

BÜYÜK ÖLÇEKLİ VERİ TABANLARINDA BİLGİ KEŞFİ

Şühedanur KAVURKACI¹, Zeynep GÜRKAŞ AYDIN², Rüya ŞAMLI³

^{1,2,3} İstanbul Üniversitesi Bilgisayar Mühendisliği Bölümü

¹ sskavurkaci@gmail.com, ² zeynepg@istanbul.edu.tr, ³ rsamli@istanbul.edu.tr

Özet: Verilerin dijital ortamlarda depolanmasının kolaylaşması ve yaygınlaşmasıyla birlikte, elde bulunan veri miktarı çok büyük boyutlara ulaşmıştır. Bu veri yığınlarından anlamlı bilgileri edinmek hayati önem taşımaktadır. Ancak geleneksel yapılarla ve sorgularla büyük veri yığınlarından anlamlı bilgi edinmek neredeyse imkansız hale gelmiştir. Bu yüzden, büyük boyutlardaki verileri işleyebilen teknikleri kullanmak gerekmektedir. Veri madenciliği, böyle durumlarda kullanılan, büyük boyutlardaki veri yığınlarından yararlı olabilecek saklı örüntüleri bulma işlemidir. Bu çalışmada veri madenciliğine genel bir bakış sunulacaktır.

Anahtar Sözcükler: Veri Madenciliği, Bilgi keşfi, Örüntü Bulma

1. Giriş

Günümüzde bilişim teknolojileri yaşantımızın büyük bir kısmında yer almaktadır. Büyük kapasiteli veri kayıt cihazlarının maliyetlerinin düşük olması da bireysel olarak veri kayıtlarının artmasına olanak sağlamıştır. Ancak veri depolama alışkanlığı sadece bireysel olarak yaygınlaşan bir alışkanlık değildir, milyonlarca şirket, veri depolama alışkanlığını kazanmış durumdadır.

Veri kayıtlama için kullanılacak cihazların özellikle internet içerisinde kullanılıyor olması, internet üzerinde internet kullanıcılarının internet kullanım süresi içerisindeki tercihlerini, kararlarını, alışverişlerini, ve alışkanlıklarını veri olarak depolama kolaylığı sağlamıştır. Bu durumda internet kullanıcılarının her adımı, veri tabanındaki bir kayda karşılık gelir.

Bütün bu gelişmeler; verilerin çokluğuyla verileri anlamlandırma kavramları arasında büyük bir uçuruma sebep olmuştur. Depolanmış veri miktarının artmasına karşılık, bu yığınlardan anlam çıkarılma ihtimali gittikçe azalmıştır. Bu nedenle geleneksel yöntemlerle veri yığınları arasında saklı bulunan ve özellikle şirket stratejileri için yararlı olabilecek bilgileri bulmak imkansız hale gelmiştir. Veri madenciliği çalışmaları bu noktada anlam kazanmıştır.

Aslında ekonomistler, istatistikçiler ve iletişim mühendisleri gibi bazı mesleklerdeki kişiler uzun zamandır veriler içinde otomatik olarak bulunabilecek, tanımlanabilecek ve tahminler için kullanılacak örüntü kavramı üzerinde çalışmaktadırlar. Diğer bir deyişle bu konu üzerindeki çalışmalar yeni değildir. Ancak iş sahasında bu alanda çalışmalara son zamanlarda önem vermeye başlanmıştır. Çünkü şirketlerin müşterilerinin bu sahaya ilişkili her hareketi kayıt altına alınmaya başlamıştır ve elde bulunan bu kayıtlardan müşteriye hizmet ve pazarlama noktasında bir çok fikir alınması mümkün olabilir.

Bütün bu nedenlerin toplamında, bir çok şirketin çok büyük boyutlarda sahip olduğu veri yığınlarını anlamlı hale getirmeleri için tek umut olarak veri madenciliği görülmeye başlanmıştır [1].

Bu çalışmanın ilk bölümünde veri madenciliğinin ve ilgili bazı kavramların tanımı yapılacaktır. Veri madenciliğinin farklı alanlardaki kullanımları hakkında bilgiler verilecektir. Ardından veri madenciliğinde kullanılan modellere değinilecektir. İkinci bölümde, bir veri madenciliği projesinin yaşam döngüsü, üçüncü bölümde veri madenciliğinin mevcut pazardaki yeri ve standartları, dördüncü bölümde günümüzde oluşmakta olan trendler anlatılacaktır.

2. Veri Madenciliği ve Modelleri

Bu bölümde veri madenciliğinin tanımı ve modellerine değinilmiştir.

2.1 Veri Madenciliği Nedir?

Veri madenciliği; OLAP (Online Analytical Processing), kurumsal raporlama ve ETL (Extract-Transform-Load) ile birlikte İş Zekası ürün ailesinin en önemli üyesidir. Veri madenciliği, büyük veritabanları içerisinde kullanıldığı alan için önemli olan ancak kolay anlaşılacak örüntülerin çıkarılması demektir. Bu örüntülerin bir diğer özelliği de kullanılabilir olmalarıdır. Örüntüler yapıları gereği içinde ilişkiler, kurallar, değişim düzenleri bulundurlar ve istatistiksel olarak değerlidirler. Bu örüntüleri veritabanlarında keşfetmek için otomatik veya kimi zaman yarı otomatik yöntemler kullanılır.

Veri madenciliği aynı zamanda, veri tabanlarında bilgi keşfi olarak da adlandırılabilir [2][3].

Bu yöntem ile müşteri istek ve gereksinimlerinin öncelikleri belirlenmekte ve müşteri sesine göre ürünle ilgili özellikler önem sırasına göre

sıralanmaktadır. Böylece tasarımcı ürünün tasarımında teknik ya da estetik nedenlerden dolayı müşteri istek ve gereksinimleri arasında tercih yapma durumunda kaldığı zaman, bu sıralamayı incelemekte ve bu sıraya göre ürünü tasarlamaktadır. Bunun sonucunda, hem zaman kaybı önlenmekte hem de ürün en çok istenilen özellikleri içerecek şekilde tasarlanmakta ve üretilmektedir.

2.2 Veri madenciliğinin kullanım nedenleri ve alanları

Veri madenciliğinin günümüzde önemli olmasının nedenleri aşağıdaki şekilde sıralanabilir:

- Ulaşılabilen veri miktarının çok büyük hacimde olması nedeniyle bu verilerin içindeki yararlı bilgilere geleneksel yollarla ulaşılabilir değildir.
- Pazarda artan rekabet nedeniyle müşterilere daha iyi hizmet verme ihtiyacı duyulmaktadır. Hizmet kalitesini artırmanın bir yolu da müşteriyi iyi anlamak ve alışkanlıklarını iyi tespit etmektir bu alışkanlıkları tespit aşamasında veritabanlarında bulunan müşteri kayıtlarından yararlanılabilir.
- Günümüzde veri madenciliği için kullanılacak teknolojilere ulaşım kolaylığı artmıştır. Bu nedenle bir çok kurum veri madenciliği hakkında yeterli bilgiye sahip olup veri madenciliği konusunda ürün sağlayan firmalarla iletişime geçmektedir [1].

Veri madenciliği kullanım alanlarına şu örnekler verilebilir:

Pazarlama: Müşteri satın alma kayıtlarından çıkarılan örüntülerle ürün düzenlemesi yapılabilir, müşteri alışkanlıkları incelenerek müşteri ilişkileri düzenlenebilir.

Bankacılık: Kredi kartı dolandırıcılıkları tespiti ve kredi talepleri değerlendirilmesi yapılabilir.

Sigortacılık: Yeni poliçe talep edebilecek müşteriler belirlenebilir.

Tıp: Salgın hastalıklarla ilgili tespitler yapılabilir [4].

2.3 Veri Madenciliği Modelleri

Veri madenciliği modelleri, farklı şekillerde sınıflandırılabilir. Bu sınıflandırmanın ilkinde modeller tanımlayıcı (descriptive) ve tahmin edici (predictive) olarak iki gruba ayrılabilir.

Tahmin edici modellerde, sonuçları bilinen verilerden hareket ederek bir model geliştirilir ve

kurulan model üzerinden sonuçları bilinmeyen veri kümeleri için sonuç değerleri tahmin edilir.

Tanımlayıcı modellerde ise karar vermeye yardımcı olacak mevcut verilerdeki örüntüler tanımlanır.

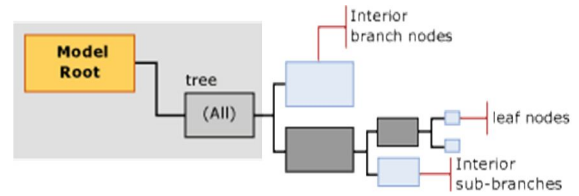
Sınıflandırma yöntemlerinin diğerinde ise modeller aşağıdaki şekilde beş gruba ayırmak mümkündür [1][5].

- Sınıflama (Classification) ve Regresyon (Regression)
- Kümeleme (Clustering)
- Birliktelik (Association)
- Dizi analizleri (Sequence Analysis)
- Sapma analizleri (Deviation Analysis)

2.3.1 Sınıflama Ve Regresyon

Sınıflama ve regresyon modelleri veri madenciliğinde en çok kullanılan modellerdir. Veriler, belli sınıflara ayrılır ve bu sınıflar üzerinde örüntü tespitleri yapılır. Bu yöntemle müşteriler belli sınıflara ayrılıp sınıfların özellikleri üzerinden yeni stratejiler geliştirilebilir. Risk analizleri yapılabilir.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler; karar ağaçları (Decision Trees), yapay sinir ağları (Artificial Neural Networks), genetik algoritmalar (Genetic Algorithms), K-En yakın komşu (K-Nearest Neighbor), bellek temelli nedenleme (Memory Based Reasoning), Naive-Bayes olarak gruplandırılabilir.



Karar Ağaçlarının genel yapısı

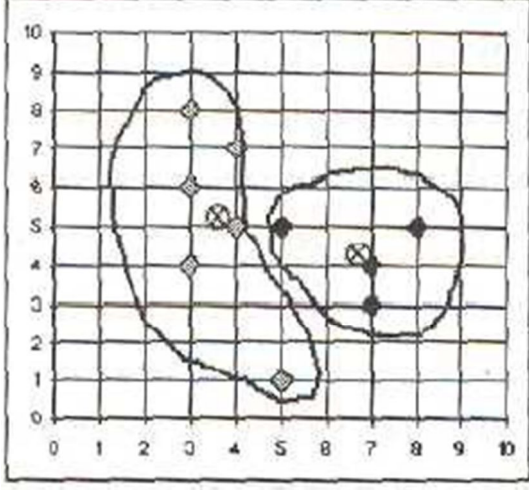
2.3.2 Kümeleme

Birbirinden çok farklı özelliklere sahip olan kümelerin bulunması yöntemidir. Kümelerin içindeki elemanlar birbirlerine benzer özellikler gösterirler. Veriler herhangi bir sınıfa dahil değildir. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin önişlemi olarak kullanılabilir. Marketlerde farklı müşteri gruplarının keşfedilmesi ve bu grupların alışverişle ilgili örüntülerinin bulunması bu uygulamaya bir örnektir.

Kümeleme yöntemleri şu şekilde sınıflandırılabilir:

- Bölme yöntemleri (Partitioning methods)
- Hiyerarşik yöntemler (Hierarchical methods)

- Yoğunluk tabanlı yöntemler (Density-based methods)
- Grid tabanlı yöntemler (Grid-based methods)
- Model tabanlı yöntemler (Model-based methods) [6]



Şekil 2: Kümeleme [6]

2.3.3 Birliktelik Kuralları

Birliktelik kuralları analizine aynı zamanda pazar sepeti analizi de denebilir. Eş zamanlı olarak gerçekleşen olayları inceler.

Bir müşterinin yaptığı tüm alışverişlerdeki ürünler arasındaki birliktelikler bulunarak müşterinin satın alma alışkanlıkları analiz edilebilir. Bu tür bilgilerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisi gibi bilgileri ortaya çıkarır ve market yöneticileri bu bilgiler ışığında etkin satış stratejileri geliştirmeye başlarlar.

Büyük veri tabanlarında birliktelik kuralları bulunurken şu iki işlem yapılmalıdır:

- a. Sık tekrarlanan öğelerin bulunması
- b. Sık tekrarlanan öğelerden güçlü birliktelik kurallarının oluşturulması [1][6]

2.3.4 Dizi Analizleri

Dizi analizleri farklı serilerde örüntüler bulmak için kullanılan yöntemlerdir. Bir dizi, farklı değerler serilerinden oluşur. Örneğin bir DNA dizisi A, G, C ve T gibi 4 farklı durumun farklı dizilmesiyle oluşan serilerin birleşimidir.

Dizi analizleri ve birliktelik kuralları analizleri arasında belirli durumların kümeleri üzerinden işlem yapılması sayesinde bir benzerlik vardır denilebilir. Ancak dizi analizleri, durumlar arası geçişleri incelerken birliktelik kuralları analizleri eş zamanlı ve birbirinden bağımsız oluşan durumları inceler [1].

2.3.5 Sapma Analizleri

Milyonlarca işlem arasından anormal olan durumları bulmak ve tanımlamak çok zor bir işlemdir. Diğerlerinden farklı davranan bu anormal durumları bulmak için sapma analizleri kullanılabilir. Bu yöntemin en çok kullanıldığı yerlerden biri kredi kartı dolandırıcılıklarını tespit etme sürecidir. Ayrıca, ağız boş yere işgal edilip edilmediğini kontrol ederken, üretim hatalarını analiz ederken de kullanılabilir.

Bu yöntem görselleştirme veya istatistiksel tekniklerle uygulanabilir. Doğrusal regresyon yöntemi de analiz işlemi için uygun olan bir diğer yöntemdir. Bu yöntemin en çok bilinen uygulaması istisna sapmasıdır. İstisna sapması, kredi kartı yolsuzluklarının tespiti için yaygın olarak kullanılan yöntemlerden biridir.[10]

Sapma analizi için standart bir teknik henüz bulunamamıştır. Bu teknik için araştırmalar devam etmektedir [1].

3. Veri Madenciliği Proje Döngüsü

Bir veri madenciliği proje döngüsü veri toplama işlemi ile başlar. Bu verilerin temizlenmesi ve yeniden yapılandırılmasıyla model keşfine hazır hale gelir. Uygulanılan birtakım yöntemlerle proje için uygun olabilecek model adayları belirlenir ve geçerlilik testleriyle en uygun model bulunmaya çalışılır. Model bulunduktan sonra raporlamaya geçilir. Raporlanan verilere göre tahminler yapılır. Tahminlerin ardından uygulamanın entegrasyonu gerçekleştirilir ve uygulamanın kullanıldığı süreç içerisinde zaman aralıklarıyla uygulama gözden geçirilir.

3.1 Veri Toplama

Veri madenciliği projeleri verinin toplanması ile başlar [7]. Bu yüzden veri toplama, en önemli aşamalardan biri olarak ifade edilebilir.

Verinin toplanmasında kullanılan kaynaklar ve veri türleri şunlardır:

- İlişkisel veritabanları
- Veri ambarları
- İşlemsel veritabanları
- Uzaysal veritabanları
- Metin veritabanları ve Multimedya veritabanları
- İnternet

3.2 Verinin Temizlenmesi ve Yeniden Yapılandırılması

Bu aşama yoğun bir şekilde, veri kaynağıyla ilgili işlemleri içerir. Verinin temizlenmesi verinin gürültülerden arındırılması diğer bir deyişle yanlış ya da uç değerlere sahip verilerin temizlenmesi

anlamına gelir.

Verinin temizlenmesi ve yeniden yapılandırılmasında kullanılan yöntemler

- Veri türünün dönüştürülmesi
- Sürekli kolonların dönüştürülmesi
- Gruplama
- Kümeleme
- Kayıp verilerin işlenmesi
- Uç verilerin ortadan kaldırılması olarak özetlenebilir.

3.3 Model Oluşturma

Verilerin arasındaki gürültüler temizlendikten ve değişkenler düzenlendikten sonra elde bulunan veriler model oluşturmada kullanılabilir hale gelmiş olur. Model oluştururken projenin hedeflerinin ne olduğu ve bu hedeflere yönelik ne tür verilerin kullanılacağı net olarak belirlenmiş olmalıdır. Projede ne tür bir model kullanılacağına karar verilmelidir. Bu kararların verilmesinin ardından proje türüne uygun olarak veri madenciliği algoritması seçilir.

Yukarıda anlatılan süreç en uygun modeli bulana kadar yinelenen bir süreçtir [7].

3.4 Modelin Keşfi

Bir önceki bölümde anlatılan model oluşturma yöntemleri ile projelerde farklı modeller bulunması söz konusu olabilir. Ancak bunlar arasındaki en doğru modeli bulmak oldukça güç bir işlemdir. Bu amaçla bulunan modelleri test etmek için bazı araçlar kullanılır.

a. Basit Geçerlilik (Simple Validation):

Modelin test edilmesinde kullanılan en basit yöntem basit geçerlilik (Simple Validation) yöntemidir. Bu yöntemde verilerin %5 ile %33 arasındaki kısmı test verileri olarak ayrılır. Kalan kısım üzerinde model bulmak için algoritmalar uygulanır. Model bulunduktan sonra daha önceden ayrılmış test verileri üzerinden test işlemleri yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının bütün olay sayısına oranı hata oranı olarak ifade edilir. Doğru olay sayısının bütün olay sayısına oranı ise doğruluk oranı olarak adlandırılır. (Doğruluk oranı = 1 – Hata oranı)

b. Çapraz Geçerlilik (Cross Validation):

Sınırlı sayıda veri bulunması durumunda çapraz geçerlilik yöntemi kullanılabilir. Bu yöntemde veriler rasgele eşit iki parçaya ayrılır. İlk olarak birinci parça üzerinden model tespiti, ikinci parça üzerinden test işlemi; daha sonra da ikinci parça

üzerinden model tespiti, birinci parça üzerinden test işlemi yapılarak hata oranlarının ortalaması kullanılır [5].

c. N-Katlı Çapraz Geçerlilik:

Bu yöntemde veriler n parçaya ayrılır ve çapraz geçerlilikte uygulanan yöntem n parça için uygulanır. n gruptan bir tanesi test için, kalan n-1 tanesi model tespiti için kullanılır. Bu yöntem birkaç bin veya daha az satırdan meydana gelmiş veri tabanlarında kullanılmak için daha uygundur [5].

d. Bootstrapping:

Küçük veri kümeleri için modelin hata düzeyinin belirlenmesinde kullanılan tekniklerden biridir. Model bütün veri kümesi üzerinde belirlenir. En az 200, olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanabilir [8].

e. Risk matrisi (Risk Matrix):

Modelin doğruluk derecesinin değerlendirilmesinde kullanılan yöntemlerden biridir. Özellikle sınıflama problemlere için kurulan modellerde kullanılmaya yatkın bir tekniktir. Aşağıda örneği görülen matrisin sütunları fiili (gerçek hayattaki), satırları ise tahmini değerleri içerir. Örneğin fiilen B sınıfına ait olması gereken 46 elemanın, kurulmuş olan model tarafından 2'sinin A, 38'inin B, 6'sının C olarak sınıflandırıldığı görülmektedir [11].

	Fiili		
Tahmini	A sınıfı	B sınıfı	C sınıfı
A sınıfı	45	2	3
B sınıfı	10	38	2
C sınıfı	4	6	40

f. Kaldıraç oranı (Lift chart):

Modelin sağladığı faydanın değerlendirilmesinde kullanılan bir yöntemdir. Lift chart ile değerlerin tahmin edilmesi için model eğitilir ve veri kümesi test edilir. Değerlerin tahmin edilmesi ve hesaplanan olasılıkların grafiksel olarak gösterilmesi sonucu model durumu görülebilir.

3.5 Raporlama

Raporlama işlemi, model keşfinin ardından, modelin sonuçlarının anlaşılır şekilde görülebildiği yöntemlerden biridir. Veri madenciliği için kullanılan uygulamaların büyük kısmında raporlama için kullanılabilir araçlar bulunmaktadır. Bu araçlar yardımıyla grafiksel ve metinsel raporlar elde edilebilir.

Elde edilecek bu raporlar bulunan örüntülerin raporları olabileceği gibi, model yoluyla yapılan tahminlerin raporları da olabilir. [7]

3.6 Tahmin (Skorlama)

Veri madenciliği projelerinde model keşfedilmiş olması, başka bir deyişle işe yarayacağı düşünülen örüntünün bulunmuş olması projenin tamamlandığı anlamına gelmez. Çünkü veri madenciliği projesine başlanmasının amaçlarının arasında projeden elde edilecek sonuçlara göre yeni stratejiler geliştirmek, ya da önemli tespitlerde bulunmak gibi amaçlar vardır. Bahsedilen amaçlara ulaşmak için veri yığınlarından çıkarılmış olan örüntüler üzerinden tahminler yapılmalı ve bu tahminlere göre bir sonraki adıma geçilmelidir. Bu tahminler mevcut durumla ilgili tahminler olabileceği gibi geleceğe yönelik tahminler de olabilir. Geleceğe yönelik tahminlere, bir sonraki ay için yapılacak satış tahmini örneği verilebilir.

3.7 Uygulamanın Entegrasyonu

Bu adım tüm projenin zeka kısmını oluşturur. Ayrıca analiz döngüsünün son adımıdır. Veri madenciliği uygulamaları bir çok uygulama içine yerleştirilebilir. Bulunan örüntülerle kurulan modeller risk analizi, dolandırıcılık tespiti gibi uygulamalarda doğrudan kullanılabilir [7].

3.8 Modelin Yönetimi

Modelin oluşturulmasının ve entegrasyonunun ardından eğer model statik olarak kullanılmıyorsa kurulan bu modelin kullanım süresi boyunca izlenmesi gerekir. Çünkü zaman içerisinde bu modellerin entegre edildiği sistemlerin özellikleri ve ürettikleri veriler değişebilir. Bu değişimlere göre kullanılan model yeniden düzenlenmelidir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılabilir. Bu grafiklere göre model düzenlemesine gidilir [7].

4. Veri Madenciliği Pazarı ve Standartları

Bu bölümde veri madenciliği pazar durumuna ve veri madenciliği uygulamalarının standartlarına değinilmiştir.

4.1 Veri Madenciliği Pazarı

Şirketlerin veri madenciliğini daha etkin kullanmaya başlamasıyla veri madenciliği uygulamaları daha çok yayılmaya başlamıştır, bu da sürekli genişleyen bir pazar oluşumunu beraberinde

getirmektedir.

Veri madenciliği hizmeti veren bazı büyük firmalar ve ürünleri şunlardır:

SAS (Statistical Analysis System): Pazar payının büyük bir kısmını elinde tutan SAS firmasının piyasaya çıkardığı ürünler, veri analizleri için kullanılabilir bir çok istatistiksel fonksiyon barındırmaktadır. Ayrıca, SAS Script adı verilen çok güçlü bir script diline sahiptir [12].

SPSS (Statistical Package for the Social Sciences): SPSS, SPSS tabanı ve Karar Ağaçları gibi ürünleri de içinde bulunduran bir çok veri madenciliği ürününe sahiptir. SAS gibi istatistiksel alanda kullanılabilir en iyi ürünleri içinde barındırır [13].

IBM (International Business Machines Corporation) : IBM, Intelligent Miner adını verdiği veri madenciliğinde kullanılabilir bir ürün sunmaktadır. Intelligent Miner, bir çok algoritma ve görüntüleme araçlarını içerir. Ayrıca Data Mining Group (DMG) tarafından tanımlanmış olan Predictive Modeling Markup Language (PMML) için de veri madenciliği modelleri üretmiştir [14].

PMML dosyaları, örüntü modellerini ve düzenlenmiş veri setlerinin istatistiklerini içeren XML dosyalarıdır.

Microsoft Corporation: Microsoft, bir ilişkisel veritabanı içerisinde veri madenciliği kullanımı imkanını sağlayan ilk veri tabanı üreticisidir. SQL Server 2000 içine iki veri madenciliği algoritması eklenmiştir. Bunlardan biri Karar Ağaçları algoritması, diğeri de kümeleme algoritmasıdır. Algoritmalarının dışında SQL Server'ın en önemli veri madenciliği özelliği ise, OLE DB'yi SQL Server içine gömmüş olmasıdır [15].

Oracle: 2000 yılında üretilmiş olan Oracle 9i sürümünde iki veri madenciliği algoritması gömülmüştür. Bu algoritmalar, birliktelik kuralları analiz algoritması ve Naive Bayes algoritmasıdır. Daha sonraki yıllarda piyasaya sürülen Oracle 10g sürümünde Oracle içerisinde daha fazla veri madenciliği araçları ve algoritmaları bulunmaktadır [16].

Angoss: Angoss ürettiği Knowledge Studio, karar ağaçları üretme, küme analizi ve bir çok tahmin edici modeli barındırmaktadır. Bu özellikler, kullanıcıya verilerini farklı açılardan göstererek kullanıcının bu verileri anlamasını da sağlamıştır. Ayrıca, bu araç buluşları açıklamak ve desteklemek için çok güçlü görüntüleme araçlarını da içermektedir [17].

KXEN: Merkezi Fransa'da bulunan bu firmanın ürünü olan SVM, regresyon, zaman serileri, kümeleme gibi bir çok özelliği içeren veri madenciliği algoritmalarına sahiptir. Ayrıca, bu ürün OLAP küpleri için çözümler de üretmiştir. Bunun yanı sıra, kullanıcılara Excel'in alışılmış özelliklerini kullanarak veri madenciliği yapabilecekleri bir eklenti çıkarmıştır [2][18].

4.2 Veri Madenciliği Standartları

Önceki yıllarda her firmanın kendine özel API (Application programming interface) standartları olması sonucunda farklı veri depoları için aynı sorgulama biçimleri kullanılması mümkün değildi. Ancak son yıllarda bağımsız firmalar tarafından üretilen ürünler veri madenciliği için depolama, API ve içerik için belirli standartlar oluşturdu [2].

Bu ürünlerden bazıları ise şunlardır:

- DM ve XML (Extensible Markup Language) için OLE DB
- Veri madenciliği için SQL/Multimedia
- Java veri madenciliği API'si
- Predictive Model Markup Language
- Crisp-DM
- Common Warehouse Metadata

5. Günümüzdeki Veri Madenciliğinin Eksiklikleri ve Yeni Trendleri

Bu bölümde günümüzde veri madenciliğinin hangi konuda yetersiz kaldığından bahsedilecek, ardından gelecek yıllarda gerçekleşme ihtimali yüksek olan gelişimler anlatılacaktır.

5.1 Veri Madenciliği Eksiklikleri

Veri madenciliği son yıllarda çok fazla konuşuluyor olsa da hala pazarda olması gerekenden daha küçük bir yere sahiptir. Veri madenciliğini kullananların çok büyük kısmı analistlerdir. Veri madenciliği hala büyük ve zorunlu olmayan bir uygulama gibi görüldüğü için çoğu kurum ve çalışanlar tarafından ilgi görmemektedir. Çoğu geliştirici için anlaşılması zor geldiğinden az sayıda kurumsal uygulama veri madenciliği özellikleri içermektedir.

Veri madenciliğinin istenen düzeyde kullanıma sahip olması için üstesinden gelmesi gereken bazı sorunlar vardır. Bunlar:

- Standart bir API'ye (Application programming interface) sahip olan yatay paketlerin bulunmaması,
- Geliştirici merkezli değil analist merkezli olması,
- Kullanıcılar için yetersiz eğitim,
- Yetersiz algoritma özellikleri,

olarak özetlenebilir.

5.2 Veri Madenciliğinde Yeni Trendler

Her ne kadar veri madenciliği kavramı ortaya çıkmalı uzun bir süre olmuş olsa da, veritabanı sistemlerinin kullanılma süresinin yanında çok yeni bir kavram olarak kalır. Şu andaki pazar payı olabileceğinden çok daha az boyuttadır. Bunun en önemli nedenlerinden biri, veri madenciliği için kullanılacak ürünlerin sadece analistlere ve veri madenciliği eğitimi almış elemanlara hitap etmesidir. Ancak son yıllarda üreticiler geliştiriciler için yeni API'ler üretmeye başlamışlardır. Bu gelişme ile birlikte önümüzdeki dönemde geliştiricilerin veri madenciliğini projelerinde etkin olarak kullanmaları gerçekleşmesi zor bir ihtimal değildir.

Önümüzde yıllarda veri madenciliği alanındaki gelişmelerin aşağıdakiler gibi olabileceği düşünülmektedir.

- Veri madenciliği uygulamaların diğer uygulamaların içine gömülmeye başlanabilir. Böylece bir çok uygulamada veri madenciliği modelleri kullanılmaya başlanarak büyük bir ilerleme kaydedilebilir [2].
- Günümüzde kullanılan veri madenciliği ürünlerinde veri madenciliğinde kullanılan algoritmaların hepsini ya da büyük bir kısmını bir arada görmek mümkün değildir. Bu problem veri madenciliği uygulamaları geliştiren firmaların üzerinde çalıştığı bir problem olmuştur. İlerleyen dönemlerde çok fazla sayıda algoritmayı ve farklı iş çözümlerini bir arada bulunduran uygulamalar geliştirilebilir [2].
- Veri madenciliği uygulaması geliştiren firmaların hepsi kendilerine özgü, dolayısıyla farklı API'ler üzerinde bu uygulamanın kullanılmasını sağlamaktadır. Ancak hepsi ortak bir grubun üyesidir. Yani hepsi veri madenciliği grubuna dahil API'lerdir. Bu API'lerin hepsi PMML'i ortak dil olarak desteklemektedir. Bu durum da PMML'in öneminin ilerleyen dönemlerde daha da artacağına bir işaret olabilir [9].
- PMML sadece veri dönüşümünde kullanılan bir standart değildir. Aynı zamanda veri depolamak için de kullanılan bir standarttır. Bu nedenle PMML formatları, metadata depolama işlemi için de kullanılabilir [9].
- XML teknolojileri kullanarak direkt API kullanımlarına gerek duymadan, farklı uygulamalar arasında veri alışverişini kolaylaştırarak, belli bir platforma bağlı kalmaksızın veri madenciliğinin tüm süreçleri yarı-otomatik hale getirilebilir [9].

6. Sonuç

Bu çalışmada veri madenciliğine giriş niteliğinde bilgiler verilmiştir. Veri madenciliğinin, depolanmış veri setlerinin gizli örüntüleri keşfetmek ve bu örüntüleri tahminlerde kullanmak anlamına geldiği anlatılmıştır. Bir veri madenciliği projesi döngüsü adım adım açıklanmıştır. Veri madenciliği teknikleri temel olarak adlandırılmıştır. Son olarak da veri madenciliğinin pazardaki yeri ve veri madenciliği için ürün tasarlayan büyük firmalar kısaca tanıtılmıştır. Veri madenciliği standartları ve yeni trendlere göz atılmıştır.

Günümüzdeki gelişmelerden yola çıkarak veri madenciliğinin sadece büyük firmalarda değil orta ve küçük ölçekli firmalarda da kullanılabilceği görülebilir. Örneğin; belediye uygulamalarında çok daha yaygın olarak kullanılarak insanlar için daha iyi yaşam standartları sağlanabilir. Bu konuda yetkililere gereken tanıtım ve eğitim verilmelidir. Veri madenciliğinin iyi entegre edilmesi sonucunda özellikle nüfus yoğunluğu çok olan yaşam bölgelerinde rahat nefes alınabilecek çözümler üretilebilir.

Teknik olarak; veri madenciliği uygulamalarına tek bir standart getirilerek, bütün uygulamalardan bağımsız, ve bütün uygulama teknolojilerini tek bir platform üzerinde toplayarak yeni bir uygulama üretilebilir. Böylece yetkin eleman eğitimi çok daha kolay hale gelecektir.

- [1] I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2005
- [2] Z. Tang, J. MacLennan, Data Mining With SQL Server 2005, 2005
- [3]http://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi
- [4] B. Işıklı, Veri Madenciliği nedir ve nerelerde Kullanılır?, 2009
- [5] W. Lee, S. J. Stolfo, K. W. Mok, A Data Mining Framework for Building Intrusion Detection Models, 2007
- [6] S. Özekes, Veri Madenciliği Modelleri ve Uygulama Alanları, 2006
- [7] B. Gürnlü, Veri Madenciliği Projelerinin Yaşam Döngüsü, 2009.
- [8]<http://www.statistics.com/resources/glossary/b/bootstrap.php>
- [9] K. J. Cios, L. A. Kurgan, Trends in Data Mining and Knowledge Discovery, 2003
- [10] İ. Aşkın, Veri madenciliğinde modeller ve kullanımları, 2010
- [11] www.bilimselkonular.com, Karar Verme ve Veri madenciliği, 2010
- [12] www.sas.com
- [13] www.spss.com
- [14] www.ibm.com
- [15] www.microsoft.com
- [16] www.oracle.com
- [17] www.angoss.com
- [18] www.kxen.com