğer disiplinlerle ilişkileri, uygulama alanları, veri madenciliğinde karşılaşılan sorunlar, veri madenciliği modelleri ve teknikleri belirtilerek veri madenciliği süreci üzerinde durulmuştur.

Uygulama bölümünde mevcut market alışverişi veri tabanı üzerinde veri madenciliği birliktelik analiz tekniği SPSS Clementine 10.1 programı Apriori ve WEKA programı Apriori, Predictive Apriori ve Tertius algoritmaları kullanılarak belirli güven ve destek değerlerine göre yapılmış, elde edilen kurallar değerlendirilerek sonuçlara ulaşılmış ve kıyaslama yapılmıştır.

1. VERİ AMBARLARI

Veri ambarları basit olarak veri madenciliği işleminin yapılacağı verilerin oluşturulduğu özel veri tabanlarıdır. Veri tabanlarındaki veriler ile analiz yapmak ve karar destek aşamasında faydalanmak, veri madenciliği ile mümkün olabilmektedir. Madenciliği yapılacak olan verinin de bazı vasıflara sahip olması gerekir. Bu vasıflar veri ambarı (Data Warehouse) ile sağlanmaktadır. Veri ambarlarının oluşturulması işlemi, verinin çeşitli kaynaklardan toplanarak, veriler içerisindeki uyuşmazlıklar ve hatalardan arındırılmasıdır.

Veri ambarları; bir kurumda gerçekleşen tüm operasyon el işlemlerin, en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen, tarihsel derinliği olan, fiziksel olarak operasyon el sistemlerden farklı ortamdaki yapılardır. [1]Günümüzün ticari işletmelerinde bilgi sistemleri canlı ve karar destek sistemleri şeklinde ikiye ayrılır.

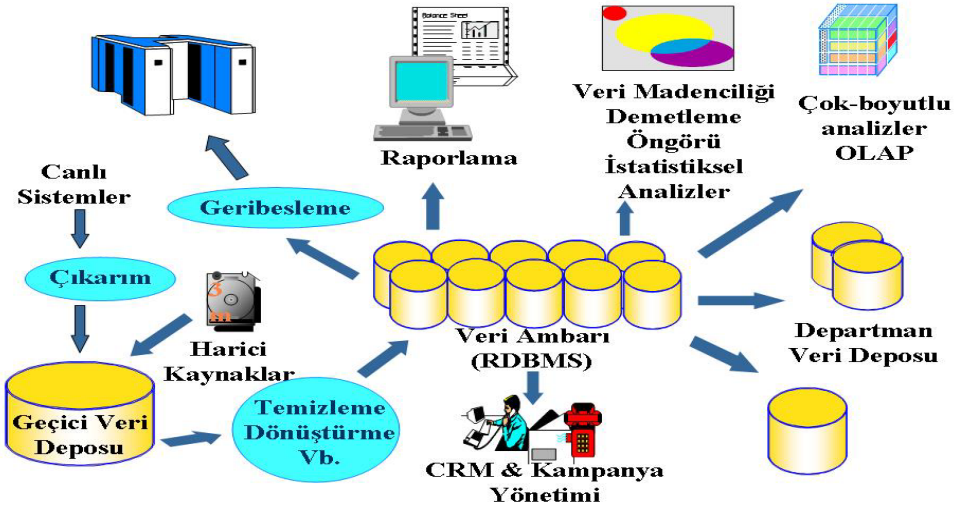
Günlük yapılan işleri gerçeklemek ve sonuçları saklamak için tasarlanan canlı sistemlerden marketlerde ya da mağazalarda stok takibi, üyelerin borçları, satış işlemleri ve ödeme kayıtları gibi güncel bilgilerin işlendiği ve tutulduğu bilgi sistemleri verilebilir. Bu tür canlı sistemlerde erişilebilirlik esastır yani veriye en kısa sürede ulaşmak, işlemleri en kısa sürede sona erdirmek hedeflenir. Örneğin süpermarkette bir satış işleminin mümkün olduğu kadar çabuk bitirilmesi istenir. Bu yüzden canlı sistemler çevrimiçi çalışma prensibi ile tasarlanırlar.

Karar Destek Sistemlerinde yer alan bilgiler çeşitli incelemelerden ve araştırmalardan geçirilerek yöneticilerin ileride işletmenin kârını ya da verimliliğini arttırması, gelecekte izlenecek şirket politikalarının belirlenmesi ve benzeri yönetimsel kararların alınmasını kolaylaştırır ve bu kararların daha doğru verilmesine yardımcı olurlar. Bu sistemlerde verilerin erişimi birinci kıstas değildir. Karar destek sistemlerinde öncelik performanstadır. Bu sistemlerde veriler, canlı sistemlere oranla

[1] Şule Özmen, İş Hayatı Veri Madenciliği ile istatistik Uygulamalarını Yeniden Keşfediyor. 2002. www.mimoza.marmara.edu.tr/~sozmen/teblig.php (Erişim Tarihi:30.08.2007)

çok daha büyük boyutlarda olduğundan verilerin incelenmesi ve bu incelemelerden sonuçlar çıkartılması, sistem kaynaklarını aşırı kullanmakta ve uzun süre almaktadır. Bu yüzden karar destek sistemlerinde, yapılacak incelemelerin ve araştırmaların performansını arttırmak için bir takım önlemler alınmış ve iyileştirmeler yapılmıştır.

Veri ambarı, bir karar destek sistemi olarak nitelendirilebilir. Veri ambarı aslında günlük işlemlerin gerçeklendiği canlı sistemlerin arka planında bulunmaktadır. Canlı sistemlerde oluşan veriler periyodik olarak veri ambarına aktarılırlar. Bu periyodun seçimi tamamen veri ambarını kullanan işletmenin ihtiyaçları doğrultusunda değişken olabilir. Dolayısı ile veri ambarı çevrimdışı olarak çalıştığından içerisindeki kayıtlar genellikle güncel olmayabilir. Veri ambarına belirli aralıklarla verileri gönderen canlı sistemlerin bir tane olma zorunluluğu da yoktur. Örneğin bir işletmenin içerisindeki farklı bölümler ve farklı günlük işlemleri gerçeklemek üzere tasarlanmış birbirlerinden habersiz ve bağımsız çalışan değişik canlı sistemlerdeki veriler, veri ambarındaki bir tek yapı içerisinden erişilebilecek bir şekilde toplanıp veri ambarına aktarılabilir. Bu aktarım süreci içerisinde veriler üzerinde veri ambarında önceden bulunan kayıtları aktarmama, kayıtlar içerisindeki bazı bilgileri değiştirmek, silmek ve benzeri bir takım işlemler de gerçeklenebilir. Bu ön işleme sonrasında değişik kaynaklardan yani değişik canlı sistemlerden toplanan veriler, anlamsal bütünlüğü sağlayacak şekilde veri ambarına yerleştirilirler. Bir veri ambarı mimarisi Şekil 1'deki gibidir.



Şekil 1. Veri Ambarı Mimarisi

Kaynak: Ahmet Cüneyt TANTUĞ, Veri Madenciliği ve Demetleme, İstanbul Teknik Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul, 2002, s.2

Şekil 1'deki gibi veri ambarına aktarılan veriler sadece işletmeye ait canlı sistemlerden gelmeyip işletmeye ait olmayan kaynaklardan da veri alımı mümkün olmaktadır. "Geri besleme" olarak gösterilen adım ile de veri ambarı ve canlı sistemler arasında bir ilişki kurulmuştur ve böylece canlı sistemler veri ambarından gelen istekleri karşılayabilmektedirler.

Veri ambarında veri oluşturulduktan sonra bu verinin elle veya gözle analizi yapılabilir. Bunun için OLAP (Online Analytical Processing) programları kullanılır. Bu programlar veriye her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak bakmayı ve incelemeyi sağlar. Böylece boyut bazında gurupluma, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik veya rapor olarak sunma olanağı sağlar.

Veri analiz tekniği seçilen veri modeli tipini ve içeriğini etkiler. Örneğin, sorgu ve raporlama becerisi kazandırılacaksa, veriyi normalize tutan bir veri modeli veriye en hızlı ve en kolay ulaşılmasını sağlar. Sorgulama ve raporlama becerisi temel olarak ilgili veri elemanlarının seçilmesi, özetlenerek kategoriler halinde gruplanması ve sonuçların sunulmasını içerir. Bu da doğrudan tablo taramasını kullanmayı gerektirir. Çok boyutlu veri analizi yapılacaksa, çok boyutlu veri yapısı uygun olacaktır. Bu tip bir analiz, verilere analiz boyutlarının birleşimleri bazında hızlı ve kolay erişimi sağlayan yapıyı destekleyen bir veri modelini gerektirir.

Veri ambarlarında beklenilen, hem organizasyonu hem de çevresini anlatan tutarlı ve yararlı bir bilgi kaynağına ulaşabilmektir. Sistemin tasarımı oluşturulurken, aşağıdaki noktalara dikkat etmek yararlı olacaktır

- Sistemin çözmesi istenen problem ayrıntılı bir biçimde tanımlanmalıdır.
- Sistemle ilgili hedefler, kısıtlamalar ve kritik başarı etkenleri sıralanmalıdır.
- Baslıca sistem bileşenleri ve ara yüzler, bileşenler arasındaki bağlantı veya iletişim yolları iyice ortaya konulmalıdır.
- Gelecekte yapılması olası iyileştirmeler, değişiklikler ve başka sistemlere geçişler hakkında öngörüler yapılmalıdır.
- Bütünsel bir geliştirme ve bakım programı ve sisteme destek verecek personel kaynağı planlanmalıdır.

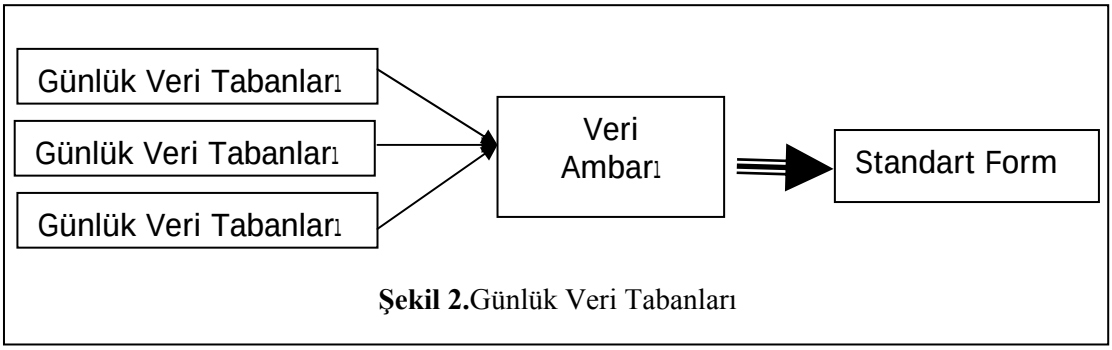
- Sistemi programa uygun bir şekilde geliştirebilmek ve uzun vadede bakımını yapıp yönetebilmek için gerekli bilgi, beceri ve diğer destek araçları belirlenmelidir. Veri ambarları uzun dönemli stratejik kararları destekler.

Bir veri ambarı yapısı; organizasyon içindeki bütün son kullanıcılara verileri ve işlem sonuçlarını sunan, en gelişmiş iletişimi sağlayan birbiriyle bütünleşik aşağıdaki alt bileşenlerden oluşur.

- Operasyonel Veri Tabanı / Harici Veri Tabanı Katmanı,
- Enformasyon Ulaşım katmanı,
- Veri Ulaşım Katmanı,
- Veri Diziin (Metadata) Katmanı,
- İşlem (process) Yönetim Katmanı,
- Uygulama Haberlesmesi Katmanı,
- Veri Ambarı Katmanı,
- Veri Sunum Katmanıdır.

1.1. Veri Ambarı ve Veri Tabanı Arasındaki Farklar

Büyük miktarda veri inceleme amacı üzerine kurulan veri madenciği ile veri tabanı olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük veri tabanlarının birleştirilmiş ve işlemeye daha uygun özetini saklamayı amaçlar (Şekil 2).



Şekil 2.Günlük Veri Tabanları

Günlük veri tabanlarından istenen özet bilgi seçilerek ve gerekli ön işlemeden sonra veri ambarında saklanır ve amaç doğrultusunda gerekli veri ambardan alınarak veri madenciliği çalışması için standart bir forma çevrilir.

Özetle veri tabanı içerisindeki bilgiler genelde anlık bilgilerdir. Yani o an için güncelliğini koruyan ancak belirli bir süre sonunda güncelliğini kaybedecek olan bilgilerdir. Ancak veri ambarı içerisindeki veriler genelde yığılarak birikirler ve verilerin geçerliliği çok daha uzun süre olmaktadır. Veri ambarı içerisinde ne kadar çok kayıt olursa yapılan incelemelerin sonucu da o kadar doğru olacaktır. Ancak veri tabanı içerisindeki kayıt adedinin çok fazla sayıda olması durumda, bu veri tabanını kullanan canlı sistemlerin performansları düşecek dolayısı ile verilere erişim çok yavaşlayacaktır ki bu, canlı sistemlerde en istenmeyen durumdur.

2.VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ ve VERİ MADENCİLİĞİ

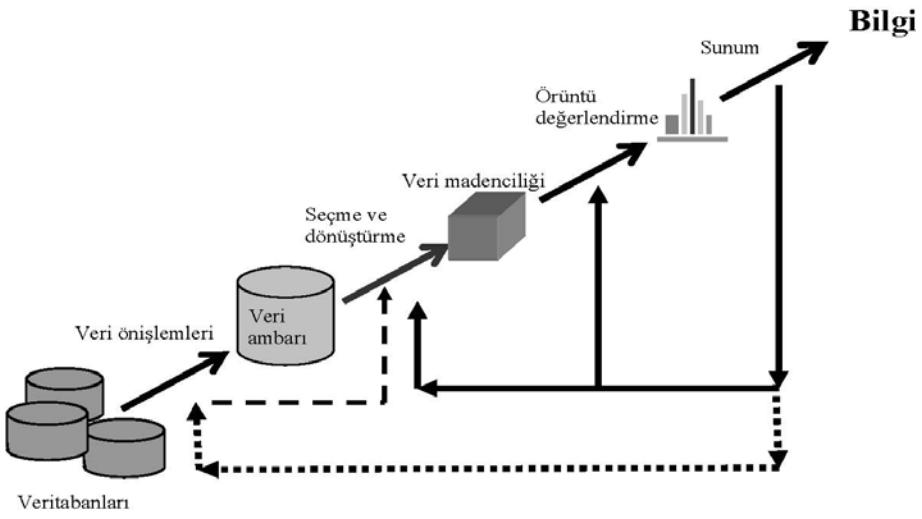
2.1. Veri Tabanlarında Bilgi Keşfi Süreci

Teknolojinin hızlı gelişimi ile birlikte saklanacak veri miktarlarının da çoğalması bu verilerden anlamlı bilgiler çıkarmak için çeşitli teoriler ve araçlara gereksinim vardır. VTBK sürecinin konusunu bu teoriler ve araçlar, oluşturmaktadır.

Veriden anlamlı örüntüler çıkarma sürecine literatürde, veri madenciliği, bilgi çıkarımı (knowledge extraction), bilgi keşfi, veri arkeolojisi ve veri örüntü işleme (data pattern processing) gibi isimler verilmektedir. İlk olarak 1989 yılında yapılan bir atölyede veri işleme sürecinde bilginin son ürün olduğunu vurgulamak için "veri tabanlarında bilgi keşfi" tanımlaması yapılmıştır.[2]

VTBK, veriden anlamlı ve yararlı bilginin çıkarıldığı süreç olarak tanımlanmakta ve VM, bu sürecin sadece bir kısmını oluşturmaktadır. Bu süreçte, büyük veri kümelerindeki düşük seviyedeki veriden yüksek seviyede bilgi çıkarımını sağlamak amaçlanmaktadır. VTBK, verinin saklanma, algoritmaların büyük veri kümelerine uygulanma ve sonuçların yorumlanma şekillerinin arandığı aşamalarıdır. Şekil 3'te bu süreç gösterilmiştir.

[2] Biçen, P. "Veri Madenciliği: Sınıflandırma ve Tahmin Yöntemlerini Kullanarak Bir Uygulama". Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.



Şekil 3. Veri Tabanlarında Bilgi Keşif Süreci

Kaynak: Han, J., Kamber, S.F.,"Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, (2001)

VTBK süreci beş aşamadan oluşmaktadır:

a) Veri Önişlemleri: Öncelikle veri temizleme denilen veriler içindeki gürültüler, tutarsızlık ve düzensizlikler giderilir. İkinci aşamada veri birleştirme işlemi uygulanır. Bu aşamada farklı kaynaklardan gelen verilerin tek bir veri ambarında toplanabilmesi için gerekli genelleme ve uyumluluk işlemleri yapılır.

b) Veri Seçme ve Dönüştürme (Data Selection): Bu aşamada, veri madenciliğinin sağlıklı yapılabilmesi için veriler üzerinde ön işlemler yapılır. Bunlar:

- Veri madenciliği konusu ile ilgili bilgi seçimi.
- Madencilik yapılacak veri türünün belirlenmesi.
- Veriler arasında hiyerarşik yapı ve genellemelerin belirlenmesi.
- Veri madenciliği sonunda bulunacak bilgi için yenilik ve ilginçlik ölçümü yöntemlerinin belirlenmesi.
- Veri madenciliği sonunda bulunacak veri için sunum ve görselleştirme araçlarının belirlenmesi.
- Önişlemlerin tamamını gerçekleyebilmek için bir veri madenciliği sorgulama dili kullanılır.

c) Veri Madenciliği: Anlamlı veri örüntüleri ortaya çıkarmak için çeşitli algoritmaların kullanıldığı aşamadır.

d) Örüntü Değerlendirme: İkinci aşamada belirlenen ilginçlik ölçüm yöntemleri kullanılarak veri madenciliği ile bulunan verilerin ilginçliği ve yararı tespit edilir.

e) Sunum: Çeşitli görselleştirme ve raporlaştırma araçları kullanılarak bulunmuş olan veriler ilgili kullanıcılara sunulur.

VTBK süreci sürekli tekrarlar, aşamalar arası atlamalar ve ileri geri hareketler içerebilmektedir. Günümüzde çoğunlukla veri madenciliği aşamasına odaklanılmakta, fakat diğer tüm aşamalar VTBK işleminin bütünlüğü açısından veri madenciliği kadar önemlidir.[3]

2.2. Veri Madenciliği

Veri madenciliği, veri tabanlarında saklanan çok çeşitli verilerden, daha önce keşfedilememiş bilgileri oraya çıkarmaktır. Veri madenciliği, kendi başına bir çözüm değil, çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli olan bilgileri sağlamaya yarayan bir araçtır.[4]

Veri Madenciliği, büyük miktarlardaki verinin içinden geleceğin doğru tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. Ayrıca, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan veri analizi tekniğidir.[5]

Veri madenciliği; veri ambarlarında tutulan çok çesitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar vermek ve eylem planını

[3] J Han,.ve S.F Kamber,., Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001.

[4] S.K Madria, W.K Bhowmick,.ve E.P NG, "Research Issues in Web Data Mining". In Proceedings of Data Warehousing and Knowledge Discovery", First International Conference. DaWak1999. ss 303-312.

[5] H Akpınar., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği". 2000. http://www.isletme.istanbul.edu.tr/dergi/nisan2000/1.htm. (Erişim Tarihi:30.08.2007)

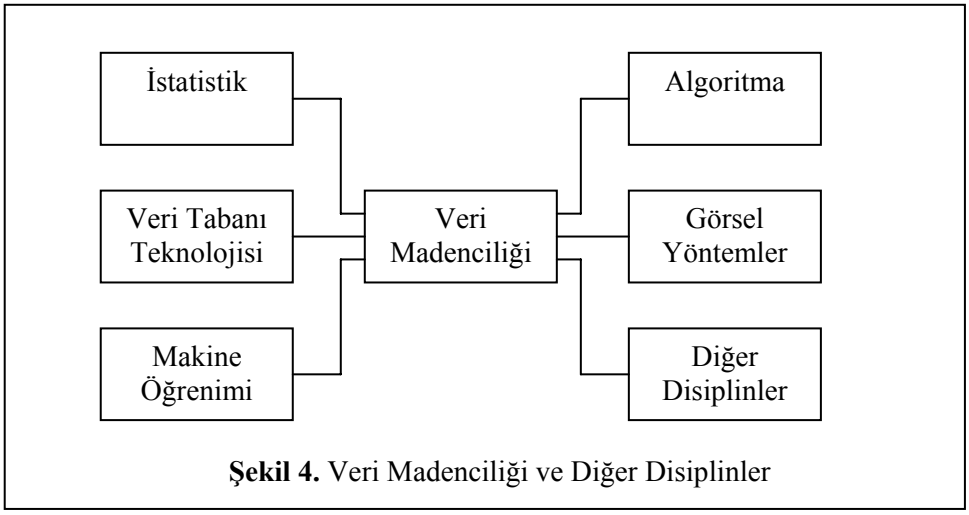
gerçekleştirmek için kullanma sürecidir. Veri Madenciliği istatistik biliminin teknolojiyle bütünleşmesi sonucu olusturulan bir araçtır.[6]

Karar Destek Sistemlerinde bir işlem sonucu oluşmuş tek bir veriye bakmak yerine bir grup müşteri bilgisini analiz ederek eğilimleri ortaya çıkarmak önemlidir. Çünkü Karar Destek Sistemlerinde tasarlanan sorgulamalar iç içe girmiş birden fazla değişken boyutuyla ilgilidir. Bu konuya şu örnekleri verebiliriz; "Son bir aydan beri bulunduğu şehirdeki büyük alışveriş merkezlerinin marketlerinde 100 milyon ve üzerinde hesap ödeyenlerin özellikleri nelerdir?" sorusuna cevap aramak, "Müşterilerin aldıkları mevcut ürünler dışında diğer hangi ürünleri almak isterler?" (Çapraz sorgulama), "En karlı müşterilerim kimlerdir ve bunların özellikleri nelerdir?", "Kurumumuzla çalışmayı bırakıp da rakibe yönelen müşterilerim kimlerdir ve (daha da önemlisi) bunların özellikleri nelerdir?" vb. Son sorunun cevabı sadece bırakıp giden müşterilerin kimler olduğunu raporlamak değil, bunları ayrıştıran özellikleri ortaya çıkaran bir model oluşturmak ve bu modeli mevcut müşterilere uygulayarak müşteriyi rakibe gitmeden önce belirleyip, gitmesini önleyecek tedbirler almaktır.

VM, gelecekteki kararlara yardımcı olmak için veritabanlarından eğilimler örüntüler ve ilişkiler bulur. Bu işlemin bizi iyi sonuçlara götüreceğinin garantisi yoktur. VM, sadece uzmanlara veriyi anlamada ve iyi karar vermede yardımcı olur. VM, araştırma ve çözümlemenin birden fazla disiplin kullanılarak yapılması yöntemidir. Makine öğrenimi, istatistik, veritabanı teknolojisi, uzman sistemler ve verilerin görüntülenmesi gibi yöntemlerin birlikte kullanıldığı, araştırma ve çözümlemelerin birden fazla disiplin kullanılarak yapılması yöntemidir.[7] Şekil 4'teki her bir disiplin bu veri keşfine kendi özünü katmaktadır.

[6] Şule Özmen, İş Hayatı Veri Madenciliği ile istatistik Uygulamalarını Yeniden Keşfediyor. 2002. www.mimoza.marmara.edu.tr/~sozmen/teblig.php (Erişim Tarihi:30.08.2007)

[7] J Maindonald, "Data Mining from a Statistical Perspective. Preprint". Australian Nat. Univ., Stat. Cons. Unit. http://www.maths.anu.edu.au/~johnm/dm/dmpaper.html. (Erişim Tarihi:"15.06.2007)



Şekil 4. Veri Madenciliği ve Diğer Disiplinler

Kaynak: Maindonald, J. Data Mining from a statistical perspective, Preprint. Australian Nat. Univ. Stat Cons Unit, (http://www.maths. anu.edu.au/~johnm/ dm/dmpaper.html

İstatistik biliminin doğusu bilgisayarın icadı ve gelişiminin öncesine dayanmaktadır. İstatistiksel yöntemlerin çoğu elle uygulanabilir. Ancak çok büyük veri kümeleri ile karşılaşıldığında ise istatistiksel yöntemleri elle uygulamak mümkün değildir. Bilgisayar, verilerin çözümlenmesi ve yorumlanması, çözümleyicilerin verilerle doğrudan ve uzun süre etkileşimde bulunmaması için gereklidir. Bundan dolayı bilgisayarlar, veriler ve çözümleyiciler arasında bir süzgeç görevi görür. Bu VM'de kullanılan algoritmalar için büyük önem taşır.[8]

İstatistiğin amacı kitle hakkında anlamlı bilgi elde etmek ve yorum yapmak, VM'nin amacı da anlamlı bilgi elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır.

VM net olarak tanımlanamayan bir alandır. Tanımı çoğunlukla tanımlayıcının arka planına, ilgi alanına ve görüşüne dayanmaktadır.

VM'nin mecazi olarak anlamı, veri dağları altındaki hazine veya altın külçelerini özel yazılımlar yardımıyla keşfetmek olarak ifade edilebilir. Tabii ki bu yeni bir düşünce değildir, EDA (Exploratory Data Analysis), çok değişkenli keşifsel çözümleme (Multivariate Exploratory Analysis) ve boyut indirgeme yöntemleri (bileşen (component), uygunluk (correspondence) ve kümeleme çözümlemeleri) gibi yöntemlerin kullanımıyla örüntülerin ve modellerin keşfine dayalı istatistiksel

[8] P Adriaans ve D Zantinge. Data Mining. Longman, Harlow: Addison Wesley. 1996..

metodoloji çok fazla gelişim göstermiştir. ASA (American Statistical Association)'nın eski başkanı J. Kettenring istatistiği, "veriden öğrenme bilimi" olarak tanımlamıştır.[9] VM'de diğer yöntemlere göre yeni olan;

- Çok fazla veri otomatik olarak biriktirilmekte ve bu verilerden kullanılabilir bilgi elde edilmektedir,
- Bilgisayar bilimlerinden gelen sinir ağları, karar ağaçları, mantık kuralları gibi çok çeşitli ve yeni yöntemler kullanılmaktadır,
- Hedef müşteri seçilerek ticari kazanç arttırılmaktadır,
- Kullanıcı dostu, çekici arayüzlü, profesyonel çözümleyiciler gibi karar verici olan, çok pahalı olmayan yeni yazılımlar mevcuttur.[10]

2.2.1. Literatürde Veri Madenciliği

Veri madenciliği 1990'lı yıllarda ortaya çıkmıştır. Bir online veritabanı olan Science Direct'te 1960'tan günümüze kadar bir literatür taraması yapıldığında veri madenciliği ile ilgili 1240'a yakın makale olduğu görülmektedir. Veri madenciliğinin özellikle 2000 yılından bu yana büyük bir gelişme gösterdiği göze çarpmaktadır. Aşağıda 2000 – 2006 tarihleri arasında veri madenciliği konusunda farklı alanlarda gerçekleştirilen uygulama örnekleri yer almaktadır.

Boginski ve Butensko, çapraz korelasyon ve kümeleme tekniklerini kullanarak hisse senedi piyasalarının yapısal özelliklerini ortaya koymuşlardır.[11]

Jiao, Zhang ve Helande, Kansai haritalama tekniği ile bir karar destek sistemi tasarlamışlardır.[12]

Jeng, Chen ve Liang, genetik algoritma ile biyolojik sistemlerin kinetik parametrelerini belirlemişlerdir.[13]

---

[9] G Saporta,. "Data Mining and Official Statistics". Quinta Conferenza Nationale di Statistica, ISTAT. Roma. 2000. ss. 15-17.

[10] a.g.e

[11] V Boginski., S Butenko. ve P Pardalos., "Mining market data: A network approach", Computers & Operations Research, 33 (11). 2006 s 3171-3184 .

[12] J Jiao, Y Zhang. ve M Helander, "Kansei mining system for affective design", Expert Systems with Applications, 30 (4). 2006. s 658-673

Facca ve Lanzi, web kütüklerinde tutulan verileri analiz etmek için makine öğrenme algoritmalarını kullanmışlardır.[14]

Hong, Park, Jon ve Rho, veri madenciliği tekniklerini kullanarak bir tedarikçi seçim modeli önermişlerdir.[15]

Huang, Chen ve Wu, kümeleme tekniklerini kullanarak dağıtım merkezleri için bir sipariş yönetim sistemi geliştirmişlerdir.[16]

Cervone, Kafatos ve Singh, veri madenciliği tekniklerini kullanarak bir deprem erken uyarı sistemi geliştirmişlerdir. [17]

Crespo ve Weber, bulanık kümelemeye dayalı veri madenciliği metodolojisi geliştirmişlerdir.[18]

Lee, Chiu, Chou ve Lu, sınıflama ve regresyon tekniklerini kullanarak bir kredi derecelendirme uygulaması gerçekleştirmişlerdir.[19]

Bellazi, Larizza ve Magni, veri madenciliği tekniklerini kullanarak hemodiyaliz servislerinin kalite ölçümünü gerçekleştirmişlerdir.[20]

Last ve Kandel, karar ağacı algoritmasını kullanarak yarı iletken endüstrisindeki bir fabrikada üretim planlama uygulaması gerçekleştirmişlerdir.[21]

---

[13] B. Jeng, J. Chen. ve T. Liang, "Applying data mining to learn system dynamics in a biological model", Expert Systems with Applications, 30 (1). 2006. s 50-58.

[14] F. Facca.ve P. Lanzi, "Mining interesting knowledge from weblogs: a survey", Data & Knowledge Engineering, 53 (3). 2005. s 225-241.

[15] G. Hong., S Park., D. Jang. ve H. Rho, "An effective supplier selection method for constructing a competitive supply relationship", Expert Systems with Applications, 28 (4) 2005. s 629-639.

[16] M. Chen, C. Huang., K. Chen. ve H. Wu., "Aggregation of orders in distribution centers using data mining", Expert Systems with Applications, 28 (3) .2005. s 453-460.

[17] G. Cervone., M. Kafatos., D. Napoletani. ve R. Singh, "An early warning system for coastal earthquakes", Advances in Space Research, 37 (4) .2006. s 636-642.

[18] F. Crespo. ve R. Weber., "A methodology for dynamic data mining based on fuzzy clustering", Fuzzy Sets and Systems, 150 (2) .2005. s 267-284.

[19] T. Lee., C. Chiu., Y. Chou. ve C. Lu., "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", Computational Statistics & Data Analysis, 50 (4) 2006. s 1113-1130.

[20] R. Bellazi., C. Larizza. ve P. Magni., "Temporal data mining for the quality assessment of hemodialysis services", Artificial Intelligence in Medicine, 34 (1): 2005. s 25-39.

[21] M. Last. ve A. Kandel., "Discovering useful and understandable patterns in manufacturing data", Robotics and Autonomous Systems, 49 (3). 2004. s 137-152.

Lian, Lai, Lin ve Yao, veri madenciliği tekniklerini montaj hattı uygulamalarında kullanmışlardır.[22]

Lin ve McClean, şirket iflaslarının tahminine yönelik veri madenciliği yaklaşımı geliştirmişlerdir.[23]

Caskey, genetik algoritma ve sinir ağları teknikleri ile bir fabrikadaki çalışma koşullarını ortaya koymuş ve bu koşulları iyileştirici işletme stratejileri önermiştir.[24]

Cox ve Lewis, çelik endüstrisindeki bir fabrikada yaptıkları uygulamada yapay sinir ağları yöntemini kullanarak ürünün istenilen kalite standartlarını sağlaması için gerekli girdi miktarını saptamışlardır.[25]

Sforna, veri madenciliği teknikleri ile bir elektrik şirketinin müşteri veritabanını analiz etmiştir.[26]

Kusiak, elektronik endüstrisinde yaptığı uygulamada imalat hatalarının tahmini için karar kurallarını kullanmıştır.[27]

2.2.2. Veri Madenciliği Uygulama Alanları

Veri madenciği her geçen gün yeni ve farklı alanlarda kullanılmaya başlanmıştır. Günümüzde yaygın olarak kullanıldığı alanlar birkaç başlık altında toplanabilir.

2.2.2.1. Pazarlama;

- • Müşterilerin satınalma örüntülerinin tesbitinde,
- • Kampanya ürünlerini belirlemede,

[22] J. Lian., M. Lai., Q. Lin., F. Yao., "Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis", Journal of Materials Processing Technology, 129 (1): 315-320 (2002)

[23] F. Lin ve S. Mcclean, "A data mining approach to the prediction of corporate failure", Knowledge-Based Systems, 14 (3). 2001. s 189-195.

[24] K. Caskey, "A manufacturing problem solving environment combining evaluation, search, and generalization methods", Computers in Industry, 44 (2) .2001 s 175-187

[25] I. Cox., R, Lewis., R, Ransing., H. Laszczewski, G. Berni, "Application of neural computing in basic oxygen steelmaking", Journal of Materials Processing Technology, 120 (1): 310-315 (2002).

[26] M. Sforna., "Data mining in a power company customer database", Electric Power Systems Research, 55 (3). 2000. s 201-209.

[27] S. Guha, R. Rastogi. ve K. Shim., "Rock: A robust clustering algorithm for categorical attributes", Information Systems, 25 (5). 2000 s 345-366

- • Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- • Pazar sepeti analizinde,
- • Çapraz satış analizleri,
- • Müşteri değerlemede,
- • Müşteri ilişkileri yönetiminde,
- • Çeşitli müşteri analizlerinde,
- • Satış tahminlerinde,

2.2.2.2. Banka ve Sigortacılık

- • Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- • Kredi kartı ve sigorta dolandırıcılıklarının tespitinde,
- • Kredi taleplerinin değerlendirilmesinde,
- • Kredi kartı harcamalarına göre müşteri profili belirlenmesinde,
- • Yeni Poliçe talep edecek müşterilerin tahmininde,
- • Risk yönetimi konusunda,
- • Riskli müşteri tipinin belirlenmesinde.

2.2.2.3. Borsa

- • Hisse senedi fiyat tahmininde,
- • Genel piyasa analizlerinde,
- • Alım-satım stratejilerinin uygunluğunda.
- • Hisse tespitlerinde.

2.2.2.4. Telekomünikasyon

- • Kalite ve iyileştirme analizlerinde,
- • Hatların yoğunluk tahminlerinde.

2.2.2.5. Sağlık ve İlaç

- • Test sonuçlarının tahmininde,
- • Ürün geliştirmede,
- • Tıbbi teşhiste,
- • Tedavi sürecinin belirlenmesinde,
- • Yeni ilaç türlerini keşfi ve sınıflandırılması.

2.2.2.6. Endüstri

- • Kalite kontrol analizlerinde
- • Lojistik uygulamalarda,
- • Üretim süreçlerinin uygunluğunda.

2.2.2.7. Bilim ve Mühendislik

- • Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesinde,
- • Yeni virüs türlerinin keşfi ve sınıflandırılmasında,
- • Gen haritasının analizi ve genetik hastalıkların tespitinde,
- • Kanserli hücrelerin tespitinde,
- • Gezegen yüzey şekillerinin, gezegen yerleşimlerinin ve yeni galaksilerin keşfinde.

2.2.3. Veri Madenciliğinde Karşılaşılan Sorunlar

Küçük veri kümelerinde ve ayıklanmış veri üzerinde hızlı ve doğru bir biçimde çalışan bir sistem çok büyük veritabanlarına uygulandığında sorun çıkabilir. Bir VM sistemi ayıklanmış veri üzerinde iyi çalışırken, aynı veriye gürültü eklendiğinde net olmayan sonuçlar oluşabilir.

Günümüzde VM sistemlerinin karşılaştığı problemler aşağıda incelenmiştir.

2.2.3.1. Veritabanı Boyutu

Veritabanı boyutunun çok büyük olması, VM sistemlerinin en önemli sorunlardan biridir. Küçük test verilerini ele alabilecek bir biçimde geliştirilmiş bir algoritmanın, çok büyük test verilerini kullanabilmesi azami dikkat gerektirmektedir. Dolayısıyla VM yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır ya da test verileri en aza indirilmelidir. Belirli bir niteliğin alan değerleri önceden sıra düzensel olarak kategorize edilir. Sonrasında ise, ilgili niteliğin değerleri aşağıdan yukarıya doğru seviye seviye güncellenir. Yani tekrarlı çokluklar çıkarılır. Oldukça sağlam bir test verisi kuramı kullanılarak çok büyük boyutlu veri öyle bir boyuta indirgenir ki, hem kaynak veri belirli bir güven aralığında temsil edilir hem de indirgenen veri kümesinin boyutu

kullanılan algoritma tarafından işlenebilir hale gelir. Son aşamada ise sürekli değerlerin belirli aralık değerlerine dönüştürülmesi ile tekrarlılık gösteren çokluklar ortadan kaldırılır.

2.2.3.2. Gürültülü Veri

Büyük veri tabanlarındaki değerlerin çoğu doğru olmayabilir. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Hatalı veri gerçek dünya veritabanlarında önemli problemler oluşturması VM yönteminin kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir.Veri kümesi gürültülü ise, sistem bozuk veriyi tanımalı ve ihmal etmelidir. Quinlan, 1986'da gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi deney yapmıştır. Deneysel sonuçlar etiketli öğrenmede etiket üzerindeki gürültü öğrenme algoritmasının performansını doğrudan etkileyerek düşmesine sebep olmuştur. Buna karşın eğitim kümesindeki nesneleri nitelikleri üzerindeki en çok %10'luk gürültü miktarı ayıklanabilmektedir .

2.2.3.3. Boş Değerler

Boş değer tanımı gereği kendisi de dahil olmak üzere hiçbir değere eşit olmayan değerdir. Çoklu veri üzerinde bir nitelik değeri boş ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Bu duruma ilişkisel veritabanında sıkça rastlanmaktadır. Bilinmeyen değer üzerinde de çalışmalar yapılmıştır. Boş değerli nitelikler veri kümesinde bulunuyorsa ya bu çokluklar tamamıyla ihmal edilmeli ya da bu çokluklarda niteliğe olası en yakın değer atanmalıdır.

2.2.3.4. Eksik Veri

Her nesnenin ayrıntılı bir biçimde tanımlandığı ve bu nesnelerin alabileceği değerler kümesinin belirtili olduğu durumlarda her bir nesnenin tanımı kesin ve yeterli olsaydı, sınıflama işlemi basitçe nesnelerin alt kümelerinden faydalanılarak yapılabilirdi. Bununla birlikte veriler kurum ihtiyaçları göz önünde bulundurularak düzenlenip toplandığında mevcut veri, gerçek hayatı yeterince yansıtmayabilir. Bu gibi

koşullarda bilgi keşfi modeli belirli bir güvenlik derecesinde tahmini kararlar alabilmelidir.[28]

2.2.3.5. Artık Veri

Kullanılan veri kümesi mevcut probleme uygun olmayan veya işe yaramayan nitelikler içerebilir. Artık nitelikleri elemek için geliştirilmiş algoritmalar, özellik seçimi olarak adlandırılır. Özellik seçimi, tümevarıma dayalı öğrenmede budama öncesi yapılan bir işlemdir. Ayrıca özellik seçimi, verilen bir ilişkinin içsel tanımını, dışsal tanımın taşıdığı bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklerle ifade edebilmektir. Özellik seçimi arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de artırır.

2.2.3.6. Dinamik Veri

Kurumsal çevrim içi veritabanlarının dinamik olması, bilgi keşfi metotları için önemli sakıncalar doğurmaktadır. Sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu bir veritabanı uygulaması olarak mevcut veritabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşer. Veritabanında bulunan verilerin kalıcı olduğu varsayılıp, çevrimdışı veri üzerinde bilgi keşif metodu çalıştırıldığında, değişen verinin elde edilen örüntülere yansıması gerektiğinden bu işlem, bilgi keşfi metodunun ürettiği örüntüleri zaman içinde değişen veriye göre sadece ilgili örüntüleri güncelleme yeteneğine sahip olmasını gerektirir.

2.2.3.7. Farklı Tipteki Verileri Ele Alma

Gerçek hayattaki uygulamalar makine öğrenmesinde olduğu gibi, yalnızca sembolik veya kategorik veri türleri değil aynı zamanda tamsayı, kesirli sayı, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam düz bir kütük veya ilişkisel veritabanlarında yer alan tablolar olabileceği gibi nesneye yönelik veritabanları, çoklu ortam veritabanları, coğrafik veritabanları vs. olabilir. Bununla birlikte veri çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini olanaksız hale getirmektedir. Bu yüzden veri tipine özgü, VM algoritmaları geliştirilmektedir.

[28] J. Han, S.F. Kamber, "Data Mining: Concepts and Techniques", MorganKaufmann Publishers, 2001

2.2.4. Veri Madenciliği Modelleri

VM'de kullanılan modeller, tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir.

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

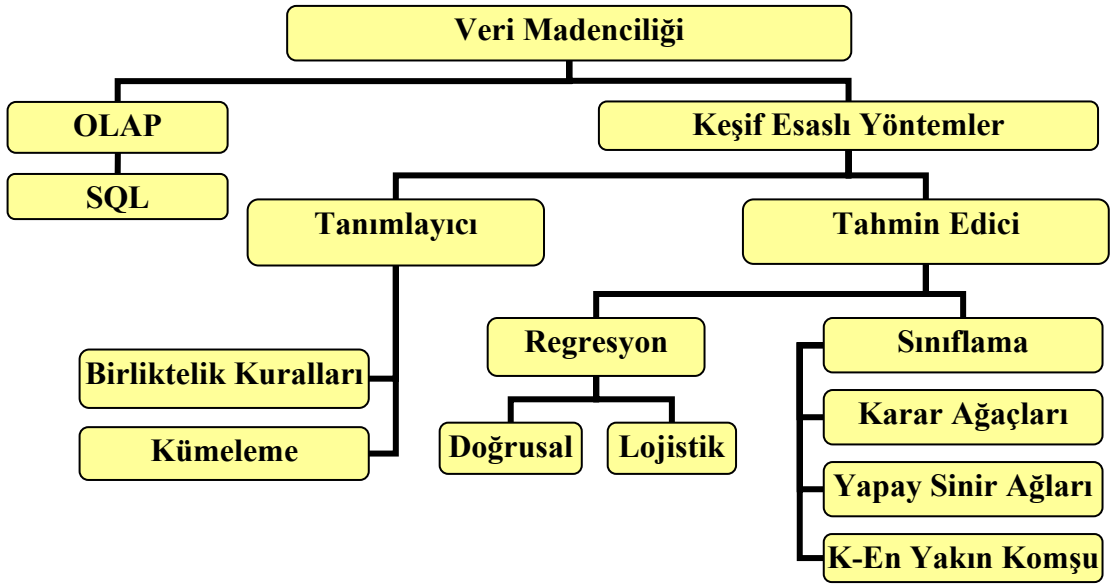
Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. X/Y aralığında geliri, evi ve arabası olan, ayrıca çocukları okul çağında olan aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir.[29]

VM modellerini gördükleri işlevlere göre,

- Sınıflama ve Regresyon,
- Kümeleme,
- Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir. Şekil 5.'de bu ilişkiler gösterilmiştir.

[29] H. Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi,2000c: 29, 1, s: 1-22.



Şekil 5. Veri Madenciliği Modelleri

2.2.4.1. Sınıflama ve Regresyon Analizi

İstenilen bir değişken bağımlı değişken ve diğerleri tahmin edici (bağımsız) değişkenler olarak adlandırılır. Amaç, girdi olarak tahmin edici değişkenlerin yer aldığı modelde, çıktı olarak bağımlı değişkenin değerinin bulunduğu anlamlı bir model kurmaktır. Bağımlı değişken sürekli yapıda değil ise problem sınıflama problemidir. Eğer bağımlı değişken sürekli yapıda ise problem regresyon problemi olarak adlandırılır.[30]

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve VM yöntemleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır.

Sınıflama ve regresyon modellerinde kullanılan başlıca yöntemler,

- Karar Ağaçları,
- Yapay Sinir Ağları,
- Genetik Algoritmalar,
- . K-En Yakın Komsu,

[30] V. Ganti, J. Gehrke, ve R. Ramakrishnan, "Mining Very Large Databases", IEEE Computer, 32, 8. 1999. s. 38-45.

- Bellek Tabanlı Yöntemler,
- Naïve-Bayes,
- Lojistik Regresyondur.[31]

2.2.4.2. Kümeleme Modelleri

Büyük bir veritabanı çok miktarda boyut, alan içerebilir ve çok karmaşık bir yapıya sahip olduğundan en iyi uygulanabilen VM yöntemleri bile bu veri yığını içerisinden anlamlı sonuçlar üretemeyebilir. Çok karmaşık ve büyük sorunları çözmekte izlenilen yöntem genellikle büyük sorunu daha küçük ve tek başına daha rahat çözülebilecek alt sorunlara bölmek ve her bir alt sorunu çözdükten sonra çözümleri birleştirerek sonuca gitmek şeklindedir. Ancak bazı durumlarda veriler öyle dağılmışlardır ki nereden bölüneceğini ve hangi şekilde alt gruplara ayrılabileceğini kestirmek mümkün değildir. Bu yüzden otomatik küme bulma yöntemleri geliştirilmiştir.

Kümeleme işlemi, heterojen yapıya sahip bir kitleyi daha homojen birkaç alt gruba ya da kümeye bölme işlemidir. Sınıflama ile kümelemeyi birbirinden ayıran en önemli fark, kümeleme işleminin sınıflama işleminde olduğu gibi önceden belirlenmiş bir takım sınıflara göre bölme yapmamasıdır. Sınıflamada her bir veri, önceden sınıflandırılmış bir takım sınıflar üzerinde yapılan bir eğitim neticesinde ortaya çıkan bir modele göre önceden belirlenmiş olan bir sınıfa atanmaktadır.

Kümeleme işleminde ise önceden tanımlanmış sınıflar ya da örnek sınıflar bulunmamaktadır. Verilerin kümelenmesi işlemi, verilerin birbirlerine olan benzerliklerine göre yapılmaktadır. Oluşan sınıfların hangi anlamları taşıdığının belirlenmesi tamamen çözümlemeyi yapan kişiye kalmıştır.

Kümeleme işlemi çoğunlukla bir başka VM uygulaması için bir ilk işlem olarak kullanılır.[32]

[31] H Akpınar,., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, 2000c: 29, 1, s: 1-22.
[32] A.C. Tantuğ, "Veri Madenciliği ve Demetleme". Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.

Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.

Literatürde birçok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve çalışmanın amacına bağlıdır.

Genel olarak başlıca kümeleme yöntemleri şu şekilde sınıflandırılabilir;

- Bölme Yöntemleri (Partitioning Methods)
- Hiyerarşik Yöntemler (Hierarchical Methods)
- Yoğunluk Tabanlı Yöntemler (Density-based Methods)
- Izgara Tabanlı Yöntemler (Grid-based Methods)
- Model Tabanlı Yöntemler (Model-based Methods)[33]

2.2.4.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Bilgisayar teknolojisi ve veritabanlarının kullanımı son yıllarda yaygınlaşmıştır. Bununla birlikte birçok kuruluş rutin olarak çok büyük miktarda veri toplamakta ve depolamaktadır. Verinin boyutları ve karmaşıklığı verinin anlaşılmasını ve çözümlenmesini zor bir hale getirdiğinden verideki bilgiyi keşfetmek için hem fırsat hem de yoğun uğraş gerekmektedir.[34]

Birliktelik kuralları ve ardışık zamanlı örüntüleri birbirinden ayıran özellik zaman kavramının uygulamada olmasıdır. Belli bir dönem boyunca nesneler arasındaki birlikteliklerin incelenmesi "ardışık zamanlı örüntü çözümlemesi" olarak da isimlendirilir.[35]

Birliktelik kuralları; ticaret, mühendislik, fen ve sağlık sektörlerinin içinde bulunduğu birçok alanda uygulanmaktadır. Birliktelik kuralları, VM araştırmalarında çok büyük yatırımlar yapılan, VM'nin özel bir uygulama alanıdır. Birliktelik kuralları

[33] S. Özekeş, "Veri Madenciliği Uygulaması". Yüksek Lisans Tezi. Marmara Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.
[34] J. Rushing, Technology Assessment Paper. 1997. http://www.cs.uah.edu/~thinke/CS 687/Fall 97/Tech/Rushing.html. (Erişim Tarihi:03.07.2007)
[35] M. Goebel, ve L. Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", ACM SIGKDD Explorations Newsletter, 1, 1. 1999. s. 20-33.

aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallardır. Birliktelik kurallarının bulunması ile pazar sepeti çözümlemesi yapılmaktadır. Pazar sepeti çözümlemesinde, nesneler müşteriler tarafından satın alınan ürünlerdir ve bir işlem (kayıt) ise birçok nesneyi içinde bulunduran tek bir satın almadır. Pazar sepeti çözümlemesinde sıklıkla beraber alınan nesneler üzerine çalışılır.[36] Bulunan kurallar ile nesnelerin birbiri ile nasıl ilişkili olduğu bilgisine ulaşılır.

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesi, müşteriye daha fazla ürün satma yollarından birisidir.[37]

Birliktelik kurallarının bulunmasında birçok yöntem vardır. Büyük veritabanlarında birliktelik kuralları bulmak için algoritma geliştirmek çok zor değildir, buradaki zorluk bu tür algoritmaların çok küçük değerli diğer birçok birliktelik kuralını da meydana çıkarmasıdır. Bulabileceğimiz olası birliktelik kuralları sayısı sonsuzdur.

Birliktelik kurallarıyla ilgili problem, birliktelik kurallarını bulmada bir eşik değeri bulmaktır. Önemsiz gürültüden değerli bilgiyi ayırabilmek ve bu eşik değerini bulabilmek çok zordur. Bu yüzden ilginç birliktelik kurallarından ilginç olmayanları ayırt edebilmek için bazı ölçütlerin belirlenmesi gereklidir. Bu ölçütler destek ve güven değerleridir.[38]

Birliktelik kuralı madenciliğinin amacı, kullanıcı tarafından belirlenen minimum destek ve güven değerlerini sağlayan kuralların bulunmasıdır.[39]

Anlamlılığı destek ve güven değerleri ile ölçülen birliktelik kuralları, "X nesnesini alan bir müşterinin muhtemelen Y nesnesini de alması" tipindeki kuralların tanımlanmasını amaçlamaktadır.[40]

---

[36] J. Rushing, Technology Assessment Paper. 1997. http://www.cs.uah.edu/~thinke/CS 687/Fall 97/Tech/Rushing.html. (Erişim Tarihi:03.07.2007)
[37] J. Han, ve Kamber, S.F., Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001
[38] P. Adriaans,.ve D. Zantinge, Data Mining. Longman, Harlow: Addison Wesley. 1996.
[39] Rushing,a.g.e.
[40] S. Brin, R. Motwani, ve C.Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations, Proceedings of the 1997 ACM SIGMOD". International Conference on Management of Data. New York, NY, USA 1997. ss. 265-276

2.2.5. Veri Madenciliği Teknikleri

2.2.5.1. Birliktelik Kuralı ve Pazar Sepeti Analizi

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder. Sepet analizi tekniği, adında da anlaşılacağı üzere, çok basit olarak hangi ürünlerin hangi ürünlerle satıldığını, hangi ürünlerin promosyona girmesi gerektiğini ortaya çıkarır ve ürün kombinasyonlarının birbirleri ile yakınlıklarını belirlemek için sepet verilerinin analizini yapar. Her ne kadar yoğunlukla satış alanında kullanılsa da, sepet analizi tekniği bunun dışında birçok alanda daha kullanılmaktadır:

- Kredi kartları ile yapılan alışverişlerin işlenmesi ve müşterilerin yapacakları potansiyel harcama kalemlerinin bulunması
- Cep telefonundaki opsiyonel hizmetlerin (GPRS, WAP, Telesekreter vb.) müşteriler tarafından tercih edilmelerine göre kârı arttırmak için hangi ürünlerin birlikte kampanyaya girmesi gerektiğini belirlemek
- Sigortacılıkta ortaya çıkan değişik tarzdaki suçların dolandırıcılık olup olmadığının belirlenmesi ve yapılacak soruşturmaya ışık tutulması
- Hastaların sağlık kayıtlarından, önerilen çeşitli tedavi birleşimlerinden doğan yan etkilerin görülmesi

Sepet analizi çoğunlukla ticari anlam taşıyan verilerin var olduğu ancak bu veri üzerinde hangi örüntülerin aranılacağının bilinmediği durumlarda bir başlangıç noktası olarak kullanılır. Bu veri içerisindeki bazı kalıplar sayesinde kazancı arttırmak üzere bir takım aksiyonlara gidilebilir.

Örneğin yurtdışında yapılan sepet analizi tekniklerine göre perşembe günleri bira ve çocuk bezi satışlarının çok fazla sayıda olduğu görülmüştür.[41] Bunun temel nedeni olarak ise evli çiftlerin hafta sonunu evde geçirmek istemeleri ve bu süre içerisinde de rahatlarını bozmamak için gerekli olması muhtemel bira ve çocuk bezini hafta sonu gelmeden almak istemeleri olarak gösterilmiştir.

[41] G. Linoff, ve M.J.A. Berry, Data Mining techniques For Marketing Sales and Customer Relationship Management, , New York: Wiley Publishıng. 2004.

Birliktelik kuralı keşfi ürün kümeleri arasındaki ilişkileri ve yakınlıkları bulur. Her bir işlem bir ürün seti olarak adlandırılır.

Birliktelik kuralı önce gelen (antecedent) ve sonra gelen (consequent) olarak adlandırılan ürün kümelerinden oluşur. Sonra gelen genellikle bir üründen oluşur. Kural tipik olarak önce gelenden sonra gelene doğru yönelen bir ok şeklinde gösterilir. Birliktelik kuralı önce gelen ürün kümesi ile sonra gelen ürün seti arasındaki yakınlığı gösterir. Birliktelik kuralına ilişkileri tanımlayan frekansa dayalı istatistik tarafından eşlik edilir. Bu ilişkileri tanımlamak için kullanılan iki istatistik sayısal kavramı destek ve güven değeridir.[42]

Bu kavramları tanımlamak için bazı sayısal terimlerin tanımlanması gereklidir. İşlem veri tabanı D, D'deki işlem sayısı N olsun. Her bir D_i ürün setidir. Destek(X) de ürün seti X'i içeren işlemlerin oranı olsun. Bu durumda;

$$Destek\ (X) = \{I \mid I \in D \wedge I \supseteq X\} \mid /N$$

Burada *I* bir ürün setidir.

Birliktelik kuralının destek değeri hem önce hem de sonra geleni içeren işlemlerin oranıdır. Yine birliktelik kuralının güven değeri sonra geleni de kapsayan önce gelen işlemlerinin oranıdır. Bir birliktelik için;

$$Destek(A \rightarrow B) = destek(A \cup B)$$

$$Güvenilirlik\ (A \rightarrow B) = destek(A \cup B) / destek(A).$$

Eğer destek değeri yeteri kadar yüksekse, güven değeri de, önce geleni ihtiva eden, sonra geleni de içerecek olan verilen herhangi bir gelecek işlemi olasılığının mantıksal tahminidir.

Örneğin bir A ürününü satın alan müşteriler aynı zamanda B ürününü da satın alı-

[42] R. Agrawal, T. Imielinski, ve A. Swami, "Mining Association Rules Between Sets Of İtems In Large Databases". *ACM SIGMOD* Conference on Management of Data. Washington, DC: ACM Press. 1993 ss. 207-216.

yorlarsa, bu durum şu birliktelik kuralı ile gösterilir[43]:

A => B [destek = %5, güven = %70]

Buradaki destek ve güven ifadeleri, kuralın ilginçlik ölçüleridir ve sırasıyla, keşfedilen kuralın kullanışlılığını ve doğruluğunu gösterirler. Yukarıdaki birliktelik Kuralı için 5% oranındaki bir destek değeri, analiz edilen tüm alışverişlerden %5'sinde A ile B ürünlerinin birlikte satıldığını belirtir. %70 oranındaki güven değeri ise A ürününü satın alan müşterilerinin %70'ının aynı alışverişte B ürününü de satın aldığını ortaya koyar.[44] Kullanıcı tarafından minimum destek ve minimum güven eşik değeri belirlenir ve bu değerleri aşan birliktelik kuralları dikkate alınır.

2.2.5.1.1. Apriori Algoritması

Büyük veri tabanları için birliktelik kuralları bulunurken, aşağıdaki iki işlem basamağı izlenir:[45]

a) Sık tekrarlanan öğeler bulunur: Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.

b) Sık tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur: Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.

Sık tekrarlanan öğeleri bulmak için kullanılan en temel yöntem Apriori Algoritmasıdır. Apriori Algoritması en az destek eşik değeri tanımlayarak dikkate alınması gereken ürün küme sayılarını azaltmayı sağlar.

Tablo 1'de tekrarlanan ürün setlerini oluşturan Apriori Algoritması görülmektedir. Üçüncü satır yeni adaylar sağlar. Bu adım ürünlerin daha önceki sıralamalarına dayanır. Bir ürün seti eğer alt kümeleri de sık tekrarlan ürün seti ise sık tekrarlanan ürün seti sayılır. Hesaplama maliyeti yüksek olan potansiyel aday olan tüm alt kümelerin sıklıklarının test edilmesi yerine aday yaratma işlemi, iki alt kümesi de sık tekrarlanan adaylar yaratır. İki alt küme yeni adayın en yüksek değerli iki elemanını iptal ederek

[43] M. J. Zaki, "Parallel and Distributed Association Mining: A Survey, IEEE Concurrency, special issue on Parallel Mechanisms for Data Mining", IEEE Vol. 7, No. 5. December 1999. s 14-25
[44] **a.g.e.**
[45] **a.g.e.**

oluşturulur. k-1 boyutlarındaki sık tekrarlanan ürün setlerinden bu tür adaylar yaratmak için çok etkin yöntemler geliştirilebilir.[46]

Tablo 1. Apriori Algoritması

1.	L_1 ={frequent one-item-item sets}
2.	**for** k =2; L_{k-1} =Ø; k++ **do begin**
3.	C_k ={{x_1, $x_{2,}$.., x_{k-2}, x_{k-1}, x_k }|{x_1, x_2 ,..., $x_{k-2,}$ x_{k-1}} $\in L_{k-1}$ $\wedge$ {x_1, $x_{2,}$..., x_{k-2} , x_k } $\in L_{k-1}$}
4.	**for all** transactions $t \in D$ **do begin**
5.	**for all** candidates $c \in C_k \wedge c \subseteq t$ **do**
6.	c.count++;
7.	**end**
8.	L_k ={$c \in C_k$ |c.count $\geq$minsup}
9.	**end**
10.	return U_k L_k ;
procedure apriori gen(Lk_1:frequent (k-1)-itemsets; min sup: minimum support)	
1.	**for each** itemset $l_1 \in L_{k-1}$
2.	**for each** itemset l2 2 Lk_1
3.	**if** (l_1[1] = l_2[1]) ^ (l_1[2] = l_2[2]) ^ ..^ (l_1[k -2] = l_2[k -2]) ^ (l_1[k -1] < l_2[k-1]) ⋈**then** {c = l_1 l_2; // join step: generate candidates
4.	**if** has infrequent subset(c;Lk_1) **then**
5.	delete c; // prune step: remove unfruitful candidate
6.	**else** add c to Ck; }
7.	**return** C_k;
procedure has infrequent subset(c: candidate k-itemset; Lk_1: frequent (k_1)-itemsets); // use prior knowledge	
1.	**for each** (k -1)-subset s **of** c
2.	**if** s$\notin L_{k-1}$ **then**
3.	**return** TRUE;
4.	**return** FALSE;

Kaynak: Data Mining Concepts and Techniques, Han, J.-Kamber, M., Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA, 2000.

[46] R. Agrawal, ve R. Srikant, "Fast Algorithms For Mining Association Rules". In Proceedings of the 20th International Conference on Very Large Databases.. San Francisco: Morgan Kaufmann. 1994 ss. 487–499.

Apriori algoritması aşağıdaki örnekle açıklanacaktır.

Tablo 2. Market Satış Noktasındaki Örnek Satış Kayıtları Veritabanı

Müşteri No	Satın Alınan Ürünler
M1	P1, P2, P6
M2	P2, P3, P5
M3	P2, P6
M4	P2, P4
M5	P3, P4, P5, P6
M6	P3, P5,P6
M7	P2, P5
M8	P2, P3,P4, P5, P6
M9	P1, P6
M10	P2,P3,P5,P6

Tablo 2'de bir marketten yapılan alışverişlerin bilgilerini içeren *D* veritabanı görülmektedir. Bu veritabanında yapılan alışverişlerin numaraları Müşteri No sütununda görülmektedir. Her alışverişte satın alınan ürünler de Satın alınan Ürünler sütununda görülmektedir. Apriori algoritmasında takip edilen basamaklar Şekil 6'da gösterilmektedir.[47]

1. Algoritmanın ilk adımında, her ürün tek başına bulunduğu C_1 kümesinin elemanıdır. Algoritma, her ürünün sayısını bulmak için tüm alışverişleri tarar ve elde edilen sonuçlar Sekil 6'da Destek Sayısı sütununda görülmektedir. Tablo 2.'de görülebileceği gibi D' de P1 ürününden 2 adet, P2 ürününden 7 adet, P3 ürününden 5 adet, P4 ürününden 3 adet, P5 ürününden 6 adet ve P6 ürününden de 7 adet satıldığı görülmektedir.

2. Minimum alışveriş destek sayısının 3 olduğu varsayılırsa, tek başlarına sık tekrarlanan ürünler L_1 kümesinde görülmektedir. C_1 kümesindeki P1 ürünü hariç tüm ürünlerin destek sayısı, minimum destek eşik değeri olan 3'ten fazla olduğu için C_1 tüm

[47] J. Han, ve S.F. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001.

P1 haricindeki ürünler sık tekrarlanan ürün olarak değerlendirilir ve L_1 kümesine aktarılır.

3-. Hangi ürünlerin ikili olarak sık tekrarlandığını belirlemek için L_1 kümesindeki ürünlerin ikili kombinasyonları bulunarak C_2 kümesi oluşturulur.

4. C_2 kümesindeki ürünlerin destek sayılarını bulmak amacıyla D taranır ve bulunan değerler destek sayısı sütununda belirtilir.

5. C_2 kümesindeki ürünlerden minimum destek eşik değerini aşan ürünler L_2 kümesine aktarılır.

6. Hangi ürünlerin üçlü olarak sık tekrarlandığını belirlemek için L_2 kümesindeki ürünlerin üçlü kombinasyonları bulunarak C_3 kümesi oluşturulur. Bu durumda C_3 = {{P2,P3,P5}, {P2,P3,P6}, {P2,P5,P6}, {P3,P5,P6}} olması beklenir. Ancak Apriori algoritmasına göre, sık tekrarlanan öğelerin alt kümeleri de sık tekrarlanan öğe olması gerekmektedir. Buna göre yukarıdaki C_3 kümesindeki elemanlar sık tekrarlanan olmadığı için, yeni C_3 kümesi C_3 = {{P2,P3,P5}, {P3,P5,P6}} olur.

7. C_3 kümesindeki ürünlerin destek sayılarını bulmak amacıyla D taranır ve bulunan değerler destek sayısı sütununda belirtilir.

8. C_3 kümesindeki ürünlerden minimum destek eşik değerini aşan ürünler L_3 kümesine aktarılır.

9. Hangi ürünlerin dörtlü olarak sık tekrarlandığını belirlemek için L_3 kümesindeki ürünlerin dörtlü tek kombinasyonu {P2, P3, P5, P6} olarak belirlenir. Ancak bu kümenin kendisi ve alt kümelerinin tamamı sık tekrarlanan öğe olmadığı için C_4 kümesi boş küme olur ve Apriori tüm sık tekrarlanan öğeleri bularak sonlanmış olur.

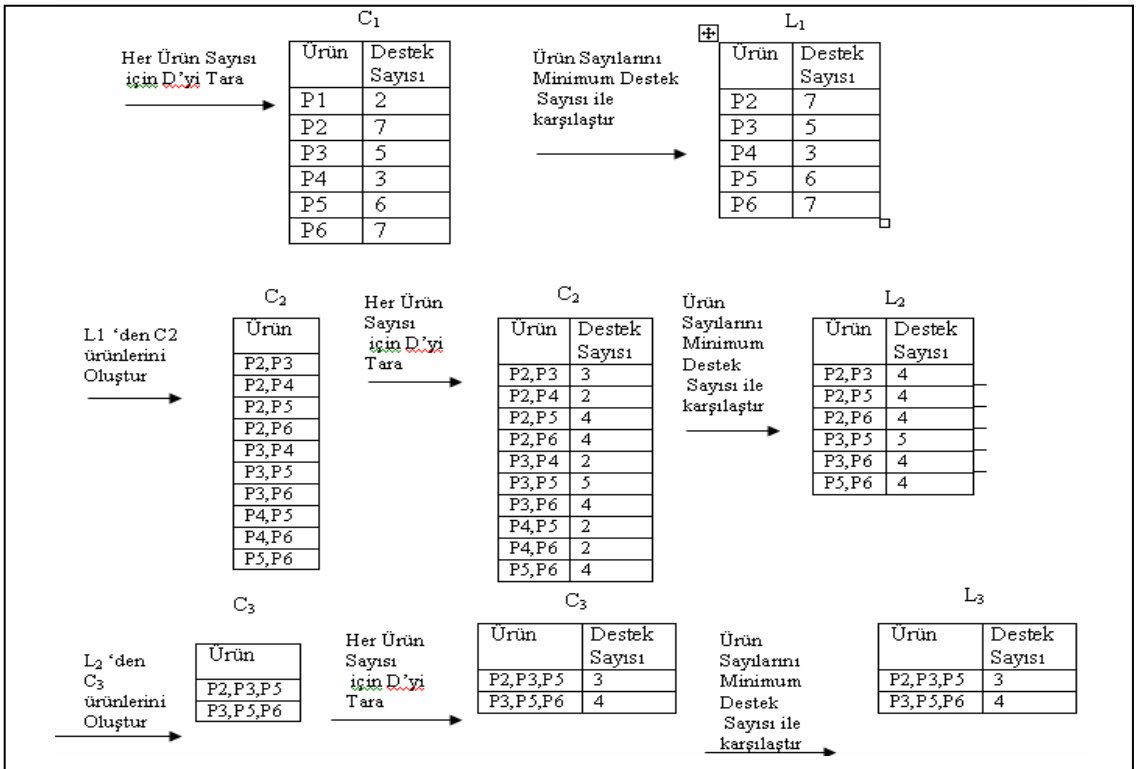
Sık tekrarlanan öğeleri bulduktan sonra, sıra birliktelik kurallarını oluşturmaya gelir. Örneğin sık tekrarlanan bir öğe için { P2, P3, P5}, boş olmayan tüm alt kümeler

{P2, P3}, {P2, P5}, {P3, P5}, {P2}, {P3},{P5} alt kümeleridir. Bu durumda Tablo 2'deki veritabanına bakarak şu birliktelik kuralları çıkartılabilir.[48]

Tablo 3. Birliktelik Kuralları

1	P2 ∧ P3 ⇒ P5	Güven = 3 / 3 = % 100
2	P2 ∧ P5 ⇒ P3	Güven = 3 / 4 = % 75
3	P3 ∧ P5 ⇒ P2	Güven = 3 / 5 = % 60
4	P2 ⇒ P3 ∧ P5	Güven = 3 / 7 = % 43
5	P3 ⇒ P2 ∧ P5	Güven = 3 / 5 = % 60
6	P5 ⇒ P2 ∧ P3	Güven = 3 / 6 = % 50

Eğer minimum güven eşik değeri %65 olarak belirlenmişse, birinci ve ikinci kurallar dikkate alınır çünkü diğer kurallar eşik değerini aşamamış olurlar.[49]



Şekil 6. Apriori Algoritmasının Uygulaması

Kaynak: Agrawal, R.,ve Srikant, R. Fast Algorithms For Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases. 1994. San Francisco: Morgan Kaufmann. ss. 487–499.

[48] J. Han,.ve S.F. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001.

[49] a.g.e.

2.2.5.1.2. Sepet Analizi

Sepet analizi, geleceğe dönük tahminlerde pek başarılı değildir. Sepet analizinin temeldeki mantığında yer alan teknikler istatistik ve olasılıktan gelmektedir.

Ticari anlam taşıyan veriler üzerinde belirli bir ürün kombinasyonunun kaç defa geçtiğinin bulunması işlemi tek başına yeterli değildir. Bu kombinasyonu işletme açısından anlamlı hale getirecek olan kilit nokta, bu kombinasyonu oluşturan kuralı bulmaktır.

Kural tanımı iki kısımdan oluşur: koşul kısmı ve sonuç kısmı;[50]

Eğer *KOŞUL* doğru ise, *SONUÇ* da doğrudur.

Pratikte işletme açısından eyleme dönüştürülebilecek kuralların sadece bir adet sonuç kısmı vardır. Yani;

Eğer çocuk bezi ve perşembe ise bira satılabilir kuralı

Eğer perşembe ise çocuk bezi ve bira kuralından daha çok faydalıdır. Çünkü sadece günün perşembe olmasından dolayı birisine çocuk bezi ve bira satmaya çalışmak anlamsız olacaktır. Aksine eğer günlerden perşembe ise ve müşteri çocuk bezi almış ise bu müşterinin bira alma olasılığı çok yüksek demektir. Bunun için işletme bira satışlarını arttırmak için perşembe günleri çocuk bezi ürünleri ile biraları beraber satmak üzere promosyona girebilir. Dolayısı ile ikinci kural işletme açısından çok daha anlamlı ve eyleme dönüştürülebilecek bir yapıya sahiptir.

Sepet analizi tekniğinin çalışma mantığına örnek olarak bir marketteki satış noktasında kaydedilen aşağıdaki alışveriş bilgilerini ele alalım.

[50] J. Han,.ve S.F. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers. 2001.

Tablo 4. Örnek Market Satış Bilgileri

Müşteri No	Satın Alınan Ürünler
1	turunç, dereotu, domates
2	enginar, ekmek
3	ekmek
4	armut, havuç, domates, patates, ekmek
5	armut, portakal, dereotu, domates, ekmek
6	şeftali, portakal, enginar, patates
7	fasülye, dereotu, domates
8	portakal, dereotu, havuç, domates, ekmek
9	armut, muz, turunç, havuç, domates, soğan, ekmek
10	armut, patates

Tablo 4 görsel olarak incelendiğinde dereotunu içeren dört alım işleminin domatesi de içerdiği ve altı alım işleminin içinde yer alan domatesin dört alım işleminde dereotuyla birlikte olduğu görülmektedir. Birliktelik kuralı bu tür benzerlikleri tanımlamaya ve karakterize etmeye çalışır. Her bir alım işlemi bir ürün seti olarak adlandırılır. Hatta alım işleminde görünmeyen farklı kombinasyonlar da ürün setidir.

Sık tekrarlanan ürün seti stratejisine göre belirlenen minimum destek ve güven değerleri araştırılan veri yığınının değerlendirilen seyrek veri için yönetilebilir biçime dönüşmesini sağlar. Sık tekrarlanan ürün seti üretimi minimum destek değeri=0.25 için Tablo 5'te gösterilmiştir.

Tablo 5. Ürün Tekrar Sayısı

armut [Destek=0.4]
havuç [Destek=0.3]
ekmek [Destek=0.6]
dereotu [Destek=0.4]
portakal [Destek=0.3]
patates [Destek=0.3]

Tablo5. devam

domates [Destek=0.6]
armut, ekmek [Destek=0.3]
armut, domates [Destek=0.3]
havuç, ekmek [Destek=0.3]
havuç, domates [Destek=0.3]
ekmek, domates [Destek=0.4]
dereotu, domates [Destek=0.4]
armut, ekmek, domates [Destek=0.3]
havuç, ekmek, domates [Destek=0.3]

Bu tablodan da minimum güven değeri = 0,5 için, Tablo 6'daki 41 birliktelik kuralı çıkarılabilir. Her bir kuralın önce geleni oktan önce ve sonuç değeri oktan sonra gösterilmiştir.

Tablo 6. Birliktelik Kuralları

→ ekmek [Destek=0.6, Güven=0.60]
→ domates [Destek=0.6, Güven=0.60]
armut → [Destek=0.4, Güven=1.00]
armut → ekmek [Destek=0.3, Güven=0.75]
armut → ekmek, domates [Destek=0.3, Güven=0.75]
armut → domates [Destek=0.3, Güven=0.75]
armut, ekmek → [Destek=0.3, Güven=1.00]
armut, ekmek, domates → [Destek=0.3, Güven=1.00]
armut, domates → [Destek=0.3, Güven=1.00]
armut, ekmek → domates [Destek=0.3, Güven=1.00]
armut, domates → ekmek [Destek=0.3, Güven=1.00]
havuç → [Destek=0.3, Güven=1.00]
havuç → ekmek [Destek=0.3, Güven=1.00]
havuç → domates, ekmek [Destek=0.3, Güven=1.00]
havuç → domates [Destek=0.3, Güven=1.00]
havuç, ekmek, domates → [Destek=0.3, Güven=1.00]
havuç, ekmek → [Destek=0.3, Güven=1.00]

Tablo 6. devam

havuç, domates → [Destek=0.3, Güven=1.00]
havuç, ekmek → domates [Destek=0.3, Güven=1.00]
ekmek → [Destek=0.6, Güven=1.00]
ekmek → armut, domates [Destek=0.3, Güven=0.50]
ekmek → armut [Destek=0.3, Güven=0.50]
ekmek → havuç [Destek=0.3, Güven=0.50]
ekmek → domates, havuç [Destek=0.3, Güven=0.50]
ekmek → domates [Destek=0.4, Güven=0.67]
ekmek, domates → [Destek=0.4, Güven=1.00]
dereotu → [Destek=0.4, Güven=1.00]
dereotu → domates [Destek=0.4, Güven=1.00]
dereotu, domates → [Destek=0.4, Güven=1.00]
portakal → [Destek=0.3, Güven=1.00]
patates → [Destek=0.3, Güven=1.00]
domates → [Destek=0.6, Güven=1.00]
domates → armut, ekmek [Destek=0.3, Güven=0.50]
domates → armut [Destek=0.3, Güven=0.50]
domates → havuç, ekmek [Destek=0.3, Güven=0.50]
domates → havuç [Destek=0.3, Güven=0.50]
domates → ekmek [Destek=0.4, Güven=0.67]
domates → dereotu [Destek=0.4, Güven=0.67]
domates, ekmek → armut [Destek=0.3, Güven=0.75]
domates, ekmek → havuç [Destek=0.3, Güven=0.75]
domates, havuç → ekmek [Destek=0.3, Güven=1.00]
havuç, domates → [Destek=0.3, Güven=1.00]
havuç, ekmek → domates [Destek=0.3, Güven=1.00]

Başlangıçtaki kullanıcının destek ve güven değerlerine göre birliktelik kuralları genellikle binlerce kuralın ortaya çıkmasına sebep olur. Bu durumda istenen, kural sayısının mümkün olduğunca azaltılması ve en ilginç kuralların tanımlanmasıdır. Genellikle ilginçlik kuralı kuralın destek değeri ile önce gelenin destek değeri ve sonucun destek değeri çarpımının farkını ilişkilendirir. Önce gelen ve sonuç birbirinden

bağımsız ise kuralın destek değeri yaklaşık olarak önce gelenin ve sonucun destek değerleri çarpımına eşittir. Önce gelen ve sonuç bağımsız ise kural güven değerinin yüksekliğiyle ilgili değildir.[51] İlginçliğin objektif ve sübjektif ölçüleri mevcuttur ve bunların ayırt edilmesi faydalıdır.[52]

Objektif ölçüler özel uygulamaya bağımsız olarak uygulanan güven ve destek değerleri gibi ölçüleri ihtiva eder. Sübjektif ölçüler ise özel içerikte kullanıcının ihtiyaç duyduğu özel bilgi ihtiyaçlarıyla kuralın ilginçliğini bağdaştırır. En yaygın objektif ölçü de yükseltmedir(Lift).

Yükseltme(A →C) =Güven (A→C) / Destek (C)

Yükseltme değerinin 1'den yüksek olması sonucun önce geleni içeren işlemlerde önce geleni içermeyenlere göre daha sık göründüğünü gösterir.

Örneğin, {domates}→{dereotu}birlikteliği düşünülecek olursa, Destek({dereotu})=0.4, Güven({domates}→{dereotu})=0.67. Bu yüzden, yükseltme({domates}→{dereotu}) = 0.67/0.4 = 1.675 olur. Tam tersine, aynı güven değeri için{domates}→{ekmek} birliktelik kuralı düşünüldüğünde; destek({ekmek})=0.6. Güven({domates}→{ekmek}) =0.67. Bu yüzden, yükseltme({domates}→{ekmek}) =0.67/0.6=1.117 olur. Yükseltmenin bu göreceli değerleri domatesin ekmeğin sıklık değerine göre dereotu sıklığı üzerinde daha fazla etkiye sahip olduğunu gösterir.

Düşük tekrar sayılı ve yüksek yükseltme değerli bir birliktelik daha düşük yükseltme değerli ve daha yüksek tekrar sayılı alternatif kuraldan daha az ilgi çekici olabilir. Çünkü sonraki daha bireyseldir ve önce gelen ve sonuç arasındaki etkileşimden kaynaklanan birimlerin sayılarındaki toplam artış daha büyüktür.[53] Bu durumda tek bir değerdeki hem gücü hem de hacmi içeren ölçü dürtü (leverage)'dır.

Dürtü(A →C) = destek(A →C) − destek(A) × destek(C)

[51] Piatetsky- G. Shapiro, Discovery, Analysis, and Presentation of Strong Rules. in Knowledge Discovery in Databases. Menlo Park, CA: AAAI/MIT Press. 1991
[52] B,.Liu, W. Hsu,., S. Chen,. ve Y. Ma, "Analyzing the subjective interestingness of association rules", IEEE Intelligent Systems, 15. 2000. s 47–55
[53] Piatetsky- G. Shapiro, **a.g.e**

Örneğin, {havuç}→{domates} ve {dereotu}→{domates} birlikteliklerini düşünürsek. Her ikisi de güven =1.0 and yükseltme =1.667 değerlerine sahiptir. Bununla birlikte, ikincisi daha fazla müşteriye uygulandığından daha büyük ilginçliğe sahip olabilir. Destek({havuç}→{domates}) = 0.3. Destek({havuç}) = 0.3. Destek({domates}) = 0.6. Bu yüzden, dürtü({havuç}→{domates}) = 0.3 − 0.3×0.6 = 0.3− 0.18 = 0.12. Destek({dereotu}→{domates}) = 0.4. Destek({dereotu}) = 0.4. Destek({domates}) = 0.6. Bu yüzden, dürtü({havuç}→ {domates}) = 0.4 − 0.4× 0.6 = 0.4 − 0.24 = 0.16. Sonrakinin birlikteliğe mutlak etkisi öncekinden daha büyüktür.

Daha küçük değerli birlikteliklerin dikkate alınmadığı minimum değer konarak tespit edilen birliktelik kümesini daha da kısıtlamak için yükseltme veya dürtü ölçüleri kullanılır.

Tablo 6'ı yakından incelendiğinde, benzer birliktelik kurallarının farklı şekiller yarattığı görülür. Örneğin son iki birliktelik kuralı aynı ürünleri fakat farklı önce gelen ve sonuç kombinasyonlarını içerirler. Herhangi bir ürün seti için, ürün seti parçaları arasındaki tüm birliktelik kuralları aynı destek fakat farklı güven, yükseltme ve dürtü değerlerine sahip olabilir. Örneğin verilen iki örnekte de 0.3 destek değerine fakat farklı güven değerine sahiptir. Her ikisinde de önce gelen ve sonuç değerleri farklı {havuç, ekmek, domates} kümesinin farklı parçalarından oluşur. Bazı uygulamalar için ürün setlerinin birbirleri ile olan ilginçlik benzerliği ve ürünlerin önce gelen ve sonuç olarak bölünmesi fayda sağlamayabilir. Örneğin, markete hangi ürünün hangi ürüne yakın yerleştirileceği düşünüldüğünde, bir kural yapısı tarafından gösterilen tesadüfi ilişkiye göre gruplamaktansa ilgili olanları yan yana gruplamak mantıklıdır. Daha önceden tanımlanmış ürün setinin kendi aralarında kesin bir düzenleme olmaksızın ürünlerin benzerliğini tanımlar. İlginçlik ölçüleri, ilginç ürün setleri otomatik olarak keşfedilecekse gereklidir. Birliktelik kuralının benzeri olan yükseltme ve dürtü aşağıdaki gibi tanımlanır:

$$\text{ürün seti-yükseltmesi } (I) = \text{destek}(I) \,/\, \prod \text{destek}(\{i\})$$

$$\text{ürün seti-dürtüsü } (I) = \text{destek}(I) - \prod \text{destek}(\{i\})$$

Ürün küme yükseltme oranı gözlenen desteğin ürünler arası korelasyon olmaması durumunda beklenecek desteğe olan oranıdır. {armut, ekmek, domates} için ürün küme

yükseltme, bu küme için geçerli destek değeri (0.3) 'ün ürün setindeki her bir ürünün destek değerlerinin çarpımına (0.4 × 0.6 × 0.6 = 0.144) bölünmesiyle 2.08 olarak bulunur. Ürün seti dürtüsü gözlenen destek ile ürünler arası korelasyon olmaması durumunda beklenecek destek arasındaki farktır.

Birliktelik kuralı keşif araştırmalarının çoğu sık tekrar eden ürün seti keşfinin alt problemlerine yoğunlaşmıştı.[54] Kısıtlamaları destek ve güven değerlerine göre gerçekleştiren birliktelikleri araştırırken sık tekrar eden ürün seti bulununca birliktelik kuralı yaratmak çok kolaydır.

Daha önce tanımlanan Apriori Algoritması ürünlerin birbirleri ile sık tekrarlanmadığı ve minimum destek değerinin yüksek olduğu durumlara karşılık gelen seyrek veri kümeleri için çok etkilidir. Ayrıca ürün küme sayısı daha az ve dolayısı ile daha az işlem gerektiği için oldukça hızlıdır. Ancak Apriori algoritması, veri yoğun olduğunda tekrar eden ürün kümeleri fazla olduğundan ve bu sebepten işlem sayısını ve hesaplama süresini arttıracağından daha az etkilidir.

Bunun azaltılabilmesi için bir yaklaşım sık tekrar eden ürün küme sayısını azaltmaktır. Örneğin sık tekrar eden kapalı ürün kümeleri yaklaşımı sadece kapalı ürün kümelerini kaydeder.[55]

Bir ürün seti I'nın destek değeri ((destek(S)) olan ve kendisinin destek değerine ((Destek (I)) eşit olan bir S üst kümesi yoksa I ürün kümesi kapalıdır. Eğer S üst kümesi destek değeri, I'nın destek değerine eşitse her iki küme de aynı işlem içerisinde birlikte görünürler demektir. Kapalı sık tekrar eden ürün kümesinden ve onun destek değerlerinden tüm sık tekrar eden ürün kümelerini ve destek değerlerini belirlemek mümkündür. Tablo 7'de sık tekrar eden kapalı ürün kümesi görülmekte ve bu kümenin Tablo 5' teki 15 ürün kümesiyle kıyaslandığında 9 kapalı sık tekrar eden ürün kümesine düştüğü görülmektedir.

---

[54] J. Han,. J. Pei,., ve Y. Yin, "Mining Frequent Patterns Without Candidate Generation". In Proceedings of the ACM SIGMOD International Conference on Management of Data SIGMOD. New York: ACMPress. 2000. ss. 1–12.
J. Pei, J. Han,.,ve R. Mao, "CLOSET:An Efficient Algorithm for Mining Frequent Closed Itemsets". In Proceedings of the ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery.. New York: ACM Press. 2000 ss. 21–30.

[55] J. Pei., J. Han, H. Lu, S. Nihi, S. Tang ve D. Yang, "H-MINE: Hyper-Structure Mining Of Frequent Patterns in Large Databases". In Proceedings of the First IEEE International Conference on Data Mining. New York: IEEE Press. 2001. ss 441–448.

Tablo 7. Kapalı Sık Tekrar Eden Ürün Kümeleri

armut [Destek=0.4]
ekmek [Destek=0.6]
portakal [Destek=0.3]
patates [Destek=0.3]
domates [Destek=0.6]
ekmek, domates [Destek=0.4]
dereotu, domates [Destek=0.4]
armut, ekmek, domates [Destek=0.3]
havuç, ekmek, domates [Destek=0.3]

İşlenmesi gereken sık tekrarlanan ürün kümeleri sayısını azaltmaktan kaynaklanan çabuk hesaplama avantajlarına ilaveten, kapalı sık tekrarlanan ürün setini destekleyen yaklaşıma göre, A→C kuralına göre yaratılan birliktelik kurallarını da A ve A U C'yi de kapalı sık tekrarlanan ürün kümesi olarak kabul eder ve birliktelik kuralı sayısını kısıtlar. Sonuçta Tablo 8'de görüldüğü gibi 41 yerine 26 birliktelik kuralı kalmış olur.

Tablo 8. Kapalı Sık tekrar Eden Ürün Kümelerinden Yaratılan Birliktelik Kuralları

→ armut, ekmek, domates [Destek=0.3, Güven=0.30]
→ armut [Destek=0.4, Güven=0.40]
→ ekmek, domates [Destek=0.4, Güven=0.40]
→ ekmek [Destek=0.6, Güven=0.60]
→ dereotu, domates [Destek=0.4, Güven=0.40]
→ portakal [Destek=0.3, Güven=0.30]
→ patates [Destek=0.3, Güven=0.30]
→ domates [Destek=0.6, Güven=0.60]
armut → [Destek=0.4, Güven=1.00]
armut → ekmek, domates [Destek=0.3, Güven=0.75]
armut, ekmek, domates → [Destek=0.3, Güven=1.00]
havuç, ekmek, domates → [Destek=0.3, Güven=1.00]
ekmek → [Destek=0.6, Güven=1.00]
ekmek → armut, domates [Destek=0.3, Güven=0.50]

Tablo 8. devam

ekmek → domates, havuç [Destek=0.3, Güven=0.50]
ekmek → domates [Destek=0.4, Güven=0.67]
ekmek, domates → [Destek=0.4, Güven=1.00]
dereotu, domates → [Destek=0.4, Güven=1.00]
portakal → [Destek=0.3, Güven=1.00]
patates → [Destek=0.3, Güven=1.00]
domates → [Destek=0.6, Güven=1.00]
domates → armut, ekmek [Destek=0.3, Güven=0.50]
domates → havuç, ekmek [Destek=0.3, Güven=0.50]
domates → ekmek [Destek=0.4, Güven=0.67]
domates → dereotu [Destek=0.4, Güven=0.67]
domates, ekmek → armut [Destek=0.3, Güven=0.75]
domates, ekmek → havuç [Destek=0.3, Güven=0.75]

Apriori Algoritmasının diğer bir dezavantajı, sık tekrarlanan ürün kümelerinin uzun olması durumunda ortaya çıkar. Bu problemin üstesinden gelmek sadece başka sık tekrar eden ürün kümesinin alt kümesi olmayan en sık tekrar eden ürün kümelerini aramakla mümkün olur. [56]

En çok tekrar eden ürün kümesi, kapalı sık tekrar eden ürün kümelerinin alt kümesi olduğundan, sık tekrar eden kapalı ürün kümelerine göre daha etkin hesaplama performansına sahiptir.

Apriori algoritmasını hızlandırmanın diğer bir yolu ise sık tekrarlanan ürün kümelerinden örnekleme (sampling) yapmaktır.

Bazı veri madenciliği uygulamalarında sık tekrarlanmayan birliktelikler daha büyük ilginçlik değerine sahip olabilir. Çünkü onlar daha yüksek değerli işlemleri ilişkilendirirler. Bu durum votka ve havyar problemi olarak adlandırılır ve Ketel Votka

[56] R. J. Jr. Bayardo,. "Efficiently Mining Long Patterns from Databases". In Proceedings of ACM'SIGMOD International Conference on Management of Data. New York: ACM Press. ss. 85–93. 1998.

ve Beluga Havyarı gibi pahalı ürünlerin arasındaki birlikteliklerin seyrek tekrarlanan ama ilgi çekici olduğu gösterilir.[57]

Minimum destek kısıtlamasını kullanmadan kurallar ortaya çıkarabilen birçok algoritma geliştirildi. Cohen algoritması [58] çok yüksek güven değerli birliktelikleri bulur.

Kural-alan araması için kullanılan çeşitli algoritmalar, doğrudan potansiyel kurallar için alan araştırması yapar.[59] Bu yapı Apriori Algoritmasının esasını teşkil eden minimum destek değerine nazaran ilginçlik kurallarını ortaya çıkarmak için geniş kıstas aralığına izin verir. Bununla birlikte, kritik bir ihtiyaç da bu kıstasın, etkin olarak tanımlanması gereken ilginçlik kurallarını içerebilen araştırma alanlarına imkan vermesidir.

Birliktelik kuralları market sepet analizinde çok geniş uygulamalarda kullanım alanına sahip olmasına rağmen sayısal değerlere doğrudan uygulanamaz. Sayısal değerlerle çalışabilmek için standart yaklaşım veritabanındaki sayısal alanların değerlerini alt gruplara ayırmaktır. Her bir alt grup birliktelik analizinde bir ürün olarak ele alınır. Veritabanındaki her bir kayıt için sayısal değerin içinde bulunduğu alt aralığa karşılık gelen ürünü kapsarmış gibi ele alınır. Çoğu sistem veriyi öncelikle gruplara ayırır, yani birliktelik analizi yapılmadan aralıklar tespit edilir. Bu tip bir yaklaşım bütün birliktelik kuralları için sayısal değer ve farklı alternatif alanları arasında birlikteliklerin içeriğinde, farklı aralık kümeleri daha uygun olmasına karşılık sadece tek bir aralık kümesi kullanılır. Srikant ve Agrawal [60] bu problemi, birliktelik kuralı keşfi süresince ayrıklaştırma aralıklarını dinamik olarak ayarlayarak çözerler. Sayısal

---

[57] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. Ullman, ve C. Yang,. "Finding Interesting Associations without Support Pruning". In Proceedings of the International Conference on Data Engineering New York: IEEE Press. 2000. ss. 489–499.

[58] **a.g.e**

[59] R. J. Jr Bayardo, R, Agrawal,. ve D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases". Data Mining and Knowledge Discovery 4. ss 217–240. 2000.

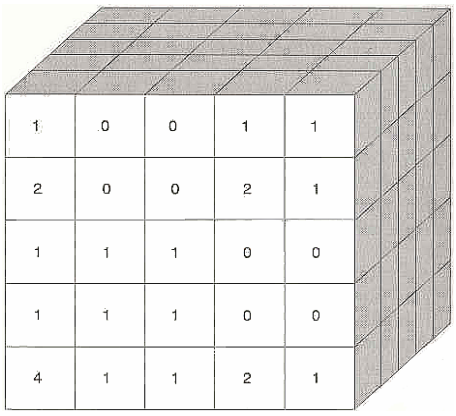
R. J., Jr. Bayardo,.ve Agrawal, R. "Mining The Most Interesting Rules". In Proceedings of the Fifth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press ss. 145–154. 1999.

G. I. Webb, "Efficient Search for Association Rules". In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press. 2000. ss. 99–107.

[60] R. Srikant,. ve R. Agrawal, "Mining quantitative association rules in large relational tables". In Proceedings of the ACM 'SIGMOD International Conference on Management of Data. New York: ACM Press. 1996. ss. 1–12

değerler için diğer bir çeşit birliktelik benzerlik kuralı darbe araştırması [61] ve etki kuralı [62] tarafından sağlanır. Bu teknikler sayısal değerler üzerinde özel etki çeşitleri ile birlikte olan önce gelenleri bulur.

İkiden fazla ürünün birliktelik analizini yapmak kolay değildir. Örneğin eğer üç ürünün birliktelik analizi yapılmak istenseydi, birliktelik tablosu iki boyutlu olmaktan çıkıp küp halini alacaktır:



Şekil 7. Üç Boyutlu Birliktelik Tablosu

Kaynak: Linoff G. ve Berry M.J.A., 2004. Data Mining techniques For Marketing Sales and Customer Relationship Management, Wiley Publishıng, New York.

Toplam beş ürünün bulunduğu bu basit analizde dahi doldurulması gereken toplam 125 adet kutu vardır. Bu sayı, küp yapısının simetriklik özelliği kullanılarak bir miktar azaltılabilse dahi gene de n adet ürün için n^3 mertebelerinde bir analiz gerektirecektir. "n" adet ürünün bulunduğu bir veri ambarında ürünlerin n'li kombinasyonlarını bulmak, n^n boyutunda analiz yapmayı gerektirmektedir.

Örneğin bir süpermarkette en az 10.000, çoğunlukla ise 20.000 ila 30.000 kalem ürün bulunur. Bu ürünlerin ikili kombinasyonlarının toplam olasılığı 50 milyon, 3'lü kombinasyonlarının olasılığı ise 100 milyara yakındır. Ayrıca süpermarketlerin satış

[61] J. H. Friedman,. ve N. I. Fisher, "Bump hunting in high-dimensional data", Statistics and Computing, 9. 1999. s 123–143.
[62] A. Aumann, ve Y. Lindell, "A Statistical Theory For Quantitative Association Rules". In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining.. New York:ACM Press. 1999. ss. 261–270.

kayıtları da çok büyük ölçektedir. Orta ölçekli (10-20 şubeli) bir süpermarketin yılda on milyonlarla ölçülen satış kayıtları bulunmaktadır. Günümüzün gelişmiş bilgisayarlarında dahi bu büyüklükteki verileri işleyip üçlü ürün kombinasyonları çıkarmak çok pahalıya mal olmaktadır. Bu verilerden 5'li ya da daha fazla kombinasyonların çıkarılması ise çok büyük şirketlerin dahi altından kalkamayacağı bir maliyet getirmektedir.

Sepet analizinin başarılı olduğu noktalar:

- Kolay ve anlaşılır sonuçlar üretir.
- Gözetimsiz veri madenciliği yöntemidir.
- Değişik boyutlardaki veriler üzerinde çalışabilir.
- Her ne kadar kayıtların sayısı ve kombinasyon seçimine göre işlem adedi artsa da sepet analizi için her adımda gerekli olan hesaplamalar diğer yöntemlere göre (genetik algoritmalar, yapay sinir ağları vb.) çok daha basittir.

Sepet analizinin başarısız olduğu noktalar:

- Sorunun boyutu büyüdükçe, gerekli hesaplamalar üstel olarak artmaktadır.
- Sepet analizinde kullanılacak doğru ürünlerin seçimi. Ürün gruplandırma (süt ürünleri, unlu mamüller vb.) biraz bilgi kaybı getirse de analizin boyutlarını küçültebilir.
- Kayıtlarda çok az rastlanan ürünleri yok sayar. Sepet analizi tekniği, en doğru sonucu, tüm ürünlerin kayıtlar içinde yaklaşık aynı frekansta göründüğü durumlarda üretmektedir.

2.2.5.2. Bellek Tabanlı Yöntemler

İnsanlar kararlarını genellikle daha önce yaşadıkları deneyimlere göre verirler. Örneğin doktorlar bir hastayı incelerken, elde ettiği bulguları daha önce tedavi ettiği benzer hastalığa yakalanmış hastalar üzerindeki deneyimlerini kullanırken bir sigorta eksperi de bir olayın sahtekarlık olup olmadığı bulmak için daha önceki sahtekarlıklarla olan benzerlikleri göz önünde bulundurmaktadır. Gözetimli bir veri madenciliği tekniği olan bellek tabanlı yöntemler de benzer şekilde deneyimleri kullanmaktadır. Bu yöntemlerde, bilinen kayıtların bulunduğu bir veri tabanı oluşturulur ve sistem yeni

gelen bir kayda komşu olan diğer kayıtları belirler ve bu kayıtları kullanarak tahminde bulunur ya da bir sınıflandırma işlemi gerçekler. Bellek yaklaşımlı yöntemlerin en önemli özelliği veriyi olduğu gibi kullanabilme yeteneğidir. Diğer veri madenciliği tekniklerinin aksine bellek tabanlı yöntemler, kayıtların formatı yerine sadece iki işlemin varlığı ile ilgilenir. İki kayıt arasındaki uzaklığı belirleyen bir uzaklık fonksiyonu ve komşu kayıtları işleyerek bir sonuç üreten bir kombinasyon fonksiyonu.

Bellek tabanlı yöntemler birçok alanda kullanılmaktadırlar:

- Sahtekarlık Tespitinde: Karşılaşılan her yeni sahtekarlık davası, önceki sahtekarlık davaları ile mutlaka benzerlik göstereceğinden bellek tabanlı yöntemler bu benzerlikleri bulur ve ilerideki işlemler için bu bulguları işaretler.
- Müşteri Tepkisi Tahmini: Belirli bir pazarlama faaliyeti ya da benzer bir faaliyete cevap verecek yeni bir müşteri, çok yüksek bir olasılıkla bu faaliyete daha önceden cevap veren müşterilere benzer davranışlar göstereceğinden bellek tabanlı yöntemler de bu davranışları tespit etmekte kullanılırlar.
- Klinik İşlemlerde: Hastalara verilebilecek en etkili tedavi yöntemi, benzer rahatsızlıkları daha önceden yaşamış hastalar üzerinde başarılı sonuçlar vermiş olan tedavi yöntemleri olduğundan bellek tabanlı yöntemler, en etkili tedavi yöntemini bulabilmektedir.

Bellek tabanlı yöntemlerin güçlü olduğu noktalar şunlardır:

- Kolayca anlaşılabilir sonuçlar üretir.
- Rasgele seçilen, hatta ilgisiz dahi olabilen verilere uygulanabilir.
- Analiz alanlarının çok olduğu durumlarda dahi efektif olarak çalışabilir.
- Eğitim kümesinin oluşturulması basittir.

Bellek tabanlı yöntemlerin zayıf olduğu noktalar da bulunmaktadır:

- Sınıflandırma ve kestirim işlemleri için kullanıldığında işlem maliyeti yüksektir.
- Eğitim kümesi için büyük miktarlarda yere ihtiyaç vardır.

- Üretilen sonuçlar seçilen uzaklık fonksiyonuna, kombinasyon fonksiyonuna ve komşu adedine doğrudan bağlıdır.

2.2.5.3. Demetleme

Kritik kararlar alınmadan önce genellikle bir adım geri atılır ve "büyük resim" görülmeye çalışılır. Ancak zaman zaman bu büyük resim anlaşılmayacak kadar çok karmaşıktır. Büyük bir veri tabanı çok miktarda boyut, yani alan içerebilir ve çok karmaşık bir yapıya sahip olduğundan en iyi yönetilen veri madenciliği teknikleri bile bu veri yığını içerisinden anlamlı sonuçlar üretemeyebilir.

Çok karmaşık ve büyük sorunların çözülmesinde izlenen yöntem genellikle büyük sorunu daha küçük ve tek başına daha rahat çözülebilecek alt sorunlara bölmek ve her bir alt sorunu çözdükten sonra çözümleri birleştirerek sonuca gitmek şeklindedir. Ancak bazı durumlarda veriler öyle dağılmışlardır ki nereden bölüneceği hangi şekilde alt gruplara ayrılabileceğini kestirmek mümkün değildir. Bu yüzden otomatik demet bulma yöntemleri geliştirilmiştir.

Veri madenciliği yöntemlerinden bir tanesi olan demetleme sadece veri madenciliğinde değil örüntü tanıma, imge işleme ve benzeri birçok değişik alanda daha kullanılmaktadır.

2.2.5.4. İlişkisel Analiz

İş dünyası, insanların, mekanların ve bütün her şeyin birbirleri ile bağlantı kurduğu bir ilişkiler dünyasıdır. Havayolu şirketleri, kargo şirketleri ve benzeri firmalar şehirleri birbirine bağlar. Haberleşme şirketleri ile diğer müşteriler birbirleri ile telefon ve benzeri şekillerde bağlantı kurarlar. Her kredi kartı müşterisi belirli mağazaları ve lokantaları tercih eder. Benzer şekilde ilişkiler her alanda çoklukla bulunmaktadır ve bu ilişkiler birçok veri madenciliği tekniğinin kullanamayacağı kadar zengin bilgi içermektedir.

İlişkisel analiz, matematiğin bir alt alanı olan graf teorisi tabanlıdır. Ayrıca ilişkisel analiz her türlü sorunu çözemez ya da her türlü veri üzerinde uygulanamaz.

İlişkisel analizi gerçeklemek üzere birkaç tane yazılım bulunmaktadır ve bu yazılımların çoğu hukuksal alanda özelleşmişlerdir. En basit ilişkisel analiz aracı olarak ise ilişkisel veri tabanları üzerinde kullanılan SQL gösterilebilir.

İlişkisel analiz yönteminin uygulanmasında ortaya çıkan bir başka sorun ise maddeler arasındaki ilişkilerin ortaya çıkarılmasıdır. Telefon çağrılarının analizinde ilişkiler açıktır: bir kişi bir başkasını arar ve burada çağrının kendisi ilişkidir. Ancak her durumda ilişkileri bulmak bu kadar kolay değildir, ilişkilerin bulunması için otomatikleşmiş bazı işlemlerin gerçeklenmesi gerekebilir.

2.2.5.5. Karar Ağaçları ve Kural Türetme

Karar ağaçları yöntemi en güçlü ve en yaygın sınıflandırma ve öngörü araçlarından birisidir. Ağaç yapılı yöntemlerin sık kullanılmasının nedeni ise yapay sinir ağlarının tersine ağaç yapılarının kuralları ifade edebilmesinden kaynaklanmaktadır. Kurallar insanların okuyup anlayabileceği herhangi bir dile çevrilebileceği gibi, bir veri tabanında belirli bir kategoriye düşen kayıtların getirilmesi için bir SQL ifadesine dönüştürülür.

Bazı uygulamalarda, sınıflandırmanın ya da öngörünün doğruluğu, önemli olan tek şeydir, örneğin doğrudan posta ilanları ile iş yapan bir firma, hangi müşterilerin kendilerine gönderilen ilanlara olumlu yanıt vereceğini öngören bir model sahibi olduğunda bu modelin nasıl veya neden çalıştığını sorgulamaz. Bazı modellerde de hem yapay sinir ağları hem de karar ağaçları bir arada kullanılmıştır.

Sık oynanan bazı oyunlarda olduğu gibi karar ağaçları da bir dizi soru sorup bunların cevapları doğrultusunda hareket ederek en kısa sürede sonuca gider. Karar ağaçları, sorduğu bir soruya gelen cevap ile soracağı diğer soruları belirler. Eğer sorular iyi seçilmiş olursa, yeni gelen bir kaydın sınıflandırılması işlemi, en az sayıda soru sorarak gerçeklenebilir.

Sorular ve bu soruların cevaplarının yönlendirdiği başka soruların bulunduğu bir ağaç yapısı olarak adlandırılan karar ağaçları ile değerlendirme yaparken yeni gelen bir kayıt, ağacın kökünden girer. Kökte test edilen bu yeni kayıt bu test sonucuna göre bir alt düğüme gönderilir. Bu süreç, yeni kayıt herhangi bir yaprak düğüme gelene kadar

devam eder. Ağacın belirli bir yaprağına gelen bütün yeni kayıtlar aynı şekilde sınıflandırılırlar. Kökten her bir yaprağa giden sadece tek bir yol vardır. Bu yol, kayıtları sınıflandırmak için kullanılan bir kuralı tanımlamaktadır. Bazı yapraklar aynı sınıflandırmayı yapabilirler fakat her bir yaprak bu sınıflandırmayı farklı nedenlere dayanarak yaparlar.

Karar ağaçlarının güçlü olduğu noktalar şunlardır:

- Karar ağaçları kolay anlaşılır sonuçlar üretir.
- Çok sayıda işlem yapılmasına gerek duymadan sınıflandırma işlemini gerçekleyebilmektedir.
- Hem sayısal hem de kategorik veriler üzerinde işlem yapabilmektedir.
- Karar ağaçları hangi alanların sınıflandırma ve kestirme için daha önemli olduğunu açık şekilde belirtmektedir.

Karar ağaçlarının zayıf olduğu noktalar ise şöyledir:

- Karar ağaçlarının tahmin için kullanıldığı durumlarda tahmin edilecek değişkenin sürekli değerler alması durumunda uygun sonuçlar üretilememektedir.

2.2.5.6.Yapay Sinir Ağları

Yapay sinir ağları, sınıflandırma, demetleme ve kestirim amaçları ile kolaylıkla kullanılabilecek genel amaçlı ve güçlü araçlardır. Ekonomik alanlardan tıbbi konulara, değerli müşterilerin belirlenmesi için yapılan demetleme işlemlerinden kredi kartlarında sahtekarlıkların belirlenmesine kadar çok geniş bir alanda uygulama alanı bulmuştur.

Yapay sinir ağlarını, insanların deneyimlerinden bir takım bilgiler çıkartması gibi kendisine verilen örneklerden bir takım bilgiler çıkartma yeteneğine sahiptir. Yapay sinir ağları öncelikle sonuçları bilinen belirli bir veri kümesi üzerinde öğrenme algoritmaları çalıştırılarak eğitilir. Bu eğitim neticesinde yapay sinir ağının içerisindeki bir takım ağırlıklar belirlenir. Bu ağırlıklar kullanılarak yeni gelen veriler işlenir ve bir sonuç üretilir. Yapay sinir ağlarının en olumsuz tarafı ise bu ağırlıkların neden ilgili değeri aldığının bilinmemesidir. Ya da çıkan sonucun neden geçerli bir sonuç olduğunu

bize açıklayamaz. Yapay sinir ağlarını kullanmak için en iyi yaklaşım onları içi bilinmeyen bir şekilde çalışan kara kutular olarak düşünmek olacaktır.

Yapay sinir ağlarının veri madenciliği açısından kuvvetli yönleri şunlardır:

- Çok geniş spektrumdaki sorunların çözümünde kullanılabilirler.
- Çok karmaşık durumlarda dahi iyi sonuçlar üretmektedirler.
- Hem sayısal hem de kategorik veriler üzerinde işlem yapabilirler.

Yapay sinir ağlarının olumsuz yönleri de vardır:

- 0 ile 1 arasında giriş verileri olması zorunludur.
- Ürettikleri sonuçların açıklamasını yapamazlar.
- Varılan sonucun olası en iyi sonuç olduğunun garantisi yoktur.

2.2.5.7. Genetik Algoritmalar

Genetik algoritmalar da yapay sinir ağları ve bellek tabanlı yöntemler gibi biyolojik işlemlerden kaynağını almıştır. Yüzyıllar boyu süren adaptasyonlar ve doğal seleksiyon ile çevre koşullarına en fazla uyum sağlayanlar hayatta kalmışlardır. Genetik algoritmaların da benzer bir çalışma biçimi vardır. Geçtiğimiz yıllar boyunca genetik algoritmalar yapay sinir ağlarının eğitilmesi, bellek tabanlı yöntemlerde kombinasyon fonksiyonunun oluşturulması gibi işlerde kullanılmışlardır.

Genetik algoritmalar açıklanabilir sonuçlar üretirler. Çok değişik tiplerdeki verileri işleme özelliğine sahip olan genetik algoritmalar optimizasyon amacı ile kullanılabilirler. Ayrıca genetik algoritmalar yapay sinir ağları ile ortaklaşa çalışarak başarılı sonuçlar üretmektedirler.

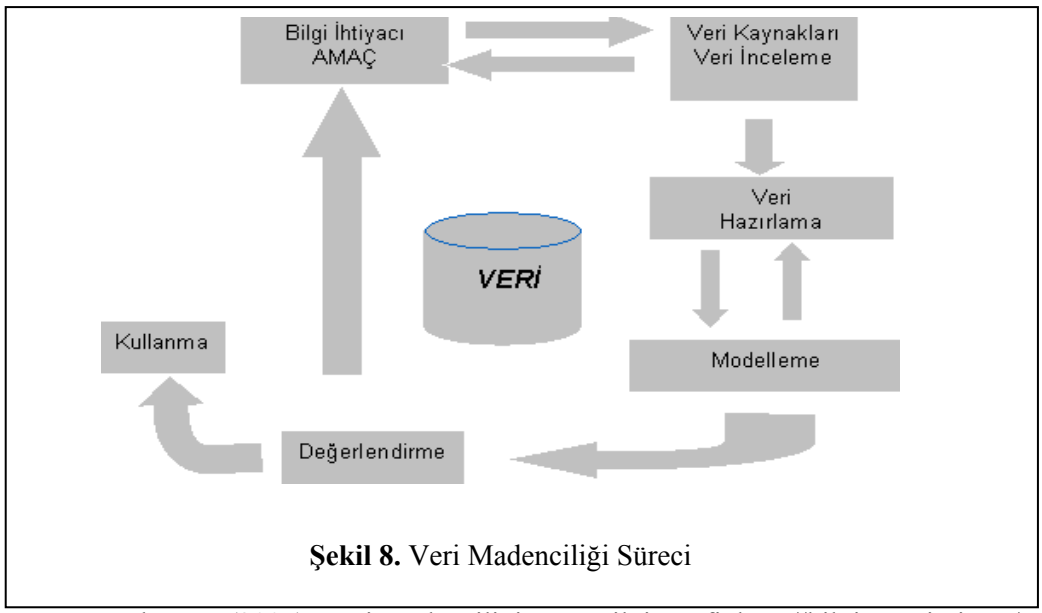
Genetik algoritmaların kullanılmalarındaki en belirgin sorunlardan biri karmaşık sorunların genetik kodlanmasının çok zor olmasıdır. Ayrıca en uygun sonucun üretildiğine dair bir garanti de bulunmamaktadır. Son olarak genetik algoritmaların çalıştırılması çok ağır bir işlem yükü getirmektedir.

2.2.6. Veri madenciliği Süreci

Veri madenciliği algoritmasının üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması olanaksızdır. Bu nedenle iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır. Başarılı bir veri madenciliği projelerinde izlenmesi gereken adımları aşağıdadır.[63] Problemin tanımlanması,

- Verilerin hazırlanması,
- Modelin kurulması ve değerlendirilmesi,
- Modelin kullanılması,
- Modelin izlenmesi.

Veri madenciliği süreci Şekil 8'de gösterilmektedir.



Şekil 8. Veri Madenciliği Süreci

Kaynak: Eker, H. (2005). Veri Madenciligi veya Bilgi Keşfi. http://bilgiyonetimi.org/m/pages/mkl_list.php?id=11 (2006)

2.2.6.1. Problemin tanımlanması

Veri madenciliği çalışmalarında başarılı olabilmek için, projenin hangi işletme amacı için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl

[63] C. Shearer, "THE CRISP-DM model: The new blueprint for data mining", Journal of Data Warehousing, 5 (4). 2000. s 13-23.

ölçüleceğinin tanımlanması gereklidir. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

Bu aşamada üretilen bilginin işletme için değerinin doğru analiz edilmesi gerekmektedir. Analistin, işletmede üretilen sayısal verilerin boyutlarını, proje için yeterlilik düzeyini ve iş süreçlerini iyi analiz etmesi gerekmektedir.

2.2.6.2. Verilerin Hazırlanması

Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50 - %85 ini harcamasına neden olmaktadır.[64]

Bu aşamada firmanın mevcut bilgi sistemleri üzerinde ürettiği sayısal bilginin iyi analiz edilmesi, veriler ile mevcut iş problemi arasında ilişki olması gerekmektedir. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile yaratıldığı da bu veriler kullanılmadan analiz edilmelidir, bu sayede analist veri kalitesi hakkında fikir sahibi olabilir.

Verilerin hazırlanması aşaması kendi içerisinde toplama, birleştirme ve temizleme, dönüştürme adımlarından meydana gelmektedir.

Veri toplama, tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, gibi veritabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

Birleştirme ve Temizleme adımında toplanan verilerde bulunan farklılıklar giderilmeye çalışılır. Hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin temizlenmesine çalışılır. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin, önemli bir uyarıcı bilgi içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir. Ancak basit

[64] U S. Piramuth. "Evaluating Feature Selection Methods For Learning in Data Mining Applications". Thirty- First Annual Hawai International Conference on System Sciences, 5. 1998. ss294.

yöntemlerle ve gelişi güzel olarak yapılacak sorun giderme işlemleri sonraki aşamalarda büyük sorunları yaratabilir.

Veri Dönüştürme safhasında kullanılacak model ve algoritma çerçevesinde verilerin tanımlama veya gösterim şeklinde değiştirilmesi gerekebilir. Örneğin kredi riski uygulamasında iş tiplerinin, gelir seviyesi ve yaş gibi değişkenlerin kodlanarak gruplanmasının faydalı olacağı düşünülmektedir.

2.2.6.3. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar tekrarlanan bir süreçtir.

Model kuruluş süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.[65]

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunmasıdır.

Öğrenme süreci tamamlandığında, tanımlanan kurallar verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak

[65] I. Witten.ve E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". San Fransisco: Morgan Kaufmann Publishers. 2000.

gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi (accuracy) belirlenir.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (simple validation) testidir. Bu yöntemde tipik olarak verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.[66]

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem çapraz geçerlilik (cross validation) testidir. Bu yöntemde veri kümesi tesadüfi olarak iki eşit parçaya ayrılır. İlk aşamada birinci parça üzerinde model eğitimi ve ikinci parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (n-fold cross validation) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır.

Küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka teknik Bootstrapping'dir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulur. Daha sonra en az 200, bazen binin üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik

[66] H.Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, 2000, c:29, 1, s: 1-22.

modeller kurulabilir. Model kuruluş çalışmalarına başlamadan önce, imkansız olmasa da hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere sayısız deneme yapılmasında yarar bulunmaktadır.

Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit ancak faydalı bir araç olan risk (yakınsaklık) matrisi kullanılmaktadır. Aşağıda bir örneği görülen bu matriste sütunlarda fiili, satırlarda ise tahmini sınıflama değerleri yer almaktadır. Örneğin fiilen B sınıfına ait olması gereken 46 elemanın, kurulan model tarafından 2'sinin A, 38'inin B, 6'sının ise C olarak sınıflandırıldığı Tablo 9'teki matriste görülmektedir.[67]

Tablo 9. Risk Matrisi

Tahmin Edilmiş Sınıf	Gerçek Sınıf		
	A Sınıfı	B Sınıfı	C Sınıfı
A Sınıfı	45	2	3
B Sınıfı	10	38	2
C Sınıfı	4	6	40

Modelin anlaşılabilirliği diğer önemli bir değerlendirme kriteridir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Seyrek olarak yorumlanamayacak kadar karmaşık olsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

Model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile edilecek olan yatırımın geri dönüş (return on investment) oranı kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçüdür.

[67] F. Aydoğan, "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi". Yüksek Lisans Tezi. Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. Ankara. 2003.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satınalma davranışını belirgin olarak etkileyecektir.

2.2.6.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmini envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.[68]

2.2.6.5. Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.[69]

Veri madenciliği süreci incelenirse Şekil 9'daki gibi sürecin dört aşamadan oluştuğu görülür..[70]

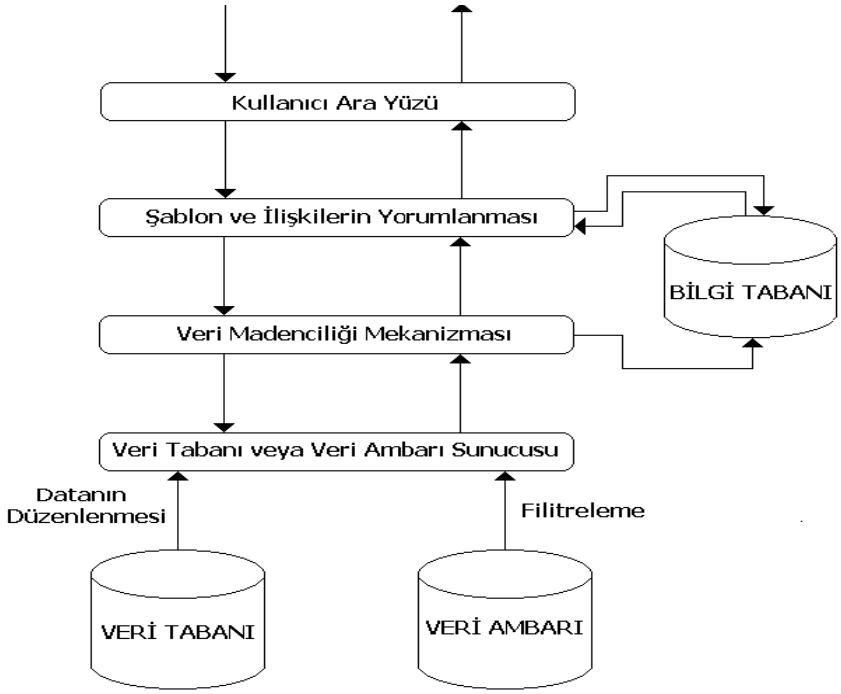
- Uygulama Alanın Ortaya Konulması
- Veri Ambarının Oluşturulması
- Modelin Kurulması ve Değerlendirilmesi
- Şablonların ve İlişkilerin Yorumlanması

---

[68] C. Shearer, "THE CRISP-DM model: The new blueprint for data mining", Journal of Data Warehousing, 5 (4). 2000. s 13-23.

[69] **a.g.e.**

[70] E. Kaya, M. Bulun ve A. Arslan, "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları". 2006 http://www.ab.org.tr/ab03/program/96.html (Erişim Tarihi:18.07.2007)



Şekil 9. Veri Madenciliği Aşamaları

Kaynak: Kaya, E.; Bulun, M.; Arslan, A. Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları. www.ab.org.tr/ab03/program/ 96.html .2006

3. UYGULAMA

Marketlerde satılan ürünlerin birbirleri ile olan ilişkileri yani hangi ürünün hangi ürünle birlikte satın alınabileceği günümüzde market stratejilerinin belirlenmesinde oldukça önemli rol oynamaktadır. Bu maksatla market yöneticileri birliktelik analizi yapılan ürünlerin yerleşimini yan yana ya da birliktelikleri tespit edilen ürünlerin arasına görülmesi veya fark edilmesi istenen başkaca ürünler koyarak farklı yöntemlere göre yapabilmektedirler.

Veri madenciliği uygulama safhasında gerçek market veri tabanı üzerinde birliktelik ve sepet analizi yapılacaktır.

3.1 Veri Madenciliği Süreci

Analizde kullanılacak olan veri bir marketin kasa kayıtları alınarak oluşturulmuş ve toplam 11507 kayıt olarak ele alınan veri tabanı bir tabloda toplanmıştır. Veriye ait özellikleri Tablo 10'te gösterilmiştir.

Tablo 10. Veri Tabanı Başlangıç Durumu

Veri Adı	Tipi ve Özellikleri
Fiş No	Sayısal değer
Tarihi	Tarih
Dönemi	Sayısal
Barkodu	Sayısal
Ürün Uzun Adı	Metin
Miktarı	Sayısal
Birim Maliyeti	Sayısal
Satış Fiyatı	Sayısal
Ödeme tipi	Sayısal Kod

Veri 913 farklı ürünü içermekte ve her bir fiş numarasına karşılık gelen kısmında ise aynı üründen birden fazla yer almaktaydı. Veri Tabanı bu aşamadan sonra veri dönüştürme işlemine tabi tutulmuştur.

Bu adımda toplanan verilerde bulunan farklılıklar giderilmiş, hatalı veya analizin yanlış yönlenmesine sebep olabilecek veriler temizlenmiştir. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin önemli bir uyarıcı bilgi içerip içermediği kontrol edilmiş ve aynı sepet içerisinde tekrar eden ürünlerde kontrollü olarak veri kümesinden atılmıştır. Ürün gruplarının isimlendirilmesindeki farklılıklar dikkate alınarak ürünler belli ürün grupları başlıkları altında toplanmış ve sıhhatli birliktelik analizine imkan sağlanmıştır.

Ayrıca Apriori algoritması kullanarak birliktelik analizi yapıldığından başlangıç durumundaki işlenmemiş veriyi analize hazır hale getirmek için veri tiplerini sayısal değerlerden kurtaracak şekilde aralık değerleri belirlemek ya da metin formatına dönüştürecek şekilde işlemek de veri dönüştürme işlemi içerisinde yapılmıştır. Bu amaçla verideki fiş no alanı aynı satırda tek bir sepet oluşturulduğundan, dönemi, barkodu, miktarı, birim maliyeti, satış fiyatı alanlarına analizde gerek duyulmadığından veri tabanından çıkarılmış, tarih alanı gün olarak metin tipine çevrilmiş, ödeme tipi ise sayısal formattan kredi kartı, nakit ya da çek bilgilerini ihtiva edecek şekilde metin formatında kümeye çevrilmiştir. Bu işlemler sonucu veri tabanının veri adı, tip ve özellikleri Tablo 11' belirtilmiştir.

Tablo 11. Verinin Birleştirilmesi, Temizlenmesi ve Dönüştürülmesi Sonucu Son Durum

Veri Adı	Tipi ve Özellikleri
Gün	Metin, küme(P.tesi,Salı, vb)
Ödeme Tipi	Metin, küme(K, N, Ç)
Cinsiyet	Metin, küme (E, B)
Biskuvi Kek Cikolata Tatlı	Boolean,Doğru Yanlış (T/F)
Gazlı icecek	Boolean,Doğru Yanlış (T/F)
Şampuan	Boolean,Doğru Yanlış (T/F)
Kahvaltılık	Boolean,Doğru Yanlış (T/F)
Margarin	Boolean,Doğru Yanlış (T/F)

Tablo 11. devam

Çay Kahve Neskafe	Boolean,Doğru Yanlış (T/F)
Temizlik	Boolean,Doğru Yanlış (T/F)
Makarna	Boolean,Doğru Yanlış (T/F)
Parfumeri	Boolean,Doğru Yanlış (T/F)
Diş MAcunu Fırçası	Boolean,Doğru Yanlış (T/F)
Süt ürünleri	Boolean,Doğru Yanlış (T/F)
Bakliyat	Boolean,Doğru Yanlış (T/F)
Kuruyemiş	Boolean,Doğru Yanlış (T/F)
Meyve suyu	Boolean,Doğru Yanlış (T/F)
Bebek Ürunleri	Boolean,Doğru Yanlış (T/F)
Et Tavuk Balık	Boolean,Doğru Yanlış (T/F)
Sıvı Yağ	Boolean,Doğru Yanlış (T/F)
Turşu	Boolean,Doğru Yanlış (T/F)
Un Nisaşta Kabartma Tozu Puding	Boolean,Doğru Yanlış (T/F)
Konserve	Boolean,Doğru Yanlış (T/F)
Şeker Tuz	Boolean,Doğru Yanlış (T/F)
Şekerleme Sakız	Boolean,Doğru Yanlış (T/F)
Traş Malzemesi	Boolean,Doğru Yanlış (T/F)
Yumurta	Boolean,Doğru Yanlış (T/F)
Elektronik	Boolean,Doğru Yanlış (T/F)
Ambalaj	Boolean,Doğru Yanlış (T/F)
Sos	Boolean,Doğru Yanlış (T/F)
Cips Kraker	Boolean,Doğru Yanlış (T/F)
Ekmek	Boolean,Doğru Yanlış (T/F)
Su	Boolean,Doğru Yanlış (T/F)

Modelin kurulmuş ve modelin izlenmesi için de uygulamada bir markete ait Aralık 2006ve Ocak 2007 arasında yer alan bir veriler kullanılmış ve yaklaşık bir aylık hareketlerin sonucunda 2000 fiş hareketi elde edilmiştir.Mağazada bir ay süresince 913 çeşit ürün satıldığı belirlenmiştir ancak gruplama ve az rastlanan ürünlerin çıkarılması sonucu bu sayı 32'ye düşürülmüştür.En az fiş hareketi bir, en fazla fiş hareketi de 54 iken en fazla fiş hareketi 22'ye dönüştürülmüştür.

3.2. Birliktelik Analizi Yapılan Programlar

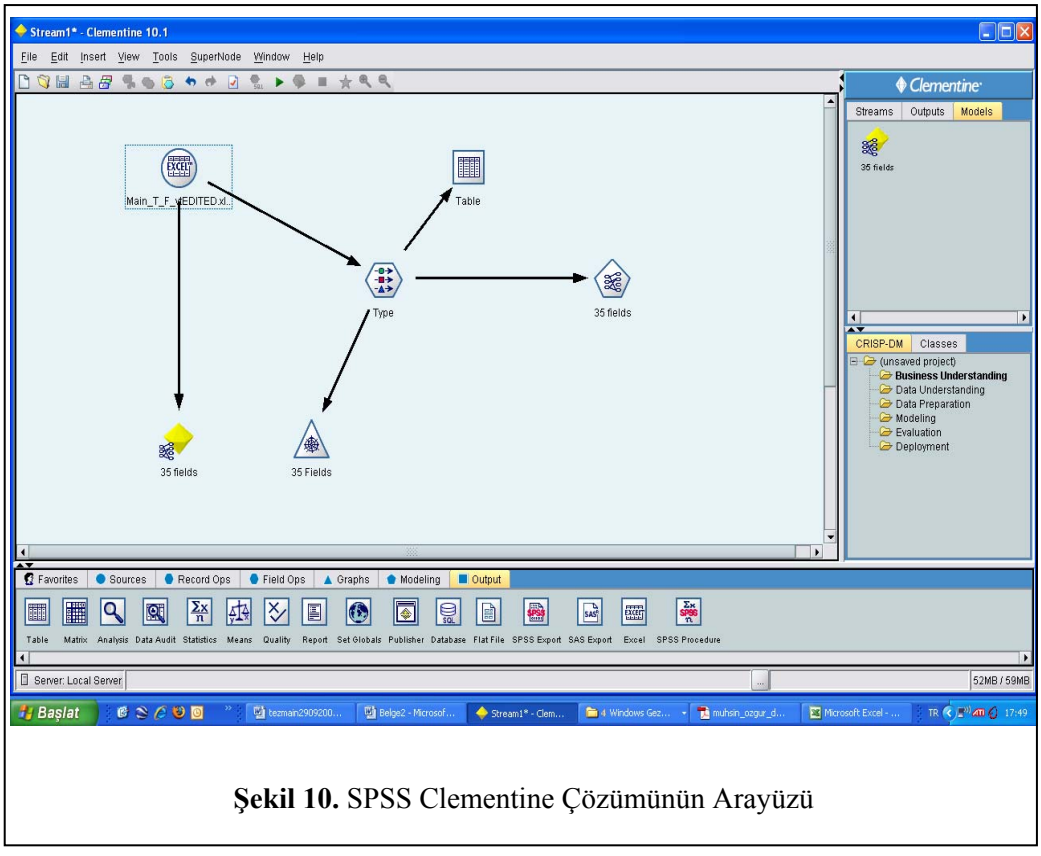
Birliktelik analizi için SPSS Clementine ve WEKA (Waikato Environment for Knowledge Analysis) programları kullanılacaktır.

3.2.1. SPSS Clementine 10.1

Bu uygulama için SPSS firmasının veri madenciliği çözümü olan SPSS Clementine 10.1 paket programı ile gerçekleştirilmiştir. Veritabanından çekilen veri kümesine Clementine 10.1'in Apriori işlemcisi ile birliktelik kuralları uygulaması gerçekleştirilmiştir.

3.2.1.1. Ürün Grupları Özellikleri Analizi

Bu uygulamada, Şekil 10'daki tasarım kullanılarak analiz sonuçları elde edilmiştir.Bu tasarım gerçekleştirilirken sadece ürün grupları seçilerek yapılan analizde destek değeri alt sınırı yaklaşık % 10 ve güven değeri alt sınırı da %50 olarak seçilmiştir. Bunun sonucunda Şekil 11'daki 17 adet destek ve güven değerlerini gösteren kurallar elde edilmiştir.



Şekil 10. SPSS Clementine Çözümünün Arayüzü

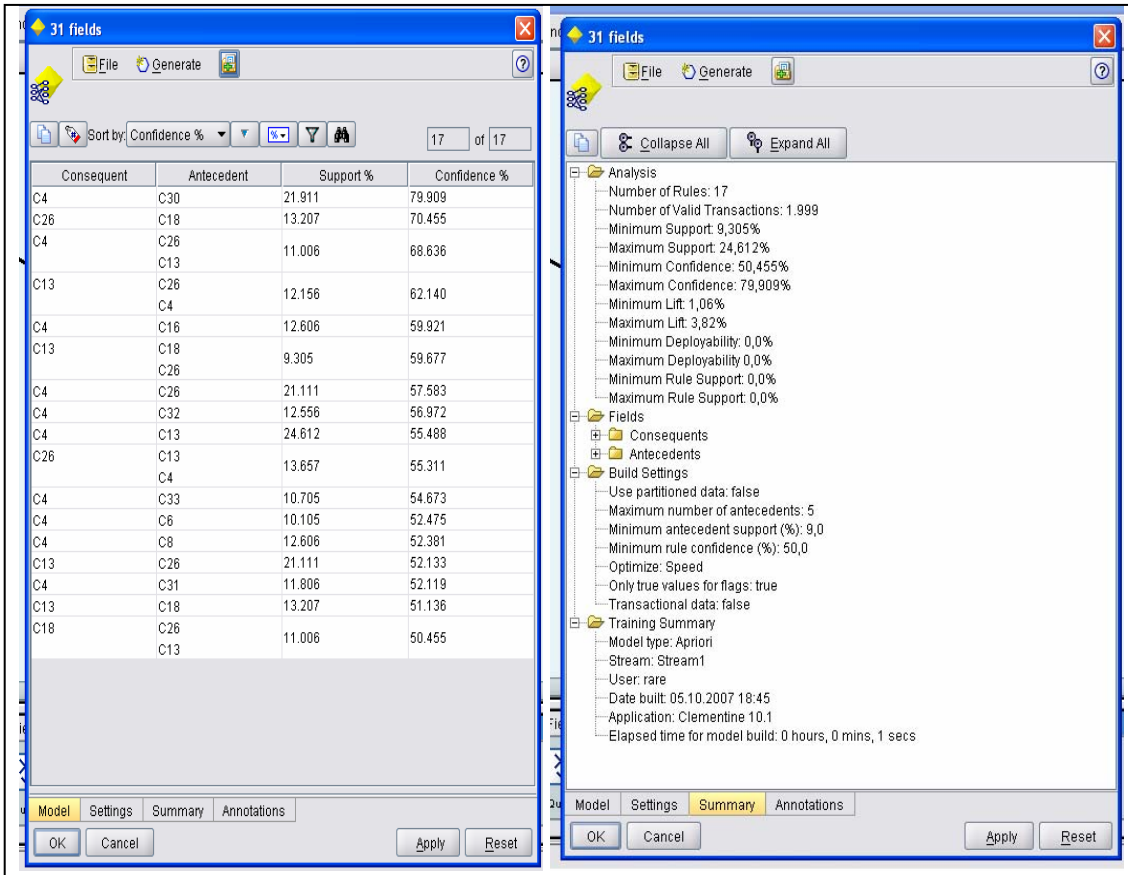
Bu uygulamadaki özellik adlarının programdaki karşılıkları Tablo 12'de gösterilmiştir.

Tablo 12 Özellik Adları Program Karşılıkları

Özellik Adı	
Gün	C1
Ödeme Tipi	C2
Biskuvi-Kek-Çikolata-Tatlı	C3
Gazlı içecekler	C4
Şampuan	C5
Kahvaltılık	C6
Margarin	C7
Çay-Kahve-Neskafe	C8
Temizlik	C9
Makarna	C10
Parfumeri	C11
Diş Macunu-Fırçası	C12
Süt ürünleri	C13
Bakliyat	C14
Kuruyemiş	C15
Meyve suyu	C16
Bebek Ürunleri	C17
Et-Tavuk-Balık	C18
Sıvı Yağ	C19
Turşu	C20
Un-Nisaşta-Kabartma-Tozu-Puding	C21
Konserve	C22
Şeker-Tuz	C23
Şekerleme-Sakız	C24
Traş Malzemesi	C25
Yumurta	C26
Elektronik	C27
Ambalaj	C28
Sos	C29
Cips-Kraker	C30

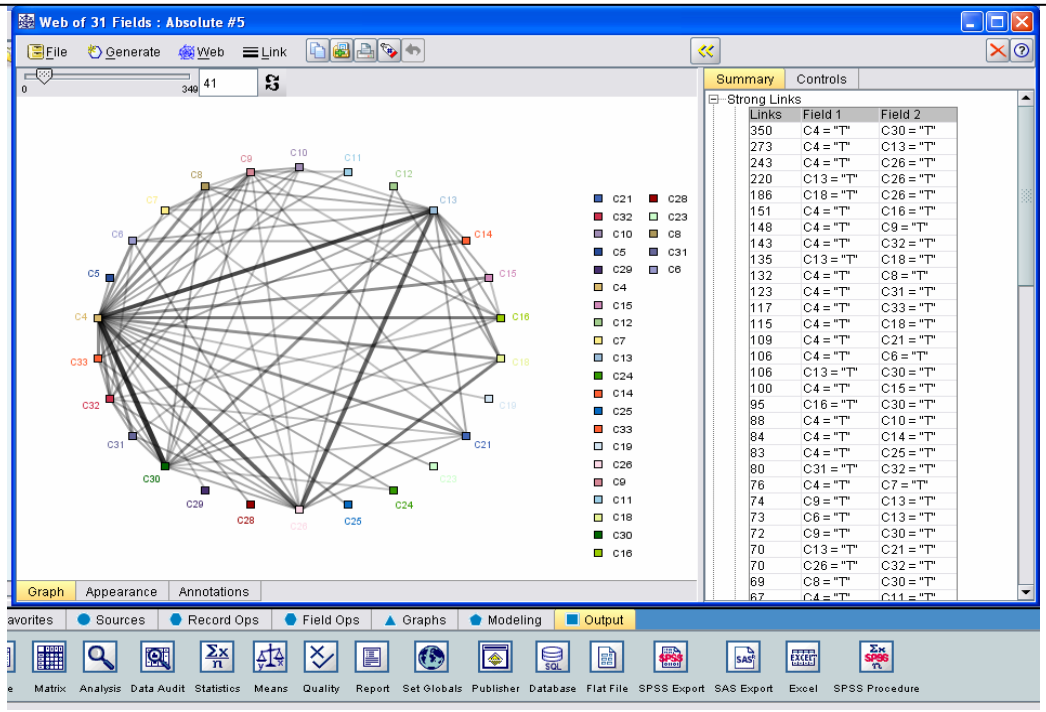
Tablo 12. devam

Ekmek	C31
Su	C32
Kağıt Temizlik	C33
Hazır Çorba	C34



Şekil 11. Birliktelik Analiz Sonuçları (Model ve Özet)

Şekil 12'de ürünler arası mutlak ilişkiler ortaya konmuş ve ilişkinin şiddeti ile çizgi kalınlığı doğru orantılı olarak gösterilmiştir.



Şekil 12. Birliktelik Analizinde Ürünler Arası İlişkiler

Bu veriler doğrultusunda;

Cips-Kraker ⇒Gazlı İçecekler kuralı için; Cips-Kraker ve Gazlı içecekler ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 21,911'dir. Cips-Kraker ürünlerini alan müşterinin % 79,909 olasılıkla Gazlı içecek ürünlerini de aldığı söylenebilir.

Et-Tavuk-Balık ⇒ Yumurta kuralı için; Et-Balık-Tavuk ürünleri ve yumurtanın toplam fiş hareketlerinde birlikte görülme olasılıkları % 13,207'dir. Et-Balık-Tavuk ürünleri alan müşterinin % 70,455 olasılıkla Yumurta da aldığı söylenebilir.

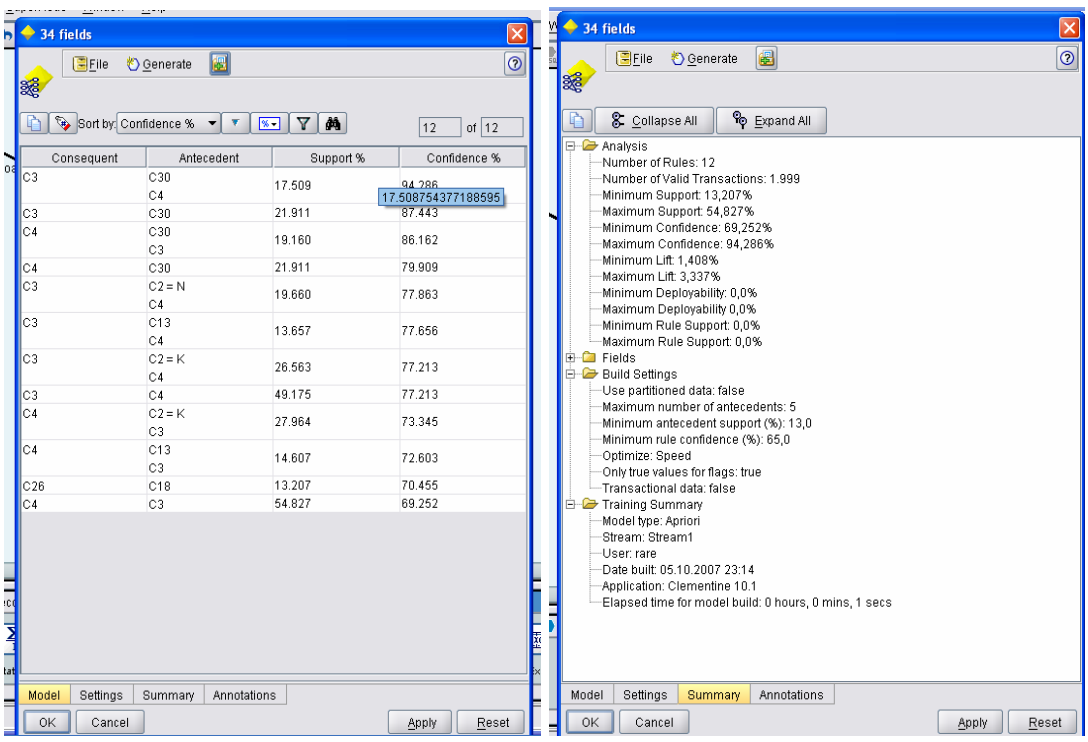
Süt ürünleri ve Yumurta ⇒ Gazlı İçecekler kuralı için; Süt ürünleri ve Yumurta ürünleri ile gazlı içecekler toplam fiş hareketlerinde birlikte görülme olasılıkları % 11,006'dir. Süt ürünleri ve Yumurta alan müşterinin % 68,636 olasılıkla Gazlı .içecek ürünlerini de aldığı söylenebilir.

Yumurta ve Gazlı içecekler ⇒ Süt ürünleri kuralı için; Yumurta ve Gazlı içecekler ürünleri ile Süt ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 12,156'dir. Yumurta ve Gazlı içecekler ürünlerini alan müşterinin % 62,140 olasılıkla Süt ürünleri de aldığı söylenebilir.

Meyve suyu ⇒ Gazlı İçecekler kuralı için; Meyve suyu Gazlı içeceklerin toplam fiş hareketlerinde birlikte görülme olasılıkları % 12,606'dir. Meyve suyu alan müşterinin % 59,921 olasılıkla Gazlı .içecek ürünlerini de aldığı söylenebilir.

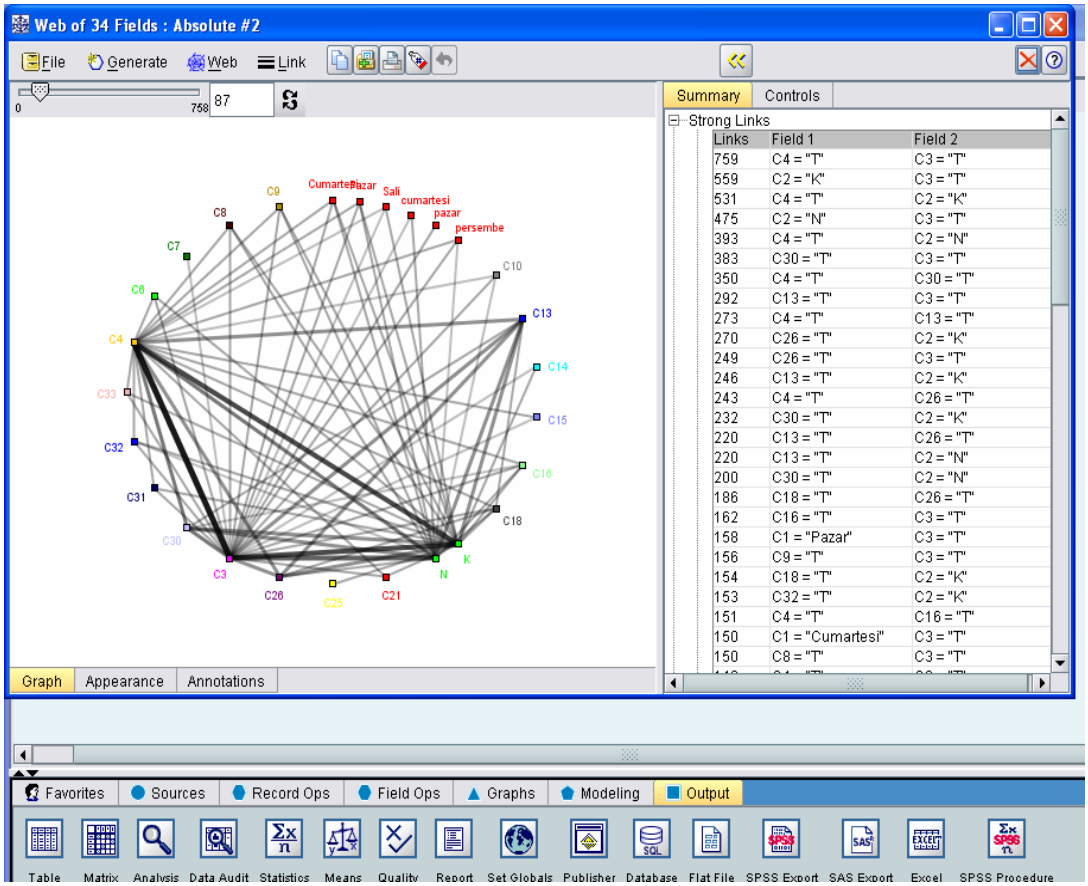
3.2.1.2. Tüm Özelliklerin Analizi

Tüm özelliklerin analizinde ürün grupları, ödeme tipi ve günü arasındaki birliktelik analizi tüm özellikler seçilerek yapılmış destek değeri alt sınırı yaklaşık % 13,207 ve güven değeri alt sınırı da %69,252 olarak seçilmiştir. Bunun sonucunda Şekil 11'deki 17 adet destek ve güven değerlerini gösteren kurallar elde edilmiştir.



Şekil 13. Tüm Özeliklerin Birliktelik Analiz Sonuçları

Şekil 14'de ürünler arası mutlak ilişkiler ortaya konmuş ve ilişkinin şiddeti ile çizgi kalınlığı doğru orantılı olarak gösterilmiştir. En fazla bağlantının gazlı içecekler ile Bisküvi-Kek-Çikolata-Tatlı arasında olduğu tespit edilmiştir.



Şekil 14. Birliktelik Analizinde Ürünler Arası İlişkiler

Bu veriler doğrultusunda;

Cips-Kraker ve Gazlı İçecekler ⇒ Bisküvi-Kek-Çikolata-Tatlı kuralı için; Gazlı içecekler, Cips-Kraker ürünleri ile Bisküvi-Kek-Çikolata-Tatlı ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 17,509'dir. Cips-Kraker ve Gazlı İçecekler ürünlerini alan müşterinin % 94,286 olasılıkla Bisküvi-Kek-Çikolata-Tatlı ürünlerini de aldığı söylenebilir.

Cips-Kraker ⇒ Bisküvi-Kek-Çikolata-Tatlı kuralı için; Cips-Kraker ve Bisküvi-Kek-Çikolata-Tatlı ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 21,911'dir. Cips-Kraker alan müşterinin % 87,443 olasılıkla Bisküvi-Kek-Çikolata-Tatlı ürünlerini de aldığı söylenebilir.

Cips-Kraker ve Bisküvi-Kek-Çikolata-Tatlı ⇒ Gazlı İçecekler kuralı için; Cips-Kraker, Bisküvi-Kek-Çikolata-Tatlı ürünleri ile Gazlı içeceklerin, toplam fiş hareketlerinde birlikte görülme olasılıkları % 19,160'dir. Cips-Kraker ve Bisküvi-Kek-Çikolata-Tatlı ürünlerini alan müşterinin % 86,162 olasılıkla Gazlı içecekleri de aldığı söylenebilir.

Cips-Kraker ⇒Gazlı İçecekler kuralı için; Cips-Kraker ve Gazlı içecekler ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 21,911'dir. Cips-Kraker ürünlerini alan müşterinin % 79,909 olasılıkla Gazlı içecek ürünlerini de aldığı söylenebilir.

Ödeme Tipi= Nakit ve Gazlı İçecekler ⇒ Bisküvi-Kek-Çikolata-Tatlı kuralı için; Ödeme Tipi nakit ve Gazlı içecekler ile Bisküvi-Kek-Çikolata-Tatlı ürünlerinin toplam fiş hareketlerinde birlikte görülme olasılıkları % 19,660'dır. Ödeme Tipi nakit ve Gazlı içecekleri alan müşterinin % 77,863 olasılıkla Bisküvi-Kek-Çikolata-Tatlı ürünlerini de aldığı söylenebilir.

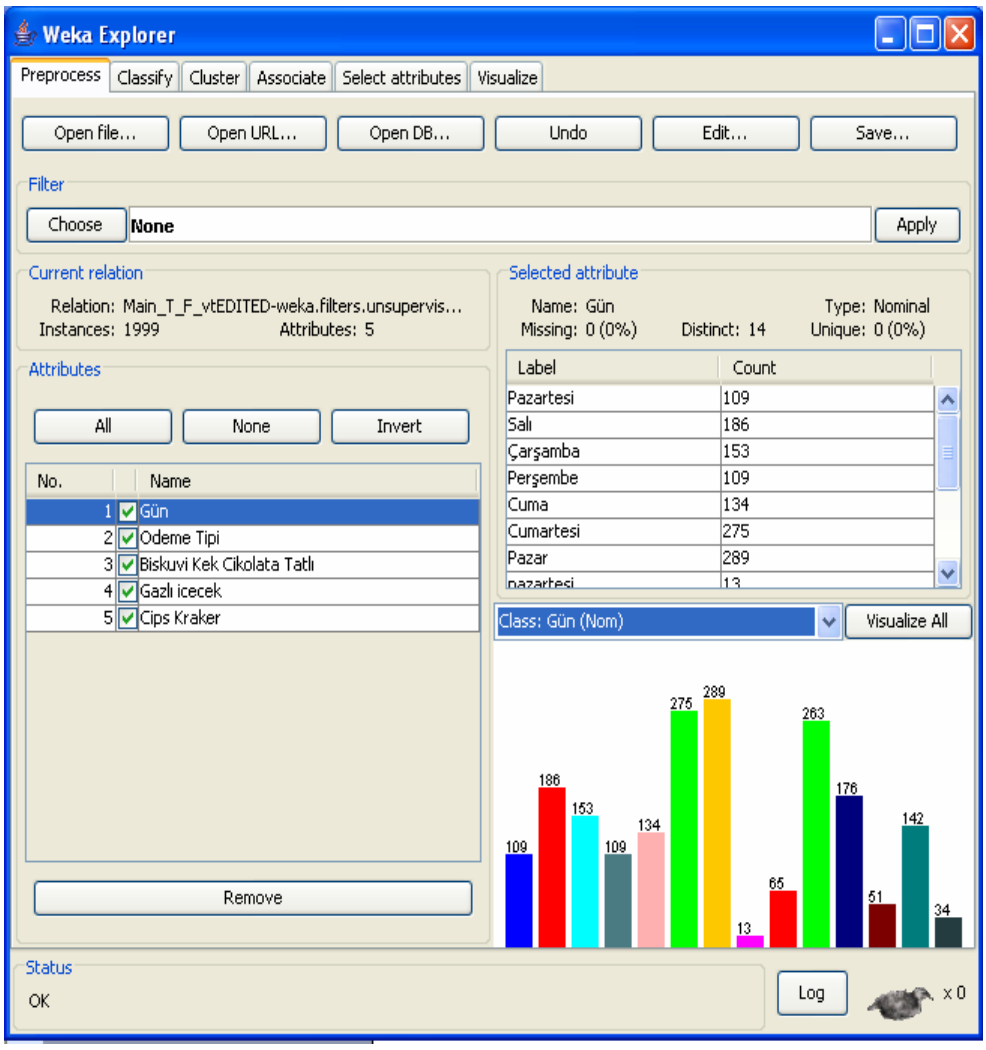
Bu çeşit birliktelik analizi tüm özellikler arasında istenilen kombinasyonlarda yapılarak farklı destek ve güven değerleri bulunarak değerlendirilir.

3.2.2. WEKA Programı

Weka makine öğrenme algoritmalarının ve veri ön işleme araçlarının bir araya getirildiği, akademik çevrelerde sıklıkla kullanılan, açık kaynak kodlu bir veri madenciliği programıdır. Yeni Zelanda'nın Waikato Üniversitesinde geliştirilmiştir. Yazılım, Java yazılım dili ile geliştirilmiştir. Büyük veya dağıtık veri tabanlarında kullanılabilir. Weka ile verinin hazırlanması, sınıflama, kümeleme, birliktelik analizi,

nitelik değerlerinin seçilmesi yapılabilmektedir. Arff, csv, c45 biçimindeki dosyalar kullanılabilmektedir.[71]

Weka programı ile tüm özellikleri kapsayan bir analiz söz konusu olduğunda içeriğinde "T" geçen özelliklerin yanı sıra "F" geçenlere göre de analiz yaptığından Apriori algoritmasında bulunan en iyi 10 kural arasına "F"lere oranla çok daha az tekrarlayan "T"ler giremediğinden Şekil 15'te görüldüğü gibi bazı özellikler seçilerek birliktelik analizi yapılmış ve sonuçlar Tablo 13'de olduğu gibi elde edilmiştir.



Şekil 15. Weka Programı Ara yüzü

[71] David Scuse ve Peter Reutemann, WEKA Experimenter Tutorial for Version 3-4, Yeni Zelanda Waikato Üniversitesi. 2007. s.1-6

Tablo 13. Weka Apriori Algoritması Birliktelik Analiz Sonucu

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     Main_T_F_vtEDITED-weka.filters.unsupervised.attribute.Remove-R5-29,31-34
Instances:    1999
Attributes:   5
              Gün
              Odeme Tipi
              Biskuvi Kek Cikolata Tatlı
              Gazlı icecek
              Cips Kraker
=== Associator model (full training set) ===
Apriori
=======

Minimum support: 0.3 (600 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 1

Best rules found:

 1. Biskuvi Kek Cikolata Tatlı=F Gazlı icecek=F 679 ==> Cips Kraker=F 644    conf:(0.95)
 2. Biskuvi Kek Cikolata Tatlı=F 903 ==> Cips Kraker=F 848    conf:(0.94)
 3. Gazlı icecek=F 1016 ==> Cips Kraker=F 928    conf:(0.91)
 4. Odeme Tipi=N 939 ==> Cips Kraker=F 739    conf:(0.79)
 5. Gazlı icecek=T 983 ==> Biskuvi Kek Cikolata Tatlı=T 759    conf:(0.77)
 6. Biskuvi Kek Cikolata Tatlı=F Cips Kraker=F 848 ==> Gazlı icecek=F 644    conf:(0.76)
 7. Biskuvi Kek Cikolata Tatlı=F 903 ==> Gazlı icecek=F 679    conf:(0.75)
 8. Odeme Tipi=K 930 ==> Cips Kraker=F 698    conf:(0.75)
 9. Biskuvi Kek Cikolata Tatlı=F 903 ==> Gazlı icecek=F Cips Kraker=F 644    conf:(0.71)
10. Gazlı icecek=F Cips Kraker=F 928 ==> Biskuvi Kek Cikolata Tatlı=F 644    conf:(0.69)
```

Burada örneğin birinci kurala bakıldığında Bisküvi-Kek-Çikolata-Tatlı ürünleri ile Gazlı içeceklerin birlikte rastlanmadığı durumun sonucu olarak Cips-Kraker bulunmama durumu güven değerinin %95 olduğu ve beşinci kurala göre de Gazlı içecekler önce gelen değerine göre Bisküvi-Kek-Çikolata-Tatlı ürünleri sonucu ortaya çıkma durumu güven değerinin %77 olduğu tespit edilmiştir.

Birliktelik analizi Predictive Apriori algoritmasına göre yapıldığında da Tablo 14'daki kurallar tespit edilmiştir.

Tablo 14. Weka Predictive Apriori Algoritması Birliktelik Analiz Sonucu

```
Scheme:       weka.associations.PredictiveApriori -N 10
Relation:     Main_T_F_vtEDITED-weka.filters.unsupervised.attribute.Remove-R5-29,31-34
Instances:    1999
Attributes:   5
              Gün
              Odeme Tipi
              Biskuvi Kek Cikolata Tatlı
              Gazlı icecek
              Cips Kraker
=== Associator model (full training set) ===


PredictiveApriori
==================


Best rules found:

 1. Gün=Salı Biskuvi Kek Cikolata Tatlı=F Gazlı icecek=F 68 ==> Cips Kraker=F 68    acc:(0.99395)
 2. Gün=perşembe Biskuvi Kek Cikolata Tatlı=T Cips Kraker=T 67 ==> Gazlı icecek=T 67    acc:(0.99391)
 3. Gün=pazar Gazlı icecek=T Cips Kraker=T 37 ==> Biskuvi Kek Cikolata Tatlı=T 37    acc:(0.98992)
 4. Gün=Cuma Odeme Tipi=N Gazlı icecek=F 35 ==> Cips Kraker=F 35    acc:(0.98928)
 5. Gün=Salı Odeme Tipi=N Gazlı icecek=F 34 ==> Cips Kraker=F 34    acc:(0.98893)
 6. Gün=cuma Gazlı icecek=T 29 ==> Biskuvi Kek Cikolata Tatlı=T 29    acc:(0.9867)
 7. Gün=perşembe Cips Kraker=T 70 ==> Gazlı icecek=T 69    acc:(0.98666)
 8. Gün=cuma Odeme Tipi=K 28 ==> Biskuvi Kek Cikolata Tatlı=T 28    acc:(0.98614)
 9. Gün=Çarşamba Odeme Tipi=C 26 ==> Cips Kraker=F 26    acc:(0.98485)
10. Odeme Tipi=C Biskuvi Kek Cikolata Tatlı=T 62 ==> Cips Kraker=F 61    acc:(0.98316)
```

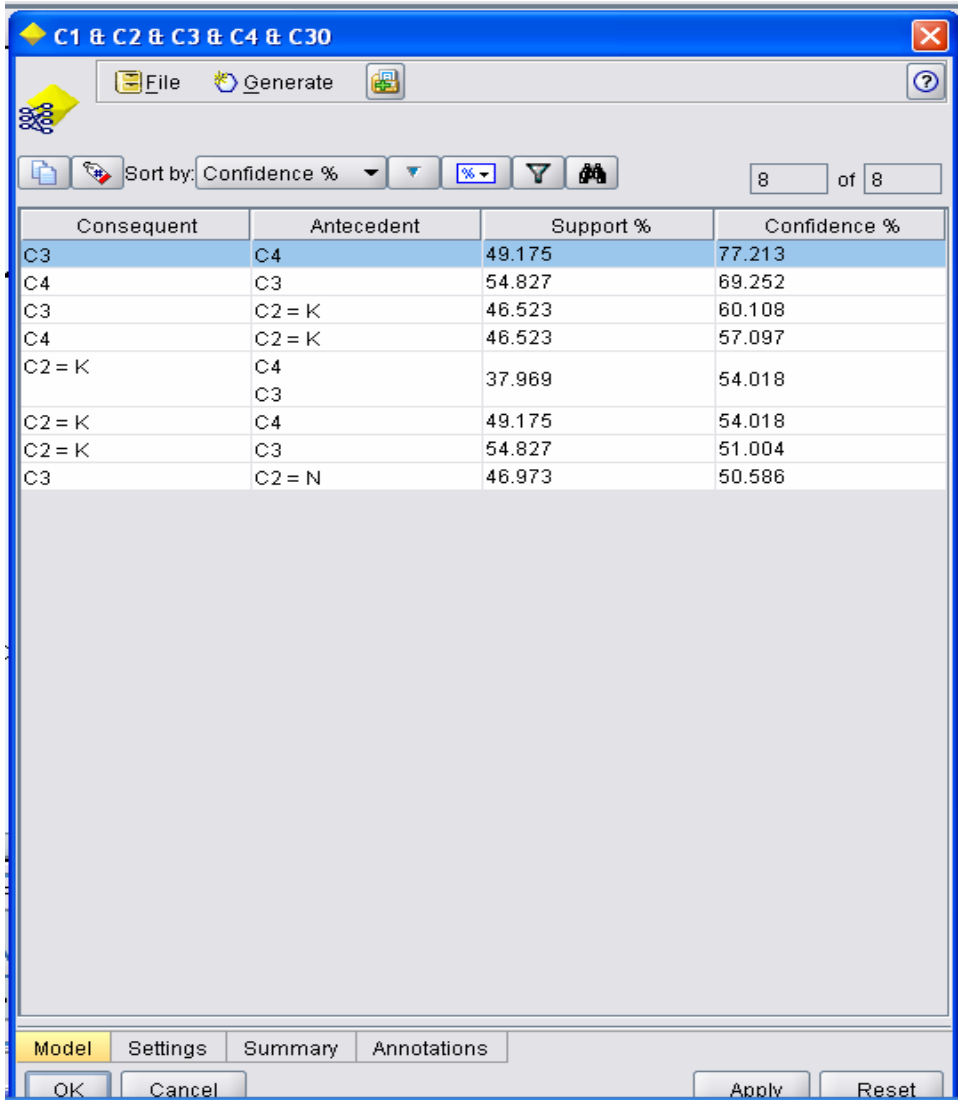
Buna göre örneğin ikinci kural değerlendirildiğinde günün perşembe olması, Bisküvi-Kek-Çikolata-tatlı ve Cips-Kraker birliktelik şartının sonucu Gazlı içecekler olma ihtimalinin doğruluk değerinin %99 olduğu tespit edilmiştir.

Birliktelik analizi Tertius algoritmasına göre yapıldığında da Tablo 15'daki kurallar tespit edilmiştir.

Tablo 15. Weka Tertius Algoritması Birliktelik Analiz Sonucu

```
=== Run information ===

Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -l 0 -p -P 0
Relation:     Main_T_F_vtEDITED-weka.filters.unsupervised.attribute.Remove-R5-29,31-34
Instances:    1999
Attributes:   5
              Gün
              Odeme Tipi
              Biskuvi Kek Cikolata Tatlı
              Gazlı icecek
              Cips Kraker
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0,462773 0,138569 */ Gazlı icecek = F ==> Gün = cumartesi or Biskuvi Kek Cikolata Tatlı = F
 2. /* 0,455288 0,118559 */ Gazlı icecek = F and Cips Kraker = F ==> Gün = cumartesi or Biskuvi Kek Cikolata Tatlı = F
 3. /* 0,452358 0,140070 */ Biskuvi Kek Cikolata Tatlı = T ==> Gün = cuma or Gazlı icecek = T or Cips Kraker = T
 4. /* 0,450357 0,140570 */ Biskuvi Kek Cikolata Tatlı = T ==> Gün = pazartesi or Gazlı icecek = T or Cips Kraker = T
 5. /* 0,449858 0,142071 */ Gazlı icecek = F and Cips Kraker = F ==> Biskuvi Kek Cikolata Tatlı = F
 6. /* 0,448837 0,095048 */ Biskuvi Kek Cikolata Tatlı = F ==> Gün = cumartesi or Gazlı icecek = F
 7. /* 0,448083 0,137569 */ Biskuvi Kek Cikolata Tatlı = T ==> Gün = salı or Gazlı icecek = T or Cips Kraker = T
 8. /* 0,444087 0,166583 */ Biskuvi Kek Cikolata Tatlı = T ==> Gün = cuma or Gazlı icecek = T
 9. /* 0,443845 0,166083 */ Biskuvi Kek Cikolata Tatlı = T ==> Gün = pazartesi or Gazlı icecek = T
10. /* 0,441768 0,112056 */ Biskuvi Kek Cikolata Tatlı = F ==> Gazlı icecek = F

Number of hypotheses considered: 5128
Number of hypotheses explored: 2954
Time: 00 min 07 s 828 ms
```

Weka Programında seçilen aynı özellikler, destek ve güven değerleri ile yapılan Apriori SPSS Clementine Programı birliktelik analizi bir fikir oluşturması için Şekil 16'te gösterilmiştir.



Şekil 16. SPSS Clementine 10.1 Programı Bazı Özellikler Birliktelik Analizi

Buna göre "T" değerleri ele alınarak değerlendirildiğinde Weka Apriori algoritması birliktelik analizindeki beşinci kural ile SPSS Clementine 10.1 Apriori birliktelik analizi birinci kuralının koşul şart ve güven değerlerinin aynı olduğu tespit edilmiştir.

4. SONUÇ

Verilerin ve veri işleme araçlarının gelişmesi sonucunda gerek ticari gerekse akademik alandaki çalışmalar bu yönde hız kazanmış ve veri analizi birçok alandan daha önemli hale gelmiştir. Sayısal ortamda biriken verilerin artık kullanılmalarının gerekliliği fark edilmeye başlanmış ve geçmiş bilgiler değerlendirilmeye alınmıştır. Özellikle gelecekteki verilerin tahmin için kullanılan bu veriler üzerinde işlem yapmak üzere birçok bilim dalı bir araya gelerek çalışmalar gerçekleştirmişlerdir.

Veri madenciliği, bu noktada çok disiplinli bir kavram olarak ortaya çıkmış ve hızla popülerlik kazanmıştır. Birçok alanda uygulanabilir olması ve ortaya koyduğu sonuçların ilgili faaliyet alanına katkılarının fark edilmesi de yaygınlaşmasında büyük rol oynamıştır.

Veri madenciliği çalışmaları, perakendecilik sektöründe müşterilerin ihtiyaçlarının tespiti ve onlara uygun çözümler bulma, sunulan hizmetlerin geliştirilmesi bunlara dayalı olarak yapılacak yatırımların ve stratejilerin tespitinde, günümüzde karar verme sürecindeki yer alanlara oldukça yardımcı olmaktadır.

Yoğun rekabet ortamında çalışan işletmeler kendilerini farklılaştırabilmek için bilgiyi etkin ve verimli bir biçimde kullanma yolunda veri madenciliğini tercih etmektedirler. Bilgiyi elde etmek için eskiden beri kullanılmakta olan yöntemlerin en önemlilerinden biri olan istatistiksel yöntemler, hak ettikleri öneme kavuşmaktadırlar. Sağladığı faydayı gören iş dünyası istatistik uygulamalarını yeniden keşfetmeye başlamışlardır.

Bu maksatla mevcut market alışverişi veri tabanı üzerinde veri madenciliği birliktelik analiz tekniği SPSS Clementine 10.1 programı Apriori ve Weka programı Aprori, Predictive Apriori ve Tertius algoritmaları kullanılarak belirli güven ve destek değerlerine göre yapılmıştır. Her bir analizin sonuçları ayrı ayrı ve birlikte incelenerek değerlendirilmiş ve sonuçları ortaya konmuştur.

Yapılan bu çalışmanın daha geniş bir dönemi kapsamasının, özellik sayılarının istenilen sonuçlara göre belirlenerek özel hale getirilmesinin, daha fazla ürün çeşidini ihtiva ederek özel sonuçların bulunmasının, daha büyük veri tabanları ile ilgili analizlerin yerel bilgisayarlardan daha çok paylaşılan ortamdaki bilgisayarların performanslarının birleştirilerek bellek ve işlemci sorunu olmadan çok daha kısa sürelerde yapılmasının ve elde edilen sonuçların perakendecilik sektöründe yaygın olarak kullanılmasının faydalı olacağı değerlendirilmektedir.

KAYNAKÇA

Kitaplar

ADRIAANS, P.ve ZANTINGE, D. **Data Mining**. Longman, Harlow: Addison Wesley. 1996.

HAN, J.ve KAMBER, S.F., **Data Mining: Concepts and Techniques.** San Francisco: Morgan Kaufmann Publishers. 2001.

LINOFF G. ve BERRY M.J.A.,. **Data Mining techniques For Marketing Sales and Customer Relationship Management,** , New York: Wiley Publishıng. 2004.

PIATETSKY-SHAPIRO, G.. **Discovery, Analysis, and Presentation of Strong Rules. in Knowledge Discovery in Databases.** Menlo Park, CA: AAAI/MIT Press. 1991.

WITTEN I.ve FRANK E. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**. San Fransisco: Morgan Kaufmann Publishers. 2000.

Süreli Yayınlar

AGRAWAL, R., IMIELINSKI, T. ve SWAMI, A.. "Mining Association Rules Between Sets Of İtems In Large Databases". **ACM SIGMOD Conference on Management of Data**. Washington, DC: ACM Press. 1993 ss. 207-216..

AGRAWAL, R.ve SRIKANT, R. "Fast Algorithms For Mining Association Rules". **In Proceedings of the 20th International Conference on Very Large Databases.** San Francisco: Morgan Kaufmann. 1994 ss. 487–499.

AKPINAR, H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İ.Ü. İşletme Fakültesi Dergisi**, c: 29, 1, s: 1-22. (2000).

APORTA, G. "Data Mining and Official Statistics". **Quinta Conferenza Nationale di Statistica,** ISTAT. Roma. 2000. ss. 15-17.

AUMANN, A., ve LINDELL, Y. "A Statistical Theory For Quantitative Association Rules". **In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining**.. New York:ACM Press. 1999. ss. 261–270.

BAYARDO, R. J., JR. "Efficiently Mining Long Patterns from Databases". **In Proceedings of ACM'SIGMOD International Conference on Management of Data**. New York: ACM Press. ss. 85–93. 1998.

BAYARDO, R. J., JR.ve AGRAWAL, R. "Mining The Most Interesting Rules". **In Proceedings of the Fifth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining**. New York: ACM Press ss. 145–154. 1999.

BAYARDO, R. J., JR., AGRAWAL, R. ve GUNOPULOS, D.. "Constraint-Based Rule Mining in Large, Dense Databases". **Data Mining and Knowledge Discovery 4**. ss 217–240. 2000.

BELLAZI R., LARIZZA C. ve MAGNI P., "Temporal data mining for the quality assessment of hemodialysis services", **Artificial Intelligence in Medicine**, 34 (1): 2005. s 25-39.

BOGINSKI V., BUTENKO S. ve PARDALOS P., "Mining market data: A network approach", C**omputers & Operations Research,** 33 (11). 2006 s 3171-3184 .

BRIN, S., MOTWANI, R. ve SILVERSTEIN, C. "Beyond Market Baskets: Generalizing Association Rules to Correlations, Proceedings of the 1997 ACM SIGMOD". **International Conference on Management of Data**. New York, NY, USA 1997. ss. 265-276.

CASKEY K., "A manufacturing problem solving environment combining evaluation, search, and generalization methods", **Computers in Industry**, 44 (2) .2001 s 175-187.

CERVONE G., KAFATOS M., NAPOLETANI D. ve SINGH R., "An early warning system for coastal earthquakes", **Advances in Space Research,** 37 (4) .2006. s 636-642.

CHEN M, HUANG C., CHEN K. ve WU H., "Aggregation of orders in distribution centers using data mining", **Expert Systems with Applications,** 28 (3) .2005. s 453-460.

COHEN, E., DATAR, M., FUJIWARA, S., GIONIS, A., INDYK, P., MOTWANI, R., ULLMAN, J., ve YANG, C. "Finding Interesting Associations without Support Pruning". **In Proceedings of the International Conference on Data Engineering New York**: IEEE Press. 2000. ss. 489–499.

COX I., LEWIS R., RANSING R., LASZCZEWSKI H., BERNI G., "Application of neural computing in basic oxygen steelmaking", **Journal of Materials Processing Technology,** 120 (1): 310-315 (2002).

CRESPO F. ve WEBER R., "A methodology for dynamic data mining based on fuzzy clustering", **Fuzzy Sets and Systems**, 150 (2) .2005. s 267-284.

SCUSE, David ve REUTEMANN,Peter, **WEKA Experimenter Tutorial for Version 3-4**, Yeni Zelanda Waikato Üniversitesi, 2007, s.1-6

FACCA F.ve LANZI P., "Mining interesting knowledge from weblogs: a survey", **Data & Knowledge Engineering,** 53 (3). 2005. s 225-241.

FRIEDMAN, J. H. ve FISHER, N. I., "Bump hunting in high-dimensional data", **Statistics and Computing, 9**. 1999. s 123–143.

GANTI, V., GEHRKE, J. ve RAMAKRISHNAN, R., "Mining Very Large Databases", **IEEE Computer**, 32, 8. 1999. pp. 38-45.

GUHA S., RASTOGI R. ve SHIM K., "Rock: A robust clustering algorithm for categorical attributes", **Information Systems,** 25 (5). 2000 s 345-366.

GOEBEL, M. ve GRUENWALD, L., "A Survey of Data Mining and Knowledge Discovery Software Tools", **ACM SIGKDD Explorations Newsletter**, 1, 1. 1999. s. 20-33.

GOUDA, K. ve ZAKI, M. J.. "Efficiently mining maximal frequent itemsets". **In Proceedings of the First IEEE International Conference on Data Mining** New York: IEEE Press. 2001.ss. 163–170.

HAN, J., PEI, J., ve YIN, Y. "Mining Frequent Patterns Without Candidate Generation". **In Proceedings of the ACM SIGMOD International**

Conference on Management of Data SIGMOD. New York: ACMPress. 2000. ss. 1–12.

HONG G., PARK S., JANG D. ve RHO H., "An effective supplier selection method for constructing a competitive supply relationship", **Expert Systems with Applications,** 28 (4) 2005. s 629-639.

JIAO J., ZHANG Y. ve HELANDER M., "Kansei mining system for affective design", **Expert Systems with Applications,** 30 (4). 2006. s 658-673.

JENG B., CHEN J. ve LIANG T., "Applying data mining to learn system dynamics in a biological model", **Expert Systems with Applications,** 30 (1). 2006. s 50-58.

LAST M. ve KANDEL A., "Discovering useful and understandable patterns in manufacturing data", **Robotics and Autonomous Systems**, 49 (3). 2004. s 137-152.

LEE T., CHIU C., CHOU Y. ve LU C., "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", **Computational Statistics & Data Analysis,** 50 (4) 2006. s 1113-1130.

LIAN J., LAI M., LIN Q., YAO F., "Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis", **Journal of Materials Processing Technology,** 129 (1): 315-320 (2002).

LIN F. ve MCCLEAN S., "A data mining approach to the prediction of corporate failure", **Knowledge-Based Systems**, 14 (3). 2001. s 189-195.

LIU, B., HSU, W., CHEN, S. ve MA, Y., "Analyzing the subjective interestingness of association rules", **IEEE Intelligent Systems,** 15. 2000. s 47–55.

MADRIA, S.K., BHOWMICK, W.K.ve NG, E.P., "Research Issues in Web Data Mining". **In Proceedings of Data Warehousing and Knowledge Discovery", First International Conference**. DaWak1999. ss 303-312

PASQUIER, N., BASTIDE, Y., TAOUIL, R., ve LAKHAL, L. "Discovering Frequent Closed Itemsets for Association Rules". **In Proceedings of the Seventh International Conference on Database Theory** .. New York: Springer. 1999. ss. 398–416

PEI, J., HAN, J.,VE MAO, R. "CLOSET:An Efficient Algorithm for Mining Frequent Closed Itemsets". **In Proceedings of the ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery**. New York: ACM Press. 2000 ss. 21–30.

PEI, J., HAN, J., LU, H., NISHIO, S., TANG, S., ve YANG, D. "H-MINE: Hyper-Structure Mining Of Frequent Patterns in Large Databases". **In Proceedings of the First IEEE International Conference on Data Mining**. New York: IEEE Press. 2001. ss 441–448.

PIRAMUTHU S. “Evaluating Feature Selection Methods For Learning in Data Mining Applications”. **Thirty- First Annual Hawai International Conference on System Sciences,** 5. 1998. ss294.

SFORNA M., “Data mining in a power company customer database”, **Electric Power Systems Research,** 55 (3). 2000. s 201-209.

SHEARER C., “THE CRISP-DM model: The new blueprint for data mining”, **Journal of Data Warehousing,** 5 (4). 2000. s 13-23.

SILBERSCHATZ, A., ve TUZHILIN, A., “What makes patterns interesting in knowledge discovery systems”, **IEEE Transactions on Knowledge and Data Engineering,** 8. 1996. s 970–974.

SRIKANT, R. ve AGRAWAL, R. “Mining quantitative association rules in large relational tables”. **In Proceedings of the ACM ’SIGMOD International Conference on Management of Data**. New York: ACM Press. 1996. ss. 1–12.

WEBB, G. I.. “Efficient Search for Association Rules”. In Proceedings of the Sixth ACM SIGKDD **International Conference on Knowledge Discovery and Data Mining.** New York: ACM Press. 2000. ss. 99–107.

ZAKI, M. J., “Parallel and Distributed Association Mining: A Survey, IEEE Concurrency, special issue on Parallel Mechanisms for Data Mining”, **IEEE** Vol. 7, No. 5. December 1999. s 14-25

Diğer Yayınlar

AKPINAR H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği". 2000. http://www.isletme.istanbul.edu.tr/dergi/nisan2000/1.htm. (Erişim Tarihi:30.08.2007).

AYDOĞAN F. "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi". Yüksek Lisans Tezi. Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. Ankara. 2003.

BİÇEN, P. "Veri Madenciliği: Sınıflandırma Ve Tahmin Yöntemlerini Kullanarak Bir Uygulama". Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.

KAYA, E.; BULUN, M. ve ARSLAN, A. "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları". 2006 http://www.ab.org.tr/ab03/program/96.html (Erişim Tarihi:18.07.2007).

MAINDONALD, J., "Data Mining from a Statistical Perspective. Preprint". Australian Nat. Univ., Stat. Cons. Unit. http://www.maths.anu.edu.au/~johnm/dm/dmpaper.html. (Erişim Tarihi:"15.06.2007).

ÖZEKEŞ, S. "Veri Madenciliği Uygulaması". Yüksek Lisans Tezi. Marmara Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.

ÖZMEN, Ş.,. "İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor". http://www.mimoza.marmara.edu.tr/~sozmen/teblig.php (Erişim Tarihi:10.09.2007)

RUSHING, J. Technology Assessment Paper. 1997. http://www.cs.uah.edu/~thinke/CS 687/Fall 97/Tech/Rushing.html. (Erişim Tarihi:03.07.2007)

TANTUĞ, A.C. "Veri Madenciliği ve Demetleme". Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü. İstanbul. 2002.