KOCAELİ ÜNİVERSİTESİ * FEN BİLİMLERİ ENSTİTÜSÜ

POZİTİF VE NEGATİF İLİŞKİLERİN VERİ MADENCİLİĞİYLE BELİRLENMESİNE YÖNELİK BİR MODEL

YÜKSEK LİSANS TEZİ

Endüstri Müh. Ahmet CİHAN

Ana Bilim Dalı: Endüstri Mühendisliği

Danışman: Prof. Dr. Alpaslan FIĞLALI

KOCAELİ, 2009

KOCAELİ ÜNİVERSİTESİ * FEN BİLİMLERİ ENSTİTÜSÜ

POZİTİF VE NEGATİF İLİŞKİLERİN VERİ MADENCİLİĞİYLE BELİRLENMESİNE YÖNELİK BİR MODEL

YÜKSEK LİSANS TEZİ

Endüstri Müh. Ahmet CİHAN

Tezin Enstitüye Verildiği Tarih: 05 Haziran 2009

Tezin Savunulduğu Tarih: 09 Temmuz 2009

Tez Danışmanı	Üye	Üye
Prof.Dr.Alpaslan Fığlalı	Yrd.Doç.Dr.Kasım Baynal	Yrd.Doç.Dr.A. Serhat Demir
(.....)	(.....)	(.....)

KOCAELİ, 2009

ÖNSÖZ VE TEŞEKKÜR

Veri depolama teknolojisindeki hızlı gelişme saklanan veri sayısının artışı ile sonuçlanmış olmasına rağmen, karar vericiye karar verme sürecinde destek olan anlamlı bilgi miktarındaki artış aynı oranda olmamıştır. Veri Madenciliği(V.M.) büyük ölçekli verileri analiz ederek veriler içinde saklı kalmış, karar vericinin kullanabileceği anlamlı bilgi ihtiyacına cevap veren yorumlama sürecidir. İşletmelerin pazar paylarını arttırmasında ve fiyatlarını belirlemesinde mevcut veriler kullanılarak yapılan çıkarımlar kullanılmaktadır. Müşterilerin davranışları ürünlerin birlikte alınması ile sonuçlanabileceği gibi bir ürünün alınması ile başka bir üründen vazgeçilmesi ile de sonuçlanabilir. Yapılan çalışmada bu ürün gruplarının tespiti için bir model önerilmiştir.

Tez çalışmamda her türlü desteğini hiçbir zaman esirgemeyen danışman hocam sayın Prof. Dr. Alpaslan FIĞLALI' ya teşekkürlerimi sunarım. Ayrıca değerli hocam Doç.Dr. Ayhan DEMİRİZ' e bilimsel katkılarından dolayı teşekkürü bir borç bilirim.

Çalışmalarım sırasında kendisinden çok şey öğrendiğim hocam Dr. Müh. Ümit TERZİ' ye teşekkür ederim. Maddi ve manevi desteklerini hiçbir zaman esirgemeyen ve bugünlere gelmemde büyük pay sahibi olan aileme sonsuz teşekkür ederim. Tezin her aşamasında desteğini esirgemeyen kardeşim Onur CİHAN' a teşekkür ederim.

Çalışmalarım sırasında bana gerekli çalışma ortamını sağlayan değerli hocalarıma teşekkürlerimi sunarım.

Bu çalışma TÜBİTAK tarafından 107M257 nolu araştırma projesi kapsamında desteklenmiştir.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

KISALTMALAR

V.M.: Veri Madenciliği
TDA: Tekil Değerlere Ayrıştırma
SMC: Simple Matching Coefficient
EJ: Extended Jaccard

POZİTİF VE NEGATİF İLİŞKİLERİN VERİ MADENCİLİĞİYLE BELİRLENMESİNE YÖNELİK BİR MODEL

Ahmet CİHAN

Anahtar Kelimeler: Veri Madenciliği(V.M), Sepet Analizi, Tekil Değerlere Ayrıştırma(T.D.A), Pozitif İlişki, Negatif İlişki.

Özet: Veri Madenciliğinde, Tekil Değerlere Ayrıştırma yöntemi matrislerin özetlenmesi amacıyla sıklıkla kullanılmaktadır. Tekil Değerlere Ayrıştırma yöntemi ile bulunan özdeğerlere pareto analizi uygulanarak hangi özvektörlerin kuralları oluşturmakta kullanılacağı tespit edilmiştir. Bu kurallar üzerinde mevcut yapıya uygun benzerlik ölçütlerinin kullanımı ile pozitif ve negatif ilişkilerin bulunmasına çalışılmıştır. Bulunan pozitif ve negatif ilişkiler, karar vericinin kuracağı modellerde kullanılabilecektir.

A MODEL FOR DETERMINING POSITIVE AND NEGATIVE RELATIONS USING DATA MINING

Ahmet CİHAN

Keywords: Data Mining, Market Basket Analysis, Singular Value Decomposition, Positive Association, Negative Association.

Abstract: Singular value decomposition technique is being widely used for summarizing matrices in data mining. The eigenvectors that will be used to construct the rules are determined by applying pareto analysis to the eigenvalues derived by singular value decomposition. Positive and negative associations are tried to be found by using similarity measures that are suitable for the existing structure to the rules. These positive and negative associations can be used by decision maker for model construction.

1.GİRİŞ

Firmalar gerek mühendislik gerek işletme çalışmaları için kullanılması muhtemel verileri saklamaya çalışmaktadırlar. Teknolojik gelişmeler doğrultusunda veri toplama ve saklama süreçleri kolaylaşmıştır. Buna karşın toplanan verilerin içerisinden yararlı bilgileri ayıklamak gerekmektedir.

Veri madenciliği, elektronik ve bilgisayar sistemlerinin hızlı gelişimi sonucu saklanabilen verilerin içerisinden işe yarar bilgilerin çıkartılması için kullanılan yöntemler bütünüdür. Elde edilen bilgilerin mühendislik ve işletme çalışmaları için kullanılması mümkün olmaktadır.

Çalışmada firmaların sattığı ürünler arası ilişkilerin bulunması için harcama veya fiyat bilgilerini dikkate alan bir yöntem geliştirilmiştir. Birçok alanda kullanılan sepet analizi tekniğinin en önemli eksiği fiyat veya harcama verilerinin kullanılmaması durumudur. Bu durum, ilişkilerin fiyata dair bir bilgi kullanılmadan bulunması ve fiyatın, müşteri tercihleri üzerindeki etkilerinin göz ardı edilmesi anlamına gelmektedir. TDA yöntemi ile harcama matrislerine ait verilerin çözümlenmesi ve müşterilerin davranışlarının tahmin edilmesi mümkün olmaktadır (Korn ve diğ., 2000).

Uygulamada bir perakende hazır giyim firmasına ait veriler kullanılmıştır. Bulunan pozitif ve negatif ilişkiler yorumlanmıştır.

2.VERİ MADENCİLİĞİ

2.1. Veri Madenciliğinin Tanımı

En çok kullanılan veri madenciliği tanımına göre veri madenciliği, büyük ölçekli veri yığınları içerisinden bilgiye ulaşma işidir (Wikipedia, 2009).

Veri madenciliğinin amacı veri yığınından kullanılabilir bilgi elde etmektir. Bu bilginin doğru, anlaşılır ve ilginç olması gerekmektedir. İlginçlikten kastedilen, keşfedilen bilginin kullanıcı için yeni, şaşırtıcı ve kullanışlı olmasıdır (Freitag, 2002). Veri yığını içerisinde değersiz yapılar da bulunmaktadır ve değerli olanlara ulaşabilmek için değersiz yapıların ayıklanması gerekmektedir (Berson ve diğ., 2000). Ayrıca veri madenciliğinin etkili kullanımı ile projelerde maliyetler azaltılıp, gelirler artırılabilir (Javovic ve diğ., 2002). Veri madenciliği hipotezleri keşfeder, sonuçları birleştirmek için insan yeteneğini kullanır (Davis,1999).

Veri madenciliği; çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir (Swift, 2001).

Veri madenciliği istatistik, yapay zekâ ve veri tabanlarında bilgi keşfi süreçlerini kullanan bir disiplindir.

2.2. Veri Madenciliğinin Özellikleri

Veri madenciliği tanımlarında belirgin olan noktalar şunlardır:

- Oldukça büyük ve karmaşık verilerin tutulduğu veritabanları ile çalışır.
- Çok farklı türlerde verileri kullanarak çözümler üretebilir.

- İstatistik, yapay zeka, makine öğrenmesi, veri tabanlarında bilgi keşfi, bilgisayar bilimi, yapı tanıma vb. gibi çeşitli disiplinlerden faydalanır.
- Önceden bilinmeyen, doğrulanabilir, etkinleştirilebilir ve amaç doğrultusunda kullanılabilir haberleşme ve bilgi arar.
- Çıkarım mekanizmasının otomatik veya yarı otomatik olarak çalışması gerekmektedir.
- Birçok endüstride farklı biçimlerde amaca yönelik olarak kullanılmaktadır.
- Farklı sorunlara farklı çözüm araçları mevcuttur.
- Sektör büyük bir hızla büyümeye devam etmektedir.

Veri madenciliği bilgi çıkarımı süreci ile iç içedir. Bilgi çıkarımı süreci şu şekildedir:

- Veri temizleme.
- Veri bütünleştirme.
- Veri seçme.
- Veri dönüşümü.
- Veri madenciliği.
- Örüntü değerlendirme.
- Bilgi sunumu.

Esasen veri madenciliği adımına kadar olan bütün adımlar veri madenciliği kapsamına dahildir. Veri madenciliği, gizli kalmış ilişkiler bulunana kadar devam etmektedir. İlişkilerin bulunması için kullanılabilecek çok sayıda model mevcuttur.
Bir veri madencisi için verilerin ne anlam ifade ettiği çok önemli değildir. Bir istatistikçi, anlamsız olarak görünen veriler ile ilgilenmeyebilir. Ancak veri madencisi için aynı veriler önem taşımaktadır. Veri madencisi, ilişkileri anlamsız gibi görünen verilerde de bulabilir. Bu durum veri madencisi ile istatistikçiyi ayıran en önemli durumdur.

2.3. Veri Madenciliğinin Tarihçesi

İşletmeler geçmişte ellerinde bulunan verileri en iyi şekilde kullanmaya çalışmış, bunun için de çeşitli yöntemler geliştirmişlerdir. Bu yöntemler verilerin taşınabilir

olmasını sağlamıştır. Verilerin toplanması işi 90' lı senelere kadar mevcut verilerin veritabanlarından sadece okunmasını sağlamış, çıkarım işlemi ise karar vericiye bırakılmıştır. Gelişimin tarihçesi Tablo 2.1 ile gösterilmektedir (Aldana, 2000).

Tablo 2.1: Veri madenciliği gelişimi (Aldana, 2000)

Gelişim Adımları	**Cevaplanan Karar Problemi**	**Kullanılabilen Teknolojiler**	**Ürün Sağlayıcıları**	**Karakteristikler**
Veri toplama (1960'lar)	"Benim toplam karım geçen yılda ne kadar arttı?"	Bilgisayarlar, Teypler, Diskler	IBM,CDC	Geriye dönük,statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel veritabanları, SQL,ODBC.	Oracle,Sybase, Informix IBM,Microsoft,	Kayıt düzeyinde geriye dönük dinamik veri dağıtımı.
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	OLAP,Çok boyutlu veritabanı sistemleri, Veri ambarları	Pilot,comshare,arbor, Cognos,microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'da ki birim satışlar muhtemelen ne olabilir, niçin?"	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot,Lockheed,IBM, SGI,SPSS Clementine, SAS,Microsoft v.s.	Geleceğe dönük,proaktif, enformasyon dağıtımı

1960'lı yıllar ve öncesinde başlayan veri toplama çalışmaları basit dosyalama işlemleriydi. Bu veriler yardımıyla karar problemlerine sınırlı zaman diliminde cevap verilebilmekteydi. Bu yıllarda sadece geriye dönük aranılan veriye ulaşılırken bu verilerden enformasyon elde edilmediği görülmektedir (Bilen, 2004).
1980'li yıllara girildiğinde ilişkisel veri tabanlarının oluşturulmaya başlanmış, SQL ve ODBC ile veri kaynaklarına ulaşım bu yıllarda gerçekleşmiştir. Ürün sağlayıcılardaki artış dikkat çekicidir (Bilen, 2004).

1990'lı yıllarda veri saklama ortamlarının hızlı gelişiminin ve ucuzlamasının sonucu olarak çok büyük miktarlarda veri saklanabilen veri ambarları oluşturulmaya başlanmış ve bu veri ambarlarından elde edilebilecek bilgiler ile karar vericiye destek sağlayacak olan karar destek sistemleri kurulmaya başlanmıştır. Değişen verilerin

farklı merkezler tarafından sorgulanmasının sağlanması yönünde büyük adımlar atılmıştır. OLAP ve çok boyutlu veri tabanları göze batan değişimleridir (Bilen, 2004).

Bugün ise veri madenciliği tam anlamı ile kullanılmaya başlanmış olup geriye dönük yapılabilen veri değerlendirmelerine ek olarak ileriye yönelik tahminlere imkan veren bilgi keşfi de yapılmaya başlanmıştır. 1960'lı yıllarda yalnızca istenilen verinin elde edilmesiyle sonuçlanan işlemler artık şimdi geleceğe dönük tahminler ve bu tahminlerin nedenlerinin açıklanmasına dönüşen işlemlere dönüşmüştür (Bilen, 2004).

2.4. Veri Madenciliğine İhtiyaç Duyulma Sebepleri

Günümüzde birçok alanda neredeyse bütün bilgiler bilgisayar sistemleri sayesinde kurulan veri tabanlarına kaydedilmektedir. Ulaşılan veri boyutlarının inanılmaz boyutlarda olduğu kabul edilmesi zorunlu olan bir gerçek olarak karşımıza çıkmaktadır. Veri madenciliği, eldeki ham veriden, anlamlı ve işe yarar bilgiyi çıkarmaya yönelik çalışmalarının bütününü oluşturmaktadır. Yıllar ilerledikçe ortaya çıkan veri yığınları içerisinden potansiyel kullanışlı bilgilerin elde edilmesi amacıyla veri madenciliği ortaya çıkmıştır. Veri madenciliği karar vericiye kullanılabilir bilgi sağlamaktadır ve bu kullanışlı bilgilerin karar vericinin işini kolaylaştırması amacıyla veri madenciliği karar destek sistemleri ile birleştirilerek kullanılmaya başlanmıştır.

2.5. Veri Madenciliğinin Uygulama Alanları

Veri madenciliği uygulanacak veriler genellikle veritabanlarında bulunmakla beraber, ilişkisel veritabanlarına, konumsal ve zamansal verilere de veri madenciliği uygulanmaktadır.

Veri madenciliği şu alanlarda yaygın olarak kullanılmaktadır:

Pazarlama:

- Müşteri gruplamasında,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,

- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- Çapraz satış analizlerinde,
- Müşteri değerlemesinde,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde,
- Hile yoluyla suç işleyen müşterilerin saptanmasında
- Kaybedilen müşterilerin geri kazanılmasında
- Kaybedilebilecek urumda olan müşterilerin tespitinde
- Sepet analizleri yardımı ile marketlerde ürünlerin raflara dağılımının belirlenmesinde.

Bankacılık:

- Farklı finansal göstergeler arasındaki gizli korelasyonlarının bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde
- Müşteri gruplamasında,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespitinde,
- Risk analizlerinde,
- Risk yönetiminde,
- Stok tahmininde,
- Kar analizinde,
- Portföy yönetiminde.

Sigortacılık:

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Perakendecilik:

- Satış noktası veri analizlerinde,
- Alış-veriş sepeti analizlerinde,
- Tedarik ve mağaza yerleşim optimizasyonunda.

Borsa:

- Hisse senedi fiyat tahmininde,
- Genel piyasa analizlerinde,
- Alım-satım stratejilerinin optimizasyonunda.

Telekomünikasyon:

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde,
- İletişim desenlerinin belirlenmesinde,
- Kaynakların daha iyi kullanılmasında,
- Servis kalitesinin arttırılmasında.

Sağlık ve İlaç:

- Test sonuçlarının tahmininde,
- Ürün geliştirmede,
- Tıbbi teşhiste,
- Tedavi sürecinin belirlenmesinde,
- DNA içerisinde genlerin sıralarının belirlenmesinde,
- Protein analizlerinin yapılmasında,
- Hastalık haritalarının hazırlanmasında,
- Hastalık tanılarında,
- Sağlık politikalarına yön verilmesinde.

Endüstri:

- Kalite kontrol analizlerinde

- Lojistikte,
- Üretim süreçlerinin optimizasyonunda.

Bilim ve Mühendislik:

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesinde.

Web hizmetleri:

- Elektronik ticaret yapan firmalar için müşteri davranışlarının belirlenmesinde,
- Web sitesini ziyaret eden kullanıcının daha önceki davranışlarına göre yönlendirilmesinde,
- Web sitesi güvenliğinin sağlanmasında,
- Kullanıcı davranışlarına göre web sitesinin yenilenmesinde,
- Kullanıcı profilinin belirlenmesinde.

İşletmelerde karar destek sistemi içerisinde bilgi çıkarımı büyük önem taşır. Özellikle pazarlama birimlerine gerekli olan bilgiler veritabanlarındaki veriler içerisinde bulunmaktadır. Ancak bu verilerin pazarlama biriminin sorularına yararlı olabilmesi için veriler bir süreçten geçirilmelidir. Pazarlama biriminin soruları genellikle en iyi müşterilerin belirlenmesi, sık alınan ürünlerin belirlenmesi, hangi ürünlerin sıklıkla birlikte alındıklarının belirlenmesi, müşteri gruplarının alışkanlıklarının belirlenmesi yönünde olmaktadır. Benzer biçimde bir yönetici, çalışanların hangilerinin daha iyi iş yaptığını, hangi grup müşterinin memnuniyetinin sağlandığını, finans sektörünün işletmeyi nasıl etkileyeceğini sorabilir. Bu sorular veri madenciliği ile elde edilecek bilgiler yardımıyla cevaplanabilir.

2.6. Veri Madenciliğinde Karşılaşılan Başlıca Problemler

Veri madenciliği girdi olarak kullanılacak ham veriyi veritabanlarından alır. Bu da veri tabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur (Aydoğan, 2003). Sınıflandırmak gerekirse başlıca sorunlar şunlardır:

- Sınırlı Bilgi: Veri tabanları genel olarak belli başlı özellik veya nitelikleri sunmak gibi amaçlar için tasarlanmışlardır. Bu sebeple, öğrenme veya çıkarım işlemlerini kolaylaştıracak özellikler bulunmayabilir.

- Veri tabanı boyutu: Veri tabanı boyutları tutulan veriler ile orantılı olarak inanılmaz bir hızla artmaktadır. Veri tabanı algoritmaları ise çok sayıda küçük veriyi ayrı ayrı işleyebilecek biçimde geliştirilmiştir. Aynı algoritmaların tutulan büyük çaplı verilerde kullanılabilmesi için çok dikkat edilmelidir. Kullanılan veri miktarının büyük çaplı olması, tahminlerin doğruluğu açısından bir avantaj olsa da dikkatsizliklere ve algoritma hatalarına davetiye çıkardığı göz ardı edilemez.

- Aykırı veri: Veri girişi sırasında oluşan kullanıcı hataları veya veri toplanması sırasında oluşan hatalara gürültü adı verilir. Güvenilir sonuçlara ulaşmak için verilerin gürültü miktarlarının az olması istenmektedir. Gürültüler geleceğe dair yapılan tahminlerin veya çıkarım mekanizmalarının doğruluğunun azalmasına neden olur. Gürültülü verilerden kurtulmak için hatalı olabilecek, çok fazla ya da çok az , aşırı uç noktalarda tutarsız veriler yerine anlamlı, aşırı uç noktalarda olmayan veriler kullanılmalıdır. Gürültülü verilerin teşhis edilmesi amacıyla anormali tespiti metotları, histogram, kümeleme analizi ve regresyon yöntemleri kullanılabilmektedir.

- Eksik veri: Veriler kayıt altına alınırken gerekli olabilecek bazı veriler kayıt edilmemiş veya kayıt edilmesi mümkün olmamış olabilir. Eksik veri bulunması durumunda eksik veri içeren kayıt veya kayıtlar veriler içerisinden çıkarılabilir, veri madenciliği için yok sayılabilir; değişkenin, verilerdeki bilinen değerlerinin ortalaması eksik veri değişkeni yerine kullanılabilir; eksik verilerdeki değişkenler, bilinen verilerdeki değişkenlerden değişkenin yapısına uygun olarak tahmin edilebilir.

Eksik veriler, istatistiksel analizler için önemli sorunlar teşkil etmektedirler. Bu durumun sebebi istatistiksel analizlerin genel olarak verilerin tümünün var olduğu

durumlar için hazırlanmış olmalarıdır. Eksik veri içeren veri setlerine istatistiksel analizler uygulanırsa bu eksik verilerin değişkenleri genellikle bilinen verilerin değişken ortalaması ile giderilmekte ve yapılan analizlerin geçerliliğini düşürmektedir.

2.7. Veri Tabanında Bilgi Keşfi Süreci

Veri tabanında bilgi keşfi, verilerden, karar verici için modeller kurularak işe yarar seviyede ve doğru bilgiler elde etmede kullanılan bir süreçtir.

2.7.1. Veri tabanı kavramı

Veri tabanı, sistematik erişim imkanı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir (Wikipedia, 2009). Belirli bir amaca yönelik düzen verilmiş kayıt ve dosyaların tümü olarak tanımlanır.
Veri tabanının genel özellikleri şunlardır:

- Veritabanları, gerçek dünya verilerini küçültülmüş biçimde tutan bir yapıdır.
- Veritabanı verilerin mantıksal olarak birbiriyle ilişkili olduğu bir topluluktur. Rasgele toplanmış, sıralanmamış, gruplanmamış verilere veritabanı olarak bakmak doğru değildir.
- Veritabanı belirlenmiş bir amaca hizmet etmek ve daha sonra verilere ulaşabilmek üzere tasarlanır ve kurulur.
- Veritabanı, herhangi bir büyüklükte ve komplekslikte olabilir.
- Veritabanı el, bilgisayar, elektronik sistemleri yardımı ile oluşturulup yönetilebilir.

Bir veri tabanı oluşturmanın faydaları şunlardır:

- Yasal zorunluluklar hariç herhangi bir evrak saklamaya gerek kalmaz.
- Bilgisayar sistemleri bilgileri daha hızlı biçimde güncelleştirebilirler.

- Yalnızca istenilen bilgiye istenilen zaman ve istenilen biçimde ulaşılabilir.
- Verilerin tek merkezden kontrolü mümkün olur.
- Verilerin gereksiz tekrarı azalır.
- Tutarsız (hatalı) bilgilerin önüne geçilmiş olur.
- Verinin paylaşımı birimler arasında daha kolay sağlanır.
- Verilerde bütünlük sağlanır.
- Raporlama işlemleri kolaylaşmış olur.

İyi bir veri tabanının özellikleri şu şekilde sıralanabilir:

- Veriler hızlı ve kolay biçimde mümkünse elektronik sistem entegrasyonu ile girilebilmelidir.
- Veriler güvenli bir şekilde saklanmalıdır.
- Veriler istenildiği zaman, istenildiği şekilde ve kolay biçimde sorgulanabilmelidir.

Veri tabanlarında bilgi keşfi; verilerden doğru, yeni, faydalı, anlaşılır modeller, kalıplar ve ilişkiler elde etmek için kullanılan özel bir süreçtir. Model elde etmek verileri en iyi biçimde temsil edebilecek modeli bulmak, böylece veri kümesine en iyi biçimde açıklayabilmektir. Süreç ise, veri tabanlarında bilgi keşfinin birçok adımdan, çeşitli yinelemelerden oluştuğunu, göstermektedir. Bilgi keşfinin test edilebilmesi ve test sonucunda kabul edilebilir bir güven düzeyi için geçerli olması, elde edilen bilginin de iş veya karar verme konularında avantajlara olanak sağlayacak şekilde faydalı ve anlaşılır olması gerekmektedir.

Geleneksel veritabanı sorgu ve raporlama araçlarının, mevcut veriler için çoğunlukla yetersiz olduğu görülmüştür. Bu durum, veri tabanlarında bilgi keşfi adı altında, sorgulama ve raporlama yanında yeni yöntem ve metotların geliştirilmesi gereksinimine sebep olmaktadır. Veri tabanlarında bilgi keşfi süreci içerisinde, en uygun modelin tespiti, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen veri madenciliği en önemli adımı oluşturmaktadır.

Veri tabanlarında bilgi keşfi işlemleri, son dönemlerde rekabetin daha da artmasının da etkisi ile veri tabanı mevcut olan işletmelerce büyük ilgi görmektedir. Bu

işletmelere örnek olarak büyük marketler, bankalar, sosyal güvenlik kuruluşları, fabrikalar, perakende satış yapan mağazalar gösterilebilir. Bu büyük veri tabanlarından veri kümelerinin analiz edilip, faydalı kalıp, ilişki ve bilgilere ulaşmak amaçlanmaktadır.

2.7.2. Veri tabanında bilgi keşfi sürecinin evreleri

İşletmelerin her şeyden önce kullanabilecekleri veriler hakkında bilgi sahibi olmaları, sürecin düzgün biçimde işlemesi için gereklidir. Veri tabanında bilgi keşfi sürecinin evreleri adım adım şu biçimdedir:

- Problemin tanımlanması
- Verilerin Hazırlanması
- Modelin Kurulması ve Değerlendirilmesi
- Modelin Kullanılması
- Modelin İzlenmesi

2.7.2.1. Problemin tanımlanması

Veri madenciliği çalışmalarında başarılı olmak için öncelikle yapılacak uygulamanın işletmenin hangi hedefi için yapılacağı ve bu hedefe ulaşmak için ne tür ilişkilerin, yapıların verilerden ortaya çıkartılması gerektiği açık bir şekilde tanımlanmalıdır. İlgili işletmenin hedefi ve mevcut problemi üzerine odaklanılmış olunmalı, açık bir biçimde ifade edilmeli, uygulama sonunda elde edilecek sonuçların ne kadar başarılı olduğuna dair ölçüm yöntemi belirlenmelidir. Hatalı tahmin yapılması durumunda katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmesi uygun görülmektedir.

Analistin, işletmede üretilen sayısal verilerin boyutlarını, proje için yeterlilik düzeyini ve iş süreçlerini iyi analiz etmesi gerekmektedir (Alataş ve Akın, 2004).

2.7.2.2. Verilerin hazırlanması

Modelin kurulması sırasında ortaya çıkacak sorunların yarısından fazlasının temel sebebi verilerin hazırlanması sırasında hatalı, eksik veya , düzensiz olmasından kaynaklanır. Bu durum sık sık geri dönüşlere ve verilerin yeniden düzenlenmesine sebep olmaktadır. Verilerin hazırlanması bu sebeple zaman ayırılması gereken bir aşamadır.

Verilerin hazırlanması kendi içerisinde şu aşamalardan oluşur:

- Veri toplama: Tanımlanan problem için gerekli olduğu düşünülen verilerin ve kaynaklarının belirlenmesi adımıdır.
- Verilere değer biçme: Verilerdeki uyumsuzluklarının belirlenmesi adımıdır.
- Verileri birleştirme ve temizleme: Verilerdeki uyumsuzlukların giderilmesi adımıdır.
- Verilerden örneklem seçimi: Kurulacak modele bağlı veri seçiminin yapıldığı adımdır. Bu adımda modelin test edilmesi için de ayrıca bir veri seti oluşturulmalıdır.
- Verilerin dönüştürülmesi: Verilerin ilgilenilen özellikleri korunarak modele uygun hale getirilmesi adımıdır.

2.7.2.3. Modelin kurulması ve değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi için çok sayıda model kurularak denenmeli veya yeni bir model geliştirilmelidir. Veri hazırlama aşamasında yapılan hatalar bu aşamayı zorlaştıracaktır. Verilerin hazırlanması ve modelin kurulması, performans açısından en uygun model bulunana kadar tekrar edilir.

2.7.2.4. Modelin kullanılması

Kurulan ve performans açısından uygun görülen model tek başına çalışabilecek bir sistem olabileceği gibi bir sistemin alt sistemi durumunda da olabilir. Kurulan

modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında karar vericiye destek amacıyla doğrudan kullanılabileceği gibi, malzeme ihtiyaç planlaması, kurumsal kaynak planlaması gibi süreçlerin alt sistemleri biçiminde de kullanılabilir.

2.7.2.5. Modelin izlenmesi

Zaman içerisinde kurulan modele gelen verilerin yapısı veya özellikleri değişebilir. Bu değişimin sebebi verileri sağlayan sistemlerin de değişmesidir. Dolayısıyla kurulan model geçerliliğini yitirebilecektir. Modelin sürekli izlenmesi modeli destekleyen, veri sağlayan sistemlerdeki değişimin tespit edilmesine yardımcı olur. Sistemlerin değişmesi ve modelin geçerliliğini yitirmesi durumunda model yeniden düzenlenebilir veya baştan tekrar kurulabilir.

2.8. Veri Madenciliği Modelleri

Veri madenciliği modelleri işlevlerine göre üç temel grupta toplanmaktadır:

- Sınıflama
- Kümeleme
- Birliktelik kuralları ve sıralı örüntüler

Kullanılan modeller literatürde ayrıca iki başlık altında toplanabilmektedir: (Gürbüz ve diğ, 2008)

- Tahmin edici
- Tanımlayıcı

Veri madenciliği modelleri açısından sınıflandırma şu biçimde ele alınacaktır:

- Matris cebri temelli modeller
- İstatistik temelli modeller

- Makine öğrenmesi temelli modeller

2.8.1. Matris cebri temelli modeller

Matris cebri, doğrusal sistemlerin modellenmesinde ve bilgi çıkarımında sıklıkla kullanılmaktadır. Bu modeller, kümeleme, boyut azaltma gibi farklı amaçlar için kullanılmaktadır.

2.8.1.1. LU ayrıştırma

Satır ve sütunları doğrusal bağımsız olan nxn boyutlu bir A kare matrisi denklem 2.1 ile gösterilen ayrıştırmaya tabi tutulabilir.

$$P\,A = L\,U \tag{2.1}$$

Bu denklemde P permutasyon matrisi, L köşegeninde 1 değerleri bulunan alt üçgen matris, U üst üçgen matris olacaktır (Eldєn, 2007). P permutasyon matrisi, A matrisinin köşegenleştirilmesinde kullanılmaktadır. Köşegenleştirme işlemi hem A matrisinin kümelenmesinde hem de karar vericinin matrisin davranışını incelemesinde yardımcı olur. Köşegenleştirilen harcama veya satış matrisi üzerinde müşteri grupları veya ürün grupları daha iyi görülebilir.

2.8.1.2. LDL^T ayrıştırma

Herhangi bir simetrik ve pozitif tanımlı A matrisi denklem 2.2 ile gösterilen ayrıştırmaya tabi tutulabilir.

$$A = L\,D\,L^T \tag{2.2}$$

L köşegeninde 1 değerleri bulunan alt üçgen matris, D köşegeninde pozitif elemanlar bulunan matris olacaktır (Eldєn, 2007). $D^{1/2}$ matrisi tanımlanırsa denklem ***** tanımlanabilir.

$$A = L\,D\,L^T = L\,D^{1/2}\,D^{1/2}\,L^T) = U^T\,U \quad (2.3)$$

U üst üçgen matris olmaktadır. Bu ayrıştırma ayrıca Cholesky ayrıştırma olarak ta adlandırılır (Eldєn, 2007).

2.8.1.3. Tekil değerlere ayrıştırma

Bir matrisin özdeğerleri ve özvektörleri, matrisin karakteristiğini belirleyen en önemli özelliklerdir. Bir A matrisinin özdeğerleri ve özvektörleri, denklem 2.4 çözümündeki sabit λ değerleri ve u vektörleridir.

$$Au = \lambda\,u \quad (2.4)$$

Diğer bir ifade ile özvektörler, A matrisi ile çarpıldıklarında genlikleri hariç değişime uğramayan vektörlerdir. Özdeğerler ise ölçeklendirme faktörüdür. Eşitlik ayrıca, I birim matris olmak üzere, denklem 2.5 biçiminde de yazılabilir.

$$(A - \lambda\,I)\,u = 0 \quad (2.5)$$

Denklem 2.5 kare bir matris için çözülürse matrisin özdeğerleri ve özvektörleri bulunabilir. Ayrıca, koşul olarak doğrusal bağımsız n adet özvektörü ve bu özvektörlere karşılık gelen n adet özdeğeri olduğu kabul edilen, diğer bir ifade ile rank(A) = n olan, özdeğerleri ve özvektörleri bilinen nxn boyutlarındaki A kare matrisinin tekrar oluşturulması da gerekebilir. Bu durumda doğrusal bağımsız özvektörler ile U sütun matrisi $U = [u_1, u_2, ..., u_n]$ olacak biçimde, bu özvektörlere karşılık gelen özdeğerler de Λ köşegen matrisinin köşegenini oluşturacak biçimde yerleştirilirse, A kare matrisi denklem 2.6 ile elde edilebilir. Benzer biçimde A kare matrisi U özvektörler matrisi, Λ özdeğerler matrisi olacak biçimde üç matrisin çarpımına ayrıştırılabilir.

$$A = U\,\Lambda\,U^{-1} \quad (2.6)$$

Daha genel biçimde herhangi bir matris üç matrisin çarpımı biçiminde ayrıştırılabilir. Bu ayrıştırma işlemi denklem 2.7 ile ifade edilebilir.

$$A = U \Sigma V^T \quad (2.7)$$

Denklem 2.7' de A matrisi mxn boyutlarında bir matris olmak üzere, U matrisi mxm boyutlarına, Σ diagonal matrisi mxn boyutlarına, V matrisi de nxn boyutlarına sahip matrisler olmaktadırlar. U ve V matrisleri için denklem 2.8 ve denklem 2.9 eşitlikleri geçerlidir.

$$UU^T = I \quad (2.8)$$

$$VV^T = I \quad (2.9)$$

U matrisindeki sütun vektörlerine sol tekil vektörler, V matrisindeki sütun vektörlerine de sağ tekil vektörler adı verilir. Bu U ile V matrislerinin sütun vektörleri doğrusal bağımsızdırlar. Ayrıca U ve V matrislerinin satır ve sütun vektörleri kendi içlerinde birbirlerine göre 90 veya 270 derecelik açıya sahiptirler. V matrisi özvektörler matrisi olarak ta adlandırılmaktadır. Σ köşegen matrisinin köşegen elemanları tekil değerler olarak isimlendirilir. Σ köşegen matrisi de tekil değerler matrisi adını alır. Bu durumda en fazla tekil değerleri ayrıştırılan matrisin doğrusal bağımsız satır veya sütun sayısı kadar tekil değer mevcuttur. Ayrıca A^TA kare matrisinin özvektörleri sağ tekil vektörler, AA^T kare matrisinin özvektörleri de sol tekil vektörlerdir.

Satın alma işlemi yapan müşteriler için satış işlemlerinin satır olarak, satış işlemindeki her bir ürünün de sütunlar ile ifade edildiği, yapılan harcama miktarının da matris hücresinde değer olarak kabul edildiği bir harcama matrisi düşünülebilir. Bu matris tekil değerlerine ayrıştırıldığında özdeğerlere karşılık gelen özvektörler, bir düzlem veya doğru belirtmektedir. Bu özdeğerler gelen herhangi bir müşteri için bilinmeyen bazı harcamaları tahmin etmekte veya müşteri davranışının ne yönde olacağını belirlemekte kullanılabilmektedir (Korn ve diğ., 2000). Bu durumda her bir özvektörü kural olarak kullanmak yerine yeterli miktarda özvektör belirlenerek kurallar oluşturulmaktadır. Basit bir sezgisel yöntem ile özdeğerlerin birikimli

toplamlarının %85 seviyesinde olduğu noktaya kadar olan özvektörler kurallar olarak kabul edilebilir (Korn ve diğ., 2000). Kuralların belirlenmesinden sonra bu kurallar bilinmeyen müşteri davranışlarının tahmini için kullanılabilir.

2.8.2. İstatistik temelli modeller

İstatistik temelli modeller birçok alanda kullanılmaktadır. Başlıca kullanım alanları yığınların karşılaştırılması, eksik verilerin düzeltilmesi ve sınıflandırmadır.

2.8.2.1 Hipotez testleri

Doğruluğu bir araştırma ya da deney ile test edilmeye çalışılan öngörülere hipotez adı verilmektedir (Wikipedia, 2009). Bir örneklem ortalaması ile örneklemin alındığı ana kütle ortalaması farkının anlamlı olup olmadığının belirlenmesinde veya bir örneklem ortalaması ile bu örneklemin çekilmiş olduğunu düşündüğümüz anakütle ortalaması etrafındaki farkın önemli olup olmadığını araştırmayı sağlayan teknikler hipotez testleridir. Ayrıca iki ana kütle ortalaması arasındaki farkın, bu ana kütlelerden seçilmiş örneklemlerin arasındaki farka hipotez testleri uygulanarak farkın önemli olup olmadığı da anlaşılabilmektedir. Hipotez testi sayesinde örnek istatistiklerine dayanılarak ana kütle parametreleri hakkında belirli bir güven seviyesine kadar karar verilebilir (Kartal, 2006).

Adım adım bir hipotez testi şu aşamalar izlenerek yapılmaktadır:

- Hipotezlerin oluşturulması
- Güven seviyesinin α belirlenmesi
- Örneklem dağılımının belirlenmesi
- Ret alanının ve kritik değerin belirlenmesi
- Karşılaştırmaların yapılması ve sonuçların yorumlanması

Hipotez testlerinde ilk olarak bir boş hipotez kurulmalıdır. H_0 olarak gösterilen boş hipotez iki anakütle arasındaki fark için kuruluyorsa iki ana kütle ortalamaları

arasında fark olmadığı görüşünü savunacaktır. Eğer hipotez bir ana kütle için kuruluyorsa ortalamanın bir değer üzerinde veya altında kalacağı biçiminde kurulabilir. Bu adımdan sonra H_0 hipotezinin ret edilmesi durumu için bir alternatif H_a hipotezi kurulacaktır. H_a hipotezi ana kütle ortalamaları arasında sadece fark olduğunu belirtebileceği gibi bu farkın yönünü, büyüklük veya küçüklük cinsinden ifade edilmesini sağlayabilir. Sonraki aşamada hipotezin kurulması ile belirli bir güven düzeyi belirlenir. Bu güven düzeyi hata yapılması durumunda yapılacak hatanın ne kadar olabileceğini belirler. Kurulmuş olan H_a hipotezine göre testin tek kuyruk testi mi çift kuyruk testi mi olacağı belirlenir. Test sonucunda dört durum oluşabilir:

- H_0 doğrudur: Hipotez doğrudur ve H_0 hipotezi kabul edilir.
- α hatası: H_0 doğru olmasına rağmen H_0 hipotezi ret edilir.
- H_0 yanlıştır: Hipotez yanlıştır ve H_0 ret edilir.
- β hatası: H_0 yanlış olmasına rağmen H_0 hipotezi kabul edilir.

Hipotez testlerinin temel varsayımları şu şekildedir:

- Örnekler bağımsız seçilmişlerdir.
- Ana kütle veya ana kütleler normal dağılıma sahiptirler.
- Ana kütlelerin varyansları eşittir.

Ret alanının belirlenmesi için normal dağılım eğrisinde belirlenen α alanına karşılık gelen z_{kritik} değerine bakılır. Test tek kuyruk veya çift kuyruk testi olabileceği için teste uygun olarak z_{kritik} değeri tablolardan okunacaktır. z_{kritik} değerine göre sonuçlar elde edilecektir. Sonuçların α veya β türü hata ile elde edilebileceği göz ardı edilmemelidir.

2.8.2.2 Karar ağaçları

Karar ağaçları veri madenciliğinde en çok kullanılan yöntemlerden birisidir. Hesaplama gücü gereksiniminin düşük olması, kolay yorumlanabilmesi ve veritabanı

sistemleri ile birlikte kolayca çalıştırılabilmeleri karar ağaçlarının sıklıkla kullanılma sebeplerinden birkaçıdır. Karar ağaçları düğümler ve bağlardan oluşan, ağ modeli yapısında ifade edilebilen ve anlaşılması oldukça kolay bir tekniktir. Karar ağaçları, kök düğümden yaprak düğüme doğru çalışır (Wei ve Chiu, 2002) Karar ağacında bulunan her bir bağın belirli bir durumsal olasılığı vardır. Bu olasılıkların gerçekleşme seviyeleri doğrultusunda, son bağlardan geriye doğru bütün bağların olasılık değerleri kök düğüme kadar hesaplanabilmektedir. Böylece karar vericiye beklenen değer seviyesi ile ilgili bilgi sağlamaktadırlar.

Geliştirilen karar ağacı algoritmaları içerisinde;

- CHAID (Chi- Squared Automatic Interaction Detector), C&RT (Classification and Regression Trees),
- ID3,
- Exhaustive CHAID,
- C4.5,
- MARS (Multivariate Adaptive Regression Splines),
- QUEST (Quick, Unbiased, Efficient Statistical Tree),
- C5.0,
- SLIQ (Supervised Learning in Quest),
- SPRINT (Scalable Parallelizable Induction of Decision Trees) başlıcalarıdır (Akpınar, 2000).

2.8.2.3. Regresyon modelleri

Regresyon temel olarak, bir matematiksel denklemin katsayılarının bilinen girdi değerlerine karşılık gelen çıktı değerlerinden en az sapma olacak şekilde bulunması amacını taşır. Doğrusal regresyonda bağımlı değişkenin değeri; lojistik regresyonda ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir (Hui ve Jha, 2000). Doğrusal regresyon denklemi denklem 2.10 ile ifade edilebilir.

$$A\,x - b = 0 \qquad (2.10)$$

Modelin amaç fonksiyonu hataların kareler toplamının en küçüklenmesidir. Bu sebeple 2.10 denkleminin ikinci normu veya kareler toplamının karekökü en küçük yapılmaya çalışılmaktadır. 2.10 denkleminin çözümü ise denklem 2.11 ile gösterilmektedir.

$$\dot{x} = (A^T A)^{-1} A^T b \tag{2.11}$$

Bunun yanı sıra eğer A matrisi tekil veya hastalıklı durumda ise çok sayıda çözüm mevcut olacaktır. Bu gibi durumlar için Levenberg-Marquardt algoritmasına dayanan bir yöntem kullanılmaktadır. Bu yöntem, hastalıklı durumu oluşturan amaç fonksiyonunun basitçe değiştirilmesi ile uygulanabilir duruma gelmektedir. Denklemin amaç fonksiyonu denklem 2.12 ile gösterilir. Amaç fonksiyonundaki Γ matrisi olarak genellikle birim matris kullanılmaktadır. Bu özel regresyon formuna da ridge regresyon veya tikhonov regülarizasyonu adı verilmektedir.

$$||A\,x - b||^2 + ||\,\Gamma\,x||^2 \tag{2.12}$$

Bu regresyon modelinin çözümü de denklem 2.13 ile yapılmaktadır.

$$\dot{x} = (A^T A + \Gamma^T \Gamma\,)^{-1} A^T b \tag{2.13}$$

Γ matrisinin ölçekleme matrisi olarak kullanılması durumunda bu matris 0 matrisi de olabilir. Bu durumda yöntem en küçük kareler yöntemine denk gelir.

Doğrusal olarak modellenmesi zor veya çok yüksek hataya sebep olan problemler için daha farklı regresyon modelleri kullanmak gerekir.

2.8.2.4. Sepet analizi

Pazar sepeti çözümlemesinde sıklıkla beraber alınan nesneler üzerine çalışılır (Rushing, 1997). İşletmede karar vericilerin hangi ürünler birlikte sıklıkla alınmış sorusuna cevap vermek için geliştirilmiş bir yöntemdir. Yöntem temel olarak sayma işlemine dayanır. Yöntemin kullanılabilmesi için işletmenin satış işlemlerine ait

kayıtları tutması gerekir. Satış verilerinde müşterilerin hangi ürünleri sıklıkla birlikte aldıkları ve hangi ürünleri birlikte almadıkları verileri mevcuttur. Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesi, müşteriye daha fazla ürün satma yollarından birisidir (Han ve Kamber, 2001).

Sepet analizi yönteminde birincil amaç ilişkilere dayalı olan kuralları bulmaktır. Birliktelik kurallarına ait örnekler aşağıda yer almaktadır:

- Müşterilerden bira satın alanların %75' i çocuk bezi de satın almaktadır.
- Düşük yağlı peynir ve yağsız süt alan müşterilerin %85' i diyet süt almaktadır.

Olası kuralların sayısı için bir genelleme yapılacak olursa; n adet ürün satan bir işletmede, A ürününü satın alan müşteriler B ürününü de satın almaktadır(A → B) biçiminde ikili ilişki sayısı n ürünün 2' li kombinasyonu kadar olacabilecektir. Benzer biçimde A ve B ürünlerini birlikte alan müşteriler C ürününü de satın almaktadır (A ve B → C) biçiminde üçlü ilişki sayısı da n ürünün 3' lü kombinasyonu kadar olabilecektir. Genelleme yapacak olursak n adet ürün satışı yapan bir firmanın teorik olarak k adet ürün içeren ilişki sayısı n ürünün k' lı kombinasyonu ile ifade edilebilecektir. Dolayısıyla bütün ilişkileri tespit etmek isteyen bir karar verici için olası durum sayısı bu kombinasyonların ayrı ayrı toplanmasını gerektirir. Bu durumda da hesaplanması gereken ilişki sayısı çok hızlı olarak arttığı görülebilir. Satışı yapılan ürün sayısı arttıkça ilişki sayısının da artacağı görülebilir. Ayrıca toplam kombinasyon sayısı aynı zamanda bir kümenin özalt küme sayısına eşittir. Bu eşitlik ile olası ilişki sayısının 2^n-1 olduğunu görülür. Çok sayıda ürün satışının yapıldığı bir işletmede olası ilişki sayısı çok fazla olacaktır.

İkame ilişkilerin de arandığı durumlarda ürünlerin hem satın alınma durumları veri setine dahil edilecek hem de satın alınmama durumları dahil edilecektir. Bu durumda veri setinde n adet ürün verisi yerine 2n adet ürün verisi olacaktır. Böylece ikame ilişkilerin bulunması da mümkün olabilecektir. Sadece A ve B olmak üzere 2 ürün satan bir işletme için veri seti {A, B, A', B'} biçiminde olacaktır. Bu veri setine dahil

olan her bir eleman ayrı birer ürünmüş gibi olacak ve 2n elemanlı kümenin özalt küme sayısı kadar olası ilişki var olacaktır.

Problemin çözüm süresi veri setinin özalt küme büyüklüğü ile birlikte çok hızlı biçimde artmaktadır. Bu sorunun çözümü için iki tür olası çözüm var olabilir. Bu çözümler:

- İlerleme sürecinde ara veri setlerindeki eleman sayının azaltılması.
- Yapılan karşılaştırma sayısının azaltılması.

olarak ifade edilebilir. İlerleme sürecinde ara veri setlerindeki eleman sayısının azaltılması için apriori prensibi geliştirilmiştir. Yapılan karşılaştırma sayısının azaltılması için de destek mekanizması kullanılmaktadır.

2.8.2.5. Apriori prensibi

Apriori prensibine göre eğer bir veri setinde sıklıkla alınan ürünler bulunuyorsa, sıklıkla alınan bu ürünlerin alt setleri de sıklıkla alınır. Bu prensibe göre veri setinde sıklık seviyesi düşük olan ürün ilişkilerinin alt ilişkilerine bakma gereksinimi ortadan kalkacak, veri setinde sıklık seviyesi yüksek olan ürün ilişkilerinin alt ilişkilerine bakma gereksinimi olacaktır. Basit bir örnek vermek gerekirse işletme {A, B, C, D, E, F} ürünlerinin satışını yapıyor olsun. Bu 5 adet ürünün olası ilişki sayısı 63 olacaktır. Ancak sadece {A, B, C} ürünlerinin sık satıldığının bulunması durumunda apriori prensibi uygulandığında {{A, B}, {A, C}, {B, C}} ikili ilişkilerine ve {{A}, {B}, {C}} ilişkilerine bakmak gerekecektir. Benzer biçimde apriori prensibine göre eğer {{D, E}} ikili ilişkisi çok etkili değil ise bu ikili ilişkiyi içeren alt ilişkiler de etkili değildir. Üçlü ilişkiler {{A, D, E}, {B, D, E}, {C, D, E}, {D, E, F}} kümesi {D, E} ikili ilişki kümesinin elemanlarını içerdiğinden bu üçlü ilişkiler kümesindeki ilişkilerin de etkili olmayacağı söylenir.

Bir ilişki kümesinin etkili olup olmadığı nasıl anlaşılabilir? Bu sorunun cevabı için bir kriter geliştirmek gerekmektedir. Temel olarak iki ölçüt belirlemiştir:

- Destek(A → B) = n(A ∩ B) / N
- Güven(A → B) = n(A ∩ B) / n(A)

Bu tanımlarda n(A ∩ B), A ve B ürünlerinin veya kümelerinin birlikte alınma sayısını, n(A) A ürününün veya kümesinin tek başına alınma sayısını, N toplam işlem miktarını göstermektedir. Ticari yazılımlar belirlenen kriterler doğrultusunda destek ve güven seviyeleri istenilen düzeyde olan ilişkileri ortaya çıkartmaktadır. Ancak ikame durumların belirlenmesi için yapay ürün olarak nitelendirilebilecek ürünler ile ilgili ilişkiler destek ve güven durumları göz önüne alınarak birçok durumda belirlenememektedir. Yapay ürünlere ait destek ve güven seviyeleri eğer müşterilerin yaptıkları alımlar çok az ise, fazlasıyla (0.90 seviyesinde) anlamlı çıkmaktadır. Bu da yanıltıcı bir sonuçtur.

İlişki madenciliği karar vericiye sadece hangi ürünlerin bir arada satıldığı ile ilgili bilgi vermektedir. Ayrıca geleneksel ilişki madenciliği harcama verilerini kullanmadığından yapılan harcamalar arasındaki ilişkiler hakkında bilgiler vermemektedir.

2.8.2.6. Benzerlik ölçütleri

Benzerlik ölçütleri genel olarak 0-1 türündeki veriler ile çalışıldığı durumlar için düşünülmüştür. Literatürde çok sayıda benzerlik ölçütü tanımlanmıştır. Bunlardan bazıları:

- Simple Matching Coefficient (SMC)
- Kosinüs
- Korelasyon
- Jaccard
- Tanimoto(Extended Jaccard)

biçiminde sayılabilir.

X ve Y, n adet 0-1 biçiminde ifade edilebilen özellik gösteren vektörler olmaları durumunda;

- f_{00}: x vektöründe 0 ve y vektöründe 0 değerinin birlikte olduğu sıklığı,
- f_{01}: x vektöründe 0 ve y vektöründe 1 değerinin birlikte olduğu sıklığı,
- f_{10}: x vektöründe 1 ve y vektöründe 0 değerinin birlikte olduğu sıklığı,
- f_{11}: x vektöründe 1 ve y vektöründe 1 değerinin birlikte olduğu sıklığı,

göstersin. Bu durumda SMC benzerlik ölçütü denklem 2.14 ile ifade edilmektedir (Tan ve diğ., 2006).

$$SMC(x, y) = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \quad (2.14)$$

Jaccard benzerlik ölçütü de denklem 2.15 ile ifade edilmektedir (Tan ve diğ., 2006).

$$J(x, y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad (2.15)$$

Tanimoto benzerlik ölçütü de denklem 2.16 ile ifade edilmektedir (Tan ve diğ., 2006).

$$EJ(x, y) = \frac{x.y}{\| x \|^2 + \| y \|^2 - x.y} \quad (2.16)$$

Korelasyon ve kosinüs benzerlik ölçütleri, sürekli veriler ile çalışabilmektedirler. Kosinüs ölçütü iki vektör arasındaki ölçünün bir göstergesidir. Korelasyon ölçütü ise iki vektör arasındaki doğrusal ilişki düzeyinin göstergesidir. Kosinüs benzerlik ölçütü denklem 2.17 ile korelasyon benzerlik ölçütü de denklem 2.18 ile ifade edilmektedir (Tan ve diğ., 2006).

$$\text{Cosine}(x, y) = \frac{x.y}{\| x \| . \| y \|} \quad (2.17)$$

$$\text{Corr}(x, y) = \frac{s_{xy}}{s_x s_y} \quad (2.18)$$

Kosinüs benzerlik ölçütünde kullanılan norm ikinci norm olarak ta bilinen euclidean normudur. Kosinüs benzerlik ölçütü -1 ile 1 arası değerler almaktadır. Kosinüs ölçütü, iki vektör aynı yönde ise 1 değeri, zıt yönlerde ise -1 değeri almakta, eğer iki vektörün dik olma durumu söz konusu ise ölçüt 0 değeri almaktadır.

Korelasyon benzerlik ölçütündeki s parametreleri x ve y vektörlerinin standart sapmalarını göstermektedir. Korelasyon benzerlik ölçütü vektörler doğrusal olarak bağımlı ise 1 değeri, bağımlılık yönü ters ise -1 değeri, doğrusal bağımlılık yok ise 0 değeri almaktadır.

Bahsedilen benzerlik ölçütleri iki değişken arasında ilginçlik olması durumunu incelemektedir. Bunun yanı sıra ilgilenilmeme durumlarını inceleyen ölçütler de mevcuttur (Savasere ve diğ., 1998).

2.8.3. Makine öğrenmesi temelli modeller

Makine öğrenmesi bilgisayar sistemlerinin gelişimi ile birlikte yinelemeli yöntemlerin artması ile daha da kullanılır duruma gelmiştir.

2.8.3.1. K-ortalamalar kümeleme algoritması

Kümeleme işlemi çoğunlukla bir başka veri madenciliği uygulaması için bir ilk işlem olarak kullanılır (Tantuğ, 2002). İstatistik ve makine öğrenmesi alanlarında, k-ortalamalar kümeleme algoritması n adet veriyi k adet kümeye bölmek için kullanılan bir kümeleme analizi yöntemidir. Böldüğü kümelere ait veriler, kümeye ait ortalamaya en yakın veriler olmaktadır.

Küme sayısı k olacak biçimde bölünmek istenen n adet verinin bulunduğu veri yapısı için k-ortalamalar algoritmasının amaç fonksiyonu denklem 2.19 ile ifade edilebilir.

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu_i \|^2 \qquad (2.19)$$

Denklem 2.19' da S, kümelere ayrılan verilerin kümelerini, μ_i de S_i kümesinin ortalamasını ifade etmektedir.

Algoritma bir başlangıç çözümü ile başlar. Başlangıç çözüm sezgisel olarak veya rastsal olarak seçilebilir. Çözüme ait merkez noktalar m_i ile gösterilir. Algoritma yinelemeli olarak aşağıdaki iki adımı gerçekleştirir:

1- Atama adımı: $S_i^{(t)} = \{x_j: ||x_j - m_i^{(t)}|| \leq ||x_j - m^*_i{}^{(t)}||; \ \forall \ i^* = 1, ..., k\}$

2- Güncelleme adımı: $m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$

Atama adımında, t yinelemesinde S_i kümesinin merkez noktasına en yakın elemanları kümeye atamaktadır.

Güncelleme adımında, her bir küme için küme elemanlarının oluşturduğu merkez noktayı hesaplamaktadır. Bu merkez nokta sonraki yinelemede atama adımında küme merkezi olarak kullanılacaktır.

En uygun çözüm bulununcaya kadar kayıtlar yeniden atanır ve küme merkezleri ayarlanır (Hui ve Jha, 2000). Yinelemeler merkez noktaların değişmemesi durumunda durdurulur ve böylece kümeleme işlemi tamamlanmış olur.

2.8.3.2. Yapay sinir ağları

Ekonomik alanlardan tıbbi konulara, değerli müşterilerin belirlenmesi için yapılan kümeleme işlemlerinden kredi kartlarında sahtekârlıkların belirlenmesine kadar çok geniş bir alanda uygulanabilmektedir (Tantuğ, 2002).

Yapay sinir ağları, biyolojik sinir sisteminin benzetiminden esinlenerek çalışmaktadır. İnsan beyni nöron olarak bilinen sinir hücreleri içermektedir. Bu hücreler birbirleri ile aksonlar adı verilen lifler ile birbirlerine bağlanmışlardır. Aksonlar, sinir hücrelerinin ürettiği sinir dürtülerini ilişkili sinir hücrelerine ileterek aktif duruma geçmelerini sağlar. Bir sinir hücresi diğer sinir hücrelerinin aksonlarına dendritler ile bağlanmıştır. Dendritler sinir hücresinin gövdesinin bir uzantısıdır. Bir dendrit ile bir aksonun birleşme noktasına da sinaps adı verilmektedir. Nörologlar insan beyninin sinir hücrelerinin aynı işaret ile uyarılması sonucu hücreler arasındaki sinaps bağlantısının gücünün değişimiyle öğrendiğini keşfetmişlerdir.

İnsan beyni yapısının analizi ile yapay sinir ağları, düğümler ve yönlü bağların birleştirilmesi ile düzenlenmiştir. Çok sayıda yapay sinir ağı modeli oluşturulmuştur. Bu ağların bir bölümü sınıflandırma problemlerinde kullanılmaktadırlar.

En temel yapay sinir ağı modeli tek katmanlı algılayıcı modelidir. Tek katmanlı algılayıcı modeli iki çeşit düğüm içermektedir:

- Giriş düğümleri: Giriş değerlerini giriş özellikleri biçiminde ağırlıklandırarak çıkış düğümüne iletmek için kullanılırlar.
- Çıkış düğümü: Modelin oluşturduğu çıktıyı temsil etmek için kullanılır.

Yapay sinir ağlarında bir düğüm nöron olarak adlandırılır. Ağırlıklandırılmış bağlar, nöronlar arası iletişimin gücünü belirlemektedir. Biyolojik sinir sistemlerindeki gibi bir tek katmanlı algılayıcı modelinin eğitilmesi, verilen veriye ait giriş-çıkış ilişkileri doğrulanıncaya kadar bağların ağırlıklarının değiştirilmesi ile sağlanır.

Bir tek katmanlı algılayıcı çıkış değerini, giriş değerlerinin ağırlıklandırılmış toplamlarından eşik değeri çıkartarak sonucun işaretine göre belirler. Örneğin model x girişlerine göre y çıkışı üretecek ise ağırlık vektörü w, eşik değeri Φ olmak üzere, y çıktısı denklem 2.20 ile belirlenir.

$$y = \text{sign}(w\, x^T - \Phi) \tag{2.20}$$

Giriş düğümleri ile çıkış düğümü arasındaki farkı belirtmek gerekir. Giriş düğümleri girişleri üzerlerinde hiçbir değişiklik yapmadan çıkış düğümüne iletmektedir. Çıkış düğümü ise matematiksel olarak girişlerin ağırlıklı toplamlarını hesaplamakta, eşik değerini bulduğu sonuçtan çıkartmakta ve oluşan yeni sonucun işaretine uygun bir çıkış değeri üretmektedir.

Tek katmanlı algılayıcı modelinin öğrenme evresinde ağırlık vektörü, tek katmanlı algılayıcı, çıkışı gerçek öğrenme çıkışları ile aynı olana kadar ayarlanır. Ağırlıkların değişimi denklem 2.21 yardımı ile yapılır.

$$w_j^{(k+1)} = w_j^{(k)} + \lambda\, (y_i - \hat{y}_i^{(k)})\, x_{ij} \tag{2.21}$$

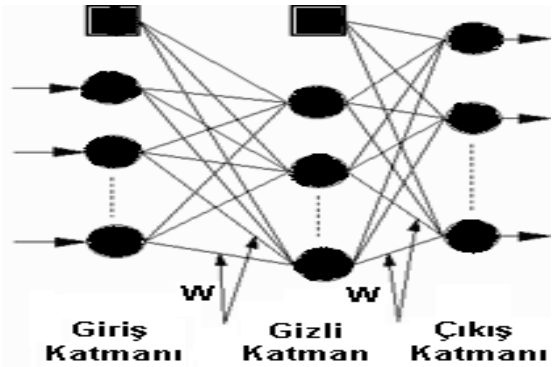
Denklem 2.21' de $w^{(k)}$, k. yinelemede ağırlık vektörünü, λ öğrenme katsayısını, x_{ij} i. öğrenme verisinin j. girişini ifade edilmektedir. Denklem 2.21 ile görülebileceği gibi ağırlık vektörünün değişimi önceki ağırlık vektörü w' ya ve tahminleme hatası olan $(y_i - y_i^{(k)})$ değerine bağlıdır. Tahminleme hatası yapılmamışsa ağırlık bir önceki yineleme değerini koruyacaktır. Eğer hata mevcutsa iki durum söz konusudur:

- Eğer olması gereken çıkış değeri y = 1 ve tahmin edilen çıkış değeri ŷ = -1 ise hata (y – ŷ) = 2 değerini alacaktır. Hatayı azaltmak için tahmin edilen çıkış değerini arttırmak gerekir. Bu sebeple pozitif olan ağırlıkların değerleri arttırılmalı, negatif olan ağırlıkların değerleri azaltılmalıdır.
- Eğer olması gereken çıkış değeri y = -1 ve tahmin edilen çıkış değeri ŷ = 1 ise hata (y - ŷ = -2) değerini alacaktır. Hatayı azaltmak için tahmin edilen çıkış değerini

azaltmak gerekir. Bu sebeple pozitif olan ağırlıkların değerleri azaltılmalı, negatif olan ağırlıkların değerleri arttırılmalıdır.

Ağırlıkların değişim hızı çok yüksek olmamalıdır. Bunun sebebi ağırlık değişiminin o yinelemeye ait giriş değerleri için olmasıdır. Ağırlıkların değişim hızının çok yüksek olması durumunda önceki verilerden elde edilen ağırlık bilgileri kaybedilmiş olacaktır. Değişim hızının çok düşük olması durumunda ise önceki verilerin etkisi çok yüksek olarak kalacak ve gerekli yineleme sayısı çok fazla olabilecektir. Öğrenme katsayısı, λ da bu ağırlıkların değişimini sınırlandırılması amacı ile kullanılmaktadır. Bu parametre adaptif olarak değiştirilebilir. İlk yinelemelerde yüksek olan λ değeri yinelemeler ilerledikçe azaltılabilir.
Tek katmanlı algılayıcı modeli bir doğru ile bölünebilecek olan sınıflandırma problemleri için öğrenme katsayısının düşük olması durumunda mutlaka yakınsama sağlayacaktır. Ancak bir doğru ile bölünemeyen sınıflandırma problemleri için tek katmanlı algılayıcı modeli yakınsama sağlayamayacaktır. Bu tür problemler için çok katmanlı yapay sinir ağları kullanılmaktadır.

Çok katmanlı yapay sinir ağları, çok sayıda tek katmanlı algılayıcı modelinin birbirine bağlanması ile oluşturulur. Basit bir çok katmanlı yapay sinir ağı yapısı Şekil 2.1 ile gösterilmektedir. Çok katmanlı algılayıcılar için esas problem hatanın hücrelere nasıl yayılacağı ve dolayısıyla hücrelerin ağırlıklarının nasıl değişeceği problemidir. Hatanın hücrelere dağıtılması için bazı algoritmalar geliştirilmiş ve çok katmanlı yapay sinir ağları ile doğrusal olarak bölümlenemeyen yüzeylerde de sınıflama problemlerinin çözümü mümkün hale gelmiştir.



Şekil 2.1: Çok katmanlı yapay sinir ağı

3. PARETO ANALİZİ

Pareto analizi temelde çıktıların %80 inin girdilerin %20 sinden geldiğini belirten pareto kanununa dayanmaktadır. Bu kanuna göre önemli olan verileri belirlemek için bir eğri çizilir ve bu eğri üzerinden önemli olan veriler hakkında karara varılır. Birçok alanda bu kanun önem taşıyan odak noktalarının belirlenmesinde kullanılmaktadır. Bir pazarlamacı için hangi müşteri grubunun daha önemli olduğunu belirlemek, bir karar verici için sonuçların hangisinin daha az maliyet ile daha yüksek gelir getirebileceği gibi durumların tespitinde kullanılmaktadır.

Pareto eğrisinin çizilmesinde kullanılan yöntem şu şekilde özetlenebilir:

- Olayların ve olayların olma sıklıkları tablo haline getirilir.
- Olayların önem sırasına göre azalan sırada tablo satırları tekrar düzenlenir.
- Tabloda birikimli sıklıklar yüzdesi ile ilgili bir kolon oluşturulur.
- Yatay eksende olaylar, düşey eksende birikimli sıklıklar yüzdesi olacak biçimde noktalar bir grafik üzerinde işaretlenir.
- İşaretlenen noktalar bir eğri olacak biçimde birleştirilir.
- Grafikte düşey eksenin %80 olduğu nokta bulunur ve yatay eksene bir dik indirilir.
- İndirilen dikmenin sol tarafı önemli olan olayları, sağ tarafı ise önemsiz olayları göstermektedir.
- En az olayların %80' inin kapsandığı kontrol edilir.

Pareto analizi sonucunda olayların önemli olanlar ve nispeten önemsiz olanlar olarak gruplanması sağlanır. Bu gruplama sonunda karar verici önemli olaylar üzerine yoğunlaşabilecektir.

4. UYGULAMA

Bir satış işletmesinde karar vericilerin çok sayıda stratejik karar vermesi gerekmektedir. Verilecek olan kararların bir çoğu işletmenin kuruluş amacı olan satışları arttırma ve kar amacına yöneliktir. Satışların ve elde edilen kar miktarının arttırılması için kullanılabilecek veriler önem taşımaktadır. Ele alınan problemin çözümü karar vericiye, işletme amaçlarına uygun olarak fiyatların nasıl belirlenmesi gerektiğine dair bir temel teşkil edecektir.

4.1. Problemin Tanımı

Veri madenciliğinin uygulanabilmesi için mevcut verilerin bir biçimde kayıt altında tutuluyor olması gerekmektedir. Bu tip veri saklama işlemleri bilgisayar sistemleri bugünkü seviyelere kadar gelişmeden önce maliyetli olmasından dolayı sadece muhasebe kayıtları gibi yasal zorunluluğu olan veriler kayıt altında tutulmakta idi. Bilgisayar sistemlerinin gelişimi ile satış verileri, satılan ürün miktarı, satılan ürünlerin maliyetleri, ürünlerden elde edilen kar miktarı, satış yapılan perakendeci ile ilgili bilgiler, hatta satışı gerçekleştiren satış sorumlusu gibi bilgiler çok düşük maliyetlere katlanılarak kurulan veritabanında saklanabilmektedir. Ham veriler içerisinden işletmenin daha çok kar elde etmesini sağlayacak verilerin çıkartılması da gerekmektedir. Bu veriler, karar vericilerin vereceği kararlar için bir dayanak teşkil edecek ve daha etkili kararların alınmasında rol oynayacaktır.

Mevcut problem perakende satış yapan bir hazır giyim firmasının satışını yaptığı ürünlerin birbirleri ile ilişkisinin bulunmasıdır. Bu ilişkiler müşteriler tarafından satın alınan ürünleri etkileyen faktörlerin tamamı bilinmeden incelenmek durumunda kalınmıştır. İşletmenin ayrıca aynı müşteriye birden çok defa satış yapması mümkündür. Ancak bu durumda işletme bir kart sistemi veya benzer bir sistem kullanmadığından dolayı, müşterisine birden çok defa yaptığı satışı belirleyememektedir.

İşletmenin satışını yaptığı ürünlerin çok sayıda olması, anlamlı ilişki bulma durumunu fazlasıyla zorlaştırmaktadır. Bu sebeple işletmenin satışını yaptığı ürünlerin modelleri problem verisi olarak kullanılmıştır. Ayrıca satılan model ürünlerin büyük bir bölümünden çok az sayıda satılabilmiştir. Bu durum da göz önüne alınarak en çok satılmış olan 50 model üzerinde analizler yapılmıştır. Bu 50 model ürünün satış zaman dilimi ve dolayısıyla herhangi bir mağazanın stoğunda o zaman diliminde bulunup bulunmadığı dikkate alınmamış ve her model ürünün, satış dönemi boyunca bütün mağazaların stoğunda bulunduğu kabulü yapılmıştır.

Veri madenciliği teknikleri kullanılarak hem harcama verileri üzerinden hem de satın alınıp alınmama verileri üzerinden ayrı ayrı pozitif ve ikame ilişkilerin aranmasına yönelik bir model geliştirilmiştir.

4.2. Ridge Regresyon Modeli

Ridge regresyon modeli doğrusal bir regresyon modelidir. En küçük kareler olarak bilinen yöntemi özel bir durumu ile de kapsayan bir modeldir. Model gerçek hayat verilerine satışı en çok yapılmış olan 50 ürün için kurulmuştur. Her bir ürün modeli teker teker girdi verisi olarak kullanılmıştır. Ayrıca regresyon modelinin özelliği dolayısıyla birim matrisin belirli bir parametre ile genişletilmesi durumu da söz konusudur. Her bir ürün modeli için belirlenmiş olan parametre aralığı ile ilgili çözüm yapılmış ve çıktılar elde edilmiştir.

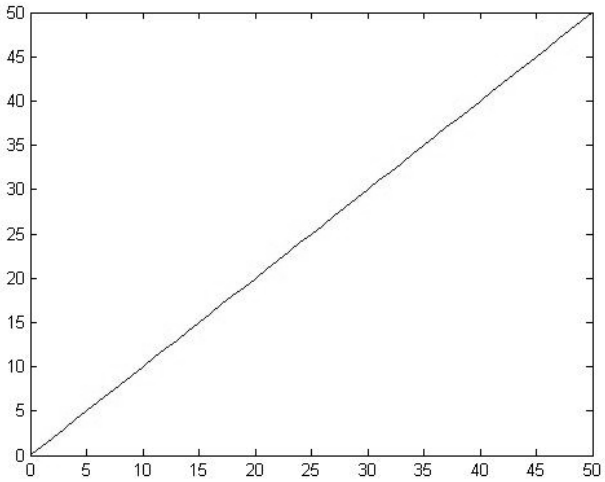
Yapılan parametre denemeleri için bulunan katsayılar arası en büyük katsayı farkı 0.0278 olarak bulunmuştur. Parametre etkisinin çok küçük olduğu sonucuna varılmıştır. Bu durum regresyon matrislerinin hastalıklı durumda olmadığını göstermektedir. Ayrıca farkın bu kadar küçük bulunması en küçük kareler yönteminin yeterli sonucu vereceğini de ortaya koymuştur..

Bulunan denklem parametreleri içerisinde pozitif ve negatif işaretler, pozitif ve negatif ilişkileri belirtebilecektir. Ancak unutulmamalıdır ki bu yöntem sadece doğrusal ilişkilere göre hareket etmekte ve sabit değerler ile oluşan farkı kapatmaya

yönelik pozitif veya negatif yöndeki sapmaları kapatmak için de katsayıları değiştirebilmektedir. Kullanılabilecek en basit yöntemlerden birisi olduğundan dolayı gerçek hayat problemine uygulanmıştır.

4.3. Geliştirilen Yöntem

Geliştirilen yöntem satışı yapılan modeller arası hem ikame ilişkilerin bulunması için hem de birliktelik kurallarının tespiti için kullanılabilmektedir. Negatif kuralların araştırılmasında diğer ürünler ile çok düşük olasılıkla satın alınan ürünlerin bulunması ile ilgilenilebilir (Savasere ve diğ., 1998). Tekil değerlere ayrıştırma(TDA) yöntemi ile harcama matrislerine ait verilerin çözümlenmesi ve müşterilerin davranışlarının tahmin edilmesi mümkün olmaktadır (Korn ve diğ., 2000). TDA yöntemi uygulanacak matris satırlarında müşteri bilgilerini, sütunlarında ürün veya model bilgilerini bulunduran ve müşterinin ürüne yaptığı harcama miktarını gösteren harcama matrisi olacaktır. TDA yönteminin kullanılması ile oluşturulan kurallar, harcama matrisini neredeyse en iyi biçimde özetlemektedir. Özet matristen bazı çıkarımlar yapılması mümkündür. Ancak bu çıkarımlar tam matris üzerinden yapılamamaktadır. Bunun sebebi ise TDA sonucunda bulunan özvektörlerin en iyi özetlemeyi yapabilmek adına birbirine dik olmalarıdır. Bunun sonucu olarak kosinüs benzerlik ölçütü bütün ürünleri ve bütün kuralları birbirleri ile ilişkisiz olarak yorumlamaktadır. Belirli bir miktarda kuralın kullanılması gereksinimi mevcuttur. Bu kural sayısının nasıl olması gerektiği pareto analizi ile belirlenebilecektir. Pareto analizinin uygulanmasında özdeğerlere ait eğri çizilmiş ve bu eğrinin kırılım noktalarındaki eğimler dikkate alınmıştır. Elde edilen özvektörlere karşılık özdeğerler sıralanmıştır. Özdeğerler azalan sırada oldukları için pareto eğrisinin en kötü durumda 45 derecelik bir açı oluşturan bir doğru biçimine gelmesi gerekmektedir. Bu durum Şekil 4.1 ile gösterilmektedir. İki özdeğerin birikimli durumlarını birleştiren doğru parçasının eğiminin 45 derecenin altına düştüğü ilk nokta bizim için kuralları belirleyen kesim noktası olarak kabul edilebilecektir.



Şekil 4.1: En kötü durumda pareto eğrisi

Geliştirilen yöntemin adım adım basit bir örnek üzerinde gösterilmesinden önce örnek uygulamalar üzerinde matris yöntemlerin nasıl kullanıldığına dair örnekler anlatılacaktır. Öncelikle LU ayrıştırma ile köşegenleştirme üzerinde durulacaktır. Bu ayrıştırma yöntemi için nxn boyutlarında tekil olmayan - rankı n olan- bir matrisin satışları ifade ettiği düşünülsün. Bu satış verilerine ait müşteri bilgilerinin karışık olarak tutulduğu ve müşteri gruplamasının veya ürün gruplamasının yapılmak istediğini varsayılsın. Örnek olarak ta Tablo 4.1 ile gösterilen M_{LU} adı verilen matrisin kullanıldığı varsayılsın.

Tablo 4.1 ile gösterilen M_{LU} matrisinin boyutları 7x7 ve doğrusal bağımsız vektör sayısı veya rankı da 7 dir. M_{LU} matrisinin LU ayrıştırmasına tabi tutulması ile P permutasyon matrisi hesaplanabilir. İşlemler için matlab yazılımı kullanılarak P matrisi Tablo 4.2 ile gösterilen matris olarak bulunmaktadır.

Hesaplanan P permutasyon matrisi ile örnek M_{LU} matrisi matris çarpımı ile çarpılırsa oluşan matris M_{LU} matrisinin köşegenleştirilmiş biçimi olacaktır. Bu çarpım işleminin sonucunda oluşan matris Tablo 4.3 ile gösterilmektedir.

Tablo 4.1: LU ayrıştırma örnek M_{LU} matrisi

	Ürün 1	Ürün 2	Ürün 3	Ürün 4	Ürün 5	Ürün 6	Ürün 7
Müşteri 1	0	0	5	6	7	0	0
Müşteri 2	0	0	0	0	1	3	5
Müşteri 3	0	0	0	8	7	6	5
Müşteri 4	0	2	3	4	0	0	0
Müşteri 5	0	0	0	0	0	3	9
Müşteri 6	1	1	1	0	0	0	0
Müşteri 7	0	0	0	0	0	7	2

Tablo 4.2: LU ayrıştırma örnek P permutasyon matrisi

0	0	0	0	0	1	0
0	0	0	1	0	0	0
1	0	0	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	1	0	0

Tablo 4.3: P*M_{LU} köşegenleştirilmiş M_{LU} matrisi

Müşteri 6	1	1	1	0	0	0	0
Müşteri 4	0	2	3	4	0	0	0
Müşteri 1	0	0	5	6	7	0	0
Müşteri 3	0	0	0	8	7	6	5
Müşteri 2	0	0	0	0	1	3	5
Müşteri 7	0	0	0	0	0	7	2
Müşteri 5	0	0	0	0	0	3	9

Görüleceği üzere P*M_{LU} sonuç matrisi aslında köşegenleştirilmiş M_{LU} matrisidir. Bu adımdan sonra karar vericinin müşteri davranışlarını ve gruplarını daha kolayca görmesi mümkündür. Birbirine benzer davranışlarda bulunmuş olan müşteriler birbirine yakın olmaktadır. Örneğin müşteri 5 ve müşteri 7 sadece ürün 6 ve ürün 7 satın almışlardır ve P*M_{LU} köşegenleştirilmiş matrisinde alt alta konumlara gelmiştir.

İkinci bir örnek ile TDA yöntemi açıklanmaktadır. n adet müşterinin satırlar ile temsil edildiği, m adet ürünün de sütunlar ile temsil edildiği bir harcama matrisi X olsun. Amaç, v_1:v_2:v_3:...:v_m biçiminde, matrisin herhangi bir veya birden çok satırında bulunan boş değerleri tahminlemeyi sağlayacak oran kurallarını bulmaktır (Korn ve diğ., 2000). Bu X matrisi Tablo 4.4 ile gösterilen matris olsun (Korn ve diğ., 2000).

Bu X matrisinin rankı 2' dir. Bu işletmede görüleceği üzere iki tür müşteri grubu bulunmaktadır. İlk müşteri grubu sadece ürün 1, ürün 2, ürün 3 satın almaktadır. İkinci müşteri grubu ise sadece ürün 4 ve ürün 5 satın almaktadır. Bu X matrisine TDA uygulanarak matrisin sağ tekil vektörleri, sol tekil vektörleri ve özdeğerleri hesaplanır. Bu matrisler üç matrisin çarpımı biçiminde Tablo 4.5 ile gösterilmektedir.

Tablo 4.4: TDA örnek X matrisi

	Ürün 1	Ürün 2	Ürün 3	Ürün 4	Ürün 5
Müşteri 1	1	1	1	0	0
Müşteri 2	2	2	2	0	0
Müşteri 3	1	1	1	0	0
Müşteri 4	5	5	5	0	0
Müşteri 5	0	0	0	2	2
Müşteri 6	0	0	0	3	3
Müşteri 7	0	0	0	1	1

Tablo 4.5: TDA ile örnek X matrisinin ayıştırılması

1	1	1	0	0		0.18	0		9.64	0		0.58	0.58	0.58	0	0
2	2	2	0	0		0.36	0									
1	1	1	0	0		0.18	0									
5	5	5	0	0	=	0.90	0	x			x					
0	0	0	2	2		0	0.53		0	5.29		0	0	0	0.71	0.71
0	0	0	3	3		0	0.80									
0	0	0	1	1		0	0.27									

İki adet müşteri grubu bulunduğu için özvektörler olarak ta bilinen sağ tekil vektör sayısı iki olmaktadır. Bulunan özvektörler kuralları oluşturmaktadır ve bazı satın alımları bilinen bir müşterinin satın alımı bilinmeyen bir ürüne ne kadar harcama yapabileceği bu kurallar ile hesaplanabilmektedir. Bu hesaplama sonucunda tahmin edilen değerlerin sütun ortalaması yönteminden daha iyi sonuçlar verdiği bilinmektedir (Korn ve diğ., 2000).

Negatif ilişkili ürünlerin bulunmasına yönelik bir model ise sıklık analizi yardımı ile önerilmiştir. Bu yöntem duruma farklı bir açıdan yaklaşmakta ve merkez ürün de denilebilecek olan üçüncü bir ürünün varlığına ihtiyaç duymaktadır. Özetlenecek olursa a ve M_o ürünleri çok sayıda birlikte alınmış, b ve M_o ürünleri de çok sayıda birlikte alınmış ancak a ve b ürünleri birlikte çok az sayıda birlikte alınmış ise bu a ve b ürünlerinin ikame olduğundan söz edilmektedir. Sepet analizi verisinde, bu yöntem, ürünler arası rekabetçi analiz için kullanılabilir (Tan, 2001).

İşletmeler fiyatlarını belirli dönemler içerisinde değiştirebilmektedir. Bir zaman diliminde bir ürüne olan talep az iken satışları arttırmak için ürünün fiyatı düşürülebilmektedir. Bu durum müşterilerin yaptıkları harcama miktarlarını etkilemektedir. İşletmeler coğrafi konumlandırma gibi sebepler ile ürünlerin aynı zaman dilimi içerisinde farklı satış merkezlerinde fiyatlarında farklılaştırmaya gidebilmektedir. Ayrıca müşterilerin bir üründen birden çok miktarda satın alma durumları da olabilmektedir.

Geliştirilen algoritmanın anlatımında örnek harcama matrisinin Tablo 4.6 ile verildiği bir durum olsun.

Örnek matriste 1. müşteri 3 para birimi harcama yaparak 1. ürünü satın almıştır ancak 3. ürüne hiç para ödememiş yani satın almamıştır. Benzer biçimde 5. müşteri 4 para birimi harcama yaparak 3. ürünü satın almıştır ancak 1. ürüne harcama yapmamıştır. 10. müşteri ise her üç ürüne de farklı miktarda para harcamış ve satın almıştır.

Örnek müşteri-ürün harcama matrisine X matrisi adı verilmesi durumunda, X matrisine TDA uygulanması ile özdeğerler ve özvektörler bulunabilir. TDA uygulaması sonucu sağ tekil vektörleri gösteren U matrisi Tablo 4.7 ile, sol tekil vektörleri veya özdeğerleri gösteren V matrisi Tablo 4.9 ile ve tekil vektörlere karşılık gelen özdeğerleri gösteren Σ matrisi de Tablo 4.8 ile gösterilmektedir.

Tablo 4.6: Geliştirilen yöntem için örnek X matrisi

	Ürün 1	Ürün 2	Ürün 3
Müşteri 1	3	1	0
Müşteri 2	2	2	0
Müşteri 3	2	1	0
Müşteri 4	5	5	0
Müşteri 5	0	1	4
Müşteri 6	0	2	2
Müşteri 7	0	1	2
Müşteri 8	0	2	5
Müşteri 9	0	3	1
Müşteri 10	1	3	4
Müşteri 11	4	2	3

Her bir özvektöre karşılık gelen özdeğerler, özvektörlerin ağırlıkları olarak düşünülebilir. Kurallar olarak bu özvektörlerden hangilerinin kullanılacağının belirlenmesi için özdeğerlere ait pareto eğrisi çizilebilir veya bütün olası kural sayıları denenebilir. Kural sayısı belirlenirken en az 2 adet özvektörün kullanılması garanti altına alınmalıdır. Bu durum karşılaştırma yapabilmek için ön gerekliliktir.

Tablo 4.7: X matrisinin sol tekil vektörleri

U =	-0,19	-0,29	0,31
	-0,2	-0,22	-0,11
	-0,14	-0,2	0,13
	-0,49	-0,56	-0,26
	-0,25	0,39	0,18
	-0,21	0,16	-0,26
	-0,15	0,18	-0,03
	-0,36	0,47	0,05
	-0,22	0,03	-0,6
	-0,41	0,26	-0,11
	-0,43	-0,09	0,57

Tablo 4.8: X matrisinin özdeğerleri

Σ =	11,61	0	0
	0	7,22	0
	0	0	3,17

Tablo 4.9: X matrisinin sağ tekil vektörleri

V^t =	-0,5	-0,64	-0,58
	-0,64	-0,17	0,75
	-0,58	0,75	0,33

Örnekteki doğru parçalarının eğimleri sırasıyla [1.58 0.98 0.43] olmaktadır ve ilk iki özvektör kural olarak seçilebilir:

Kural 1: [-0.5, -0.64, -0.58]

Kural 2: [-0,64, -0.17, 0.75]

Kullanılacak kurallara ürün bazında kosinüs ve korelasyon benzerlik ölçütleri uygulanmıştır. Hem kesikli hem de sürekli olarak iki farklı veri için örnek problem uygulaması yapılmıştır. Kesikli kural matrisi sürekli kural matrisinin signum fonksiyonuna tabi tutulması ile elde edilmiştir.

Sürekli matrise kosinüs benzerlik ölçütü uygulanarak elde edilen sonuçlar Tablo 4.10 ile, kesikli matrise kosinüs benzerlik ölçütü uygulanarak elde edilen sonuçlar Tablo 4.11 ile, sürekli matrise korelasyon benzerlik ölçütü uygulanarak elde edilen sonuçlar Tablo 4.12 ile, kesikli matrise korelasyon benzerlik ölçütü uygulanarak elde edilen sonuçlar Tablo 4.13 ile, gösterilmiştir. Kesikli matrise uygulanan korelasyon ölçütünün bazı verileri, veri yapısında standart sapmaların sıfır değeri almasından dolayı hesaplanamamıştır.

Tablo 4.10: Örnek probleme ait sürekli kurallarda kosinüs benzerlikleri

1	-1	-1
-1	1	1
-1	1	1

Tablo 4.11: Örnek probleme ait kesikli kurallarda kosinüs benzerlikleri

1	1	0
1	1	0
0	0	1

Tablo 4.12: Örnek probleme ait sürekli kurallarda korelasyon benzerlikleri

1	-1	-1
-1	1	1
-1	1	1

Tablo 4.13: Örnek probleme ait kesikli kurallarda korelasyon benzerlikleri

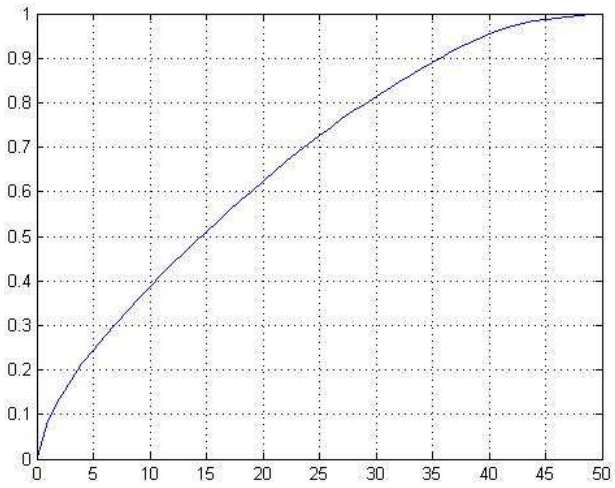
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	1

Kesikli kural vektörlerine uygulanan korelasyon, sürekli kural vektörlerine uygulanan korelasyon ve kosinüs ölçütleri ürün 1 ile ürün 3 arasında negatif ilişki olduğunu göstermektedir. Kesikli kural vektörlerine kosinüs ve korelasyon uygulanmasından ise elde edilen sonuçlar tutarsızdır. Ayrıca bu ürünlerin ürün 2 ile birlikte ayrı ayrı sıklıkla alınmalarına rağmen birlikte alınma oranlarının da düşük olduğu görülmektedir.

Gerçek veri seti ile yapılan uygulamada mevcut harcama matrisi çok daha fazla veri içermektedir. Ayrıca müşterilerin satışı yapılan ürünlerden çok az sayıda ürün almış olmaları dolayısıyla matrisin büyük bir bölümü harcama olmadığı için sıfır değeri içermektedir. Bu durum mevcut matrisin seyrek bir matris olduğu anlamına gelmekte ve seyrekliğin seviyesi ölçülebilmektedir. Tekil değerlere ayrıştırma yönteminin tahmin yapma amacı ile kullanımı neticesinde elde edilen sonuçlarda kullanılan matrisin birçok verisi sıfır olduğundan dolayı etkin sonuç elde edilemediği görülmüştür.

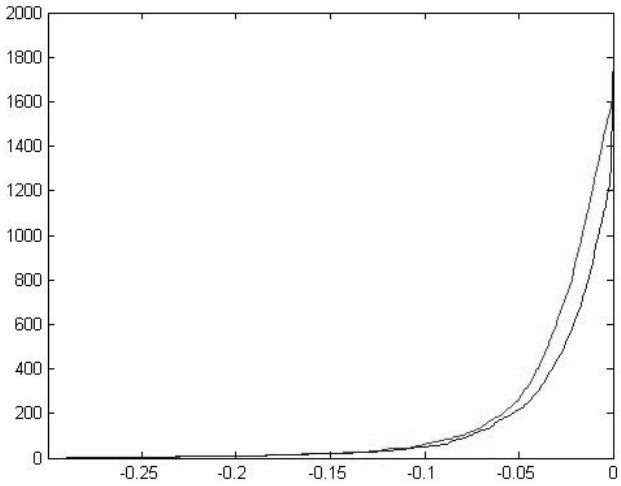
İşletme verilerine geliştirilen yöntemin uygulanması sırasında pareto eğrisinden yararlanılmıştır. Özdeğerlere karşılık gelen pareto eğrisi Şekil 4.2 ile gösterilmektedir. Daha az sayıda kural kullanılarak daha iyi sonuçların elde

edilebilmesi için pareto eğrisinin sol üst köşeye daha yakın olması gerekmektedir. Bulunan 50 adet özvektörün en yüksek özdeğere sahip 24 adetinin kural olarak kullanılabilir olduğu görülmektedir. Ancak 1 kural ile çalışmak anlamsız olacağından kural sayısının belirlenmesinde bulunan sayıya 1 eklenmiş ve en az 2 kural ile çalışılacağı neredeyse garanti altına alınmıştır. Sonraki aşamalarda kural seviyesinin daha iyi belirlenmesi amacıyla kural sayısına göre sistemin değişimlerini gözlemlemek için kural yüzeyi çizdirilmiştir.



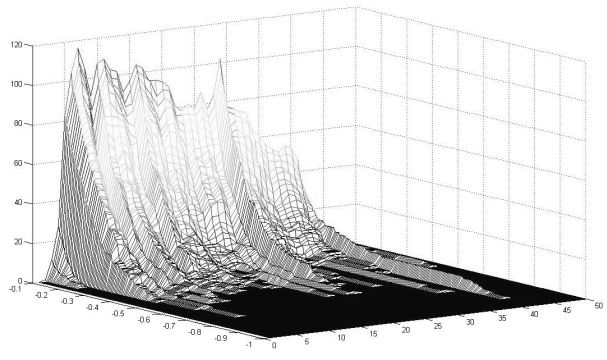
Şekil 4.2: Uygulama problemine ait pareto eğrisi

Bulunan kurallara kosinüs ve korelasyon benzerlik ölçütleri ürünlerin karşılaştırılması amacıyla uygulanmıştır. Benzerlik ölçütlerinin sonuçları arasında çok büyük farklar görülmemiştir. Bu farkların ne düzeyde olduğu ise Şekil 4.3 ile gösterilmektedir. Yatay eksen kesim düzeyini, düşey eksen bulunan kural sayısını göstermektedir. Daha çok kural bulan eğri korelasyon ölçütüne ait eğridir. Şekil 4.3 ile görüleceği gibi -0.1 seviyesine kadar arada oluşan fark çok önemsiz olmaktadır. Ancak 0 merkez noktasına yaklaşıldığı durumda fark büyük bir hızla artmaktadır.

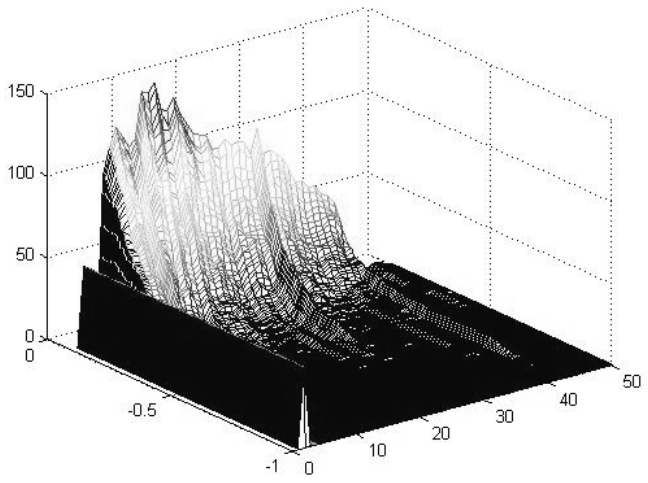


Şekil 4.3: Benzerlik ölçütlerinin belli seviyelerde bulduğu kural sayıları

Kosinüs ve korelasyon ölçütleri ile bulunan katsayıların negatif veya pozitif olan ilişkileri gösterdiği düşünülmektedir. Ancak bu ölçütlerin hangi seviyelerinin kabul edilebileceği bir öngörü olacaktır. Hiçbir ilişkinin bulunmaması durumunun katsayıların -0.1 ile 0.1 arası değerlerde iken olduğu kabul edilmiştir. -0.1 seviyesinden küçük olan katsayılar negatif ilişkileri, 0.1 seviyesinden büyük olan katsayılar ise pozitif ilişkileri göstermektedir. Bu seviyelerin değişimleri de ayrıca önem taşımaktadır. Negatif ilişkiler için çizdirilen Şekil 4.4 ve Şekil 4.5 yüzeyleri üzerinde ilişki sayısının, bu katsayıların değişimi ile nasıl değiştiği detaylı olarak görülebilir.



Şekil 4.4: Kosinüs benzerlik ölçütünün eşik-kural sayısı-ilişki sayısı grafiği



Şekil 4.5: Korelasyon benzerlik ölçütünün eşik-kural sayısı-ilişki sayısı grafiği

Kullanılacak kural sayısı belirlenirken hem kosinüs hem de korelasyon yüzeyleri için hem çok sayıda kuralın bulunduğu, hem de kararlılığın iyi olduğu 29 adet kuralın kullanılması iyi bir nokta olarak belirlenmiştir. Nitekim 29 adet kuralın, pareto eğrisinde müşteri davranışlarının %80' ini açıkladığı görülebilmektedir.
29 adet kuralın kullanıldığı, -0.1 seviyesinden düşük olan katsayıların negatif ilişki kabul edildiği durum için bulunan kurallar korelasyon ölçütü için Tablo 4.14 ile, kosinüs ölçütü için Tablo 4.15 ile gösterilmiştir.

Tablo 4.14 ve Tablo 4.15 üzerinden de görülebileceği üzere arada çok büyük bir fark çıkmamaktadır. Korelasyon ölçütü kullanılarak, kosinüs ölçütü kullanılarak bulunan toplam negatif ilişki sayısından sadece 3 adet fazla ilişki bulunabilmiştir.

Pozitif ilişkiler konusunda da durum farklı değildir. Korelasyon benzerlik ölçütünün 311 adet pozitif ilişki bulmasına karşılık kosinüs benzerlik ölçütü 297 adet pozitif ilişki bulmuştur. Bulunan ilişkiler hemen hemen aynı ilişkilerdir. Ayrıca çalışmada apriori algoritması ile 301 adet pozitif ilişki bulunmuştur. Bulunan sonuçlar arasındaki farklar sonuçlar bölümünde yorumlanmıştır.

Tablo 4.14: Korelasyon ölçütü ile bulunan negatif ilişkiler

3 8	15 20	11 26	11 31	22 37	11 44	28 47	22 49
3 11	11 21	12 26	12 31	25 37	12 44	37 47	25 49
6 11	12 21	18 26	20 31	17 38	20 44	39 47	31 49
3 12	20 21	21 26	26 31	26 38	26 44	46 47	35 49
6 12	11 22	22 26	7 33	27 38	37 44	26 48	43 49
3 15	12 22	23 26	30 33	29 38	1 46	29 48	44 49
8 15	20 22	25 26	31 33	22 38	9 46	35 48	47 49
11 15	1 23	17 28	30 34	9 39	21 46	47 48	1 50
12 15	11 23	27 28	8 35	16 39	25 46	4 49	9 50
15 16	20 23	20 30	11 35	36 39	34 46	6 49	14 50
14 18	11 25	26 30	31 35	39 40	8 47	14 49	15 50
2 20	12 25	7 31	6 37	7 43	14 47	15 49	31 50
6 20	20 25	8 31	21 37	23 43	18 47	21 49	34 50
36 50							

Tablo 4.15: Kosinüs ölçütü ile bulunan negatif ilişkiler

3 8	11 21	6 26	12 31	22 37	11 44	39 47	31 49
3 11	12 21	11 26	20 31	25 37	12 44	46 47	35 49
6 11	20 21	12 26	26 31	17 38	20 44	1 48	43 49
3 12	11 22	22 26	7 33	26 38	26 44	27 48	44 49
6 12	12 22	23 26	30 33	27 38	37 44	32 48	47 49
3 15	20 22	25 26	31 33	29 38	1 46	35 48	1 50
8 15	1 23	17 28	8 35	32 38	9 46	47 48	9 50
11 15	11 23	27 28	11 35	9 39	21 46	4 49	15 50
12 15	20 23	20 30	31 35	16 39	25 46	6 49	24 50
6 16	11 25	26 30	30 36	36 39	36 46	15 49	31 50
15 16	12 25	7 31	6 37	39 40	6 47	21 49	36 50
6 20	20 25	8 31	17 37	7 43	8 47	22 49	15 20
3 26	11 31	21 37	23 43	37 47	25 49		

5. SONUÇLAR

Yapılan çalışmada pozitif ve negatif ilişkilerin bulunması için kullanılabilecek yeni bir yöntem önerilmektedir.

Apriori algoritması ile yapılan çalışma sonucunda bulunan pozitif ilişkiler ile geliştirilen yöntem ile bulunan pozitif ilişkiler karşılaştırılmıştır. Aynı olarak tespit edilen ilişki oranı %30 civarındadır. Geliştirilen yöntem ile bulunan pozitif ilişkilerin büyük bir bölümü tek yönlü veya iki yönlü olarak bakıldığında yüksek güven değerlerine sahiptir. Ancak apriori algoritması ile bulunan ilişkilerin hem tek yönlü olması hem de güven ile destek değerlerinin düşük olması muhtemeldir. Ayrıca apriori algoritmasının harcama verileri üzerinden değil de model ürünlerin alınma durumu üzerinde çalıştığı bilinmektedir. Bu da ayrıca fiyatın etkisinin gözardı edilmesi anlamına gelmektedir. Bu duruma bir örnek vermek yararlı olacaktır. Müşteri iki adet ürün beğenmektedir. Ancak bu tür üründen sadece bir adet ürünü alabilecek kadar para harcamak istemektedir. Müşteri ürün seçimini hangi model daha ucuz ise o modelden yana kullanmaktadır. Belki de iki ürün birlikte indirimli olarak satışa sunulur ise müşteri iki ürünü de alacaktır. Bu durumu apriori algoritması bir harcama verisi değil de sadece ürünün alınıp alınmama durumunu göz önüne aldığı için belirleyememektedir. Ayrıca ürünlerin fiyatlarının birbirlerine oranla düşük veya yüksek bulunup, zaman ilerledikçe stok miktarına göre indirime gidilen durumlar da mevcuttur. Apriori algoritması bu tip durumların hiçbirisini göz önüne almamaktadır.

Geliştirilen yöntemin benzerlik verilerini kullanması ve harcama verileri ile çalışabilmesi yöntemin iyi yönleridir. Ancak sonuçların doğruluğu için apriori algoritması ile karşılaştırmada %30 gibi bir tutarlılık durumunun ortaya çıkması sonuçların çok iyi olamayabileceğini göstermektedir.

KAYNAKLAR

Akpınar H., "Veri Tabnalarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi, ***İşletme Fakültesi Dergisi***, C29, S:1-22, (2000).

Alataş, Bilal – Akın, Erhan; Veri Madenciliğinde Yeni Yaklaşımlar, ***Ya/Em-2004-Yöneylem Araştırması/Endüstri Mühendisliği XXIV Ulusal Kongresi***, 15-18 Haziran, Gaziantep-Adana (2004)

Aydoğan F., "E-ticarette veri madenciliği yaklaşımlarıyla müşteriye hizmet sunan akıllı modüllerin tasarımı ve gerçekleştirimi", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 12-16 (2003).

Berson,Alex- Smith, Stephen- Thearling, Kurt, ***Building Data Mining applications for CRM***, McGraw- Hill, USA, (2000).

Bilen Ö., "ÖSS Sınav Sonuçlarının Okul Bazında Veri Madenciliği ile İncelenmesi", Yüksek Lisans Tezi, ***Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul,10-13 (2004).

Davis B., "Data mining transformed", ***Information Week***, 751: 86 (1999).

Elden L., "Matrix Methods in Data Mining and Pattern Recognition", ***SIAM***, 23-26, 57, (2007)

Freitag alex A., Data Mining and Knowledge Discovery with Evolutionary algorithms***, Springer - Verlag Berlin Heidelberg***, Germany, (2002).

Han, J. And Kamber, M. , "Data Mining- Concept, Techniques, academic" ***PRESS, USA***, 550p (2001).

Hui S., Jha G., "Application data mining for customer service support", ***Information and Management***, 38: 1-13 (2000).

İnternet: "Hipotez Testi – Vikipedi", Vikipedi özgür ansiklopedi, http://tr.wikipedia.org/wiki/Hipotez_testi .**(Ziyaret tarihi: 18 Temmuz 2009)**.

İnternet: "Veri Madenciliği – Vikipedi", Vikipedi özgür ansiklopedi, http://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi.**(Ziyaret tarihi: 25 Temmuz 2009)**.

İnternet: "Veri Tabanı – Vikipedi", Vikipedi özgür ansiklopedi, http://tr.wikipedia.org/wiki/Veri_taban%C4%B1 .**(Ziyaret tarihi: 25 Temmuz 2009)**.

Javovic, N.- Milutinovic, V. –Obradovic, Z. “Foundations of Predictive Data Mining, Member”, ***6th seminar on neural network applications in electrical engineering Neural,*** 53, -58,IEEE, (2002).

Kartal, M., “Bilimsel Araştırmalarda Hipotez Testleri”, ***Nobel Yayın Dağıtım***, (2006)

Korn F., Labrinidis A., Kotidis Y., Faloutsos C., “Quantifiable data mining using ratio rules”, ***VLDB Journal***, 8, 254-266 (2000).

Ronald Swift; Accelerating Customer Relationship; ***Prentice Hall PTR***,(2001).

Rushing, J., 1997, CS 687 Technology Assessment Paper, www.cs.uah.edu/~thinke/CS 687/Fall 97/Tech/Rushing.html **(Ziyaret tarihi:18.Mart 2008).**

Savasere A., Omiecinski E., Navathe S., “Mining for Strong Negative Associations in a Large Database of Customer Transactions”, ***IEEE Computer Society***, 494 – 502 (1998).

Tan, P. N., Steinbach, M., Kumar, V., “Introduction to Data Mining”, ***Pearson International Edition***, 73-77, (2006).

Tantuğ A.C. ,” Veri Madenciliği ve Demetleme” , Yüksek Lisans Tezi, ***İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul,3 (2002).

Wei C., Chiu T., “Turning telecommunications call details to churn prediction: a data mining approach,”, ***Expert Systems with Applications***, 23: 103-102 (2002).

ÖZGEÇMİŞ

1983 yılında Ankara' da doğdu. İlkokul öğrenimini Ankara'da, ortaokul ve lise öğrenimini İstanbul' da tamamladı. 2002 senesinde girdiği Kocaeli Üniversitesi Endüstri Mühendisliği bölümünden, 2007 yılında mezun oldu. 2007 yılında Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği anabilim dalında yüksek lisans öğrenimine başladı. Yüksek lisans öğrenimi sırasında TÜBİTAK destekli 107M257 nolu projede yer aldı. Yayınları arasında metasezgisel optimizasyon öne çıkmaktadır. İleri seviyede C ve matlab bilmektedir. Hobi olarak amatör telsizcilik ile uğraşmaktadır.