Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 21, Temmuz-Aralık 2003, ss. 67-76.

VERİ ÖN İŞLEME

Ayşe OĞUZLAR*

ÖZET

Veri madenciliği, (data mining-DM) son 10 yılda dünyada hızla yaygınlaşmaya başlayan bir disiplinler arası disiplin olarak göze çarpmaktadır. Günümüzde artan veri sayısı, bilgisayar kullanımının yaygınlaşması ve bilgi toplumu olma yolundaki adımlar bu disiplini daha fazla gündeme getirecektir. Yurtdışında yaygın bir kullanım alanına sahip veri madenciliği ülkemizde yeni yeni uygulanıp anlaşılmaya çalışılmaktadır. Veri madenciliğinde kullanılacak verinin kalitesi sonuçları da etkileyeceğinden kullanılacak verilerin ön işlemden geçirilmesi oldukça önemli olmaktadır. Kaliteli veriler sonuçta kaliteli çıktılar üretecektir. Verilerin kalitesini artırmanın yolu ise ön işlemden geçirilmesi yoluyla veri kalitesinin artırılması yoluyla gerçekleşebilir. Bu çalışmada veri madenciliğine ilişkin genel bir bilgi verilerek verilerin ön işleme teknikleri açıklanmaya çalışılmıştır.

Anahtar Sözcükler : Veri ön işleme, veri madenciliği.

GİRİŞ

Veriler hacim olarak sayfalarca yer kaplarlar ama kullanım değerleri azdır. Oysa, sayıları düzenleyip özetlersek, harfleri düzenleyerek anlamlı cümleler haline dönüştürürsek, notaları sıralayıp bir melodi oluşturursak ve bilgisayar ekranındaki noktaları (verileri) bir araya getirerek bir ağaç resmi veya bir grafik oluşturursak ancak o zaman verileri bilgiye dönüştürmüş oluruz. Bilgi verilere göre hacim olarak daha az yer tutar ama kullanım değeri olarak daha güçlüdür (Gürsakal, 2001:48).

Günümüzde veri tabanları artık tera byte'larla ölçülmektedir. Bu ölçekte büyük veriler, stratejik öneme sahip bilgileri gizlemektedir. Veri madenciliği, büyük veri tabanlarındaki gizli bilgi ve yapıyı açığa çıkarmak için, çok sayıda veri analizi aracını kullanan bir süreçtir. Veri madenciliğinin üç farklı bakış açısı vardır: veri tabanı bakış açısı, makine öğrenim bakış açısı ve istatistiksel bakış açısı. Yazılan kitaplar ve geliştirilen bilgisayar programları da bu farklı bakış açılarına uygun olarak yapılmaktadır (Zhou, 2002:140). Veri madenciliğinin cazibesi anlaşılmaya başladıkça bu dal ile ilgili bilgisayar programları hızla artmaktadır (Goebel and Gruenwald, 1999:1). Yine de istatistik bu alanda yapılan çalışmalarda temel bir rol üstlenmektedir.

* Yrd. Doç. Dr., Uludağ Üniversitesi, Ekonometri Bölümü.

Makalenin geliş tarihi: Nisan 2003, kabul tarihi: Şubat 2004

I. VERİ MADENCİLİĞİNDE VERİLERİN ÖNEMİ

Veri madenciliği, veri tabanı teknolojisi, istatistik, makine öğrenim, örüntü tanımı, yapay sinir ağları, verilerin görselleştirilmesi ve uzaysal veri analizi gibi farklı disiplinlerde yer alan tekniklerin bir birleşimini içerir. Bu disiplinlerin aralarındaki kesin sınırları tanımlamak zor olduğu gibi, bu alanlar ile veri madenciliği arasındaki kesin sınırları tanımlamak da zordur (Hand vd., 2001:4).

Veri madenciliğinin uygulanabilmesi için kullanılacak verilerin bir takım kesin kriterleri karşılaması gerekmektedir. Aşağıdaki başlıklar veri ve uygulamada ele alınması gereken bazı özellikleri göstermektedir:

A. VERİLER ELDE EDİLEBİLİR Mİ?

Bu soru çok açık olmasına karşın, veriler elde edilebilir olmasına karşılık kolayca kullanılabilecek bir forma sahip olmayabilir. Bu durumda verilerin uygun formlara dönüştürülmeleri gerekebilir. Farklı veri tabanı veya kaynaklardaki veriler bir araya getirilebilir. Bazen veriler kağıt üzerinde de olabilir. Bu durumda veri madenciliğine başlanılmadan önce verilerin bilgisayara girilmesi gerekecektir.

B. VERİLER İLGİLİ NİTELİKLERİ KAPSAMAKTA MI?

Veri madenciliğinin hedefi ilgili niteliklerin teşhis edilmesi olduğundan bu soru anlamsızmış gibi görünebilir. Bununla beraber hangi verilerin elde edilebilir olduğuna bakılarak kayıt edilmemiş ilgili değişkenlerin belirlenmesine çalışılabilir. Kayıt edilmemiş değişkenler veri madenciliğinin kullanışlı sonuçlar üretemeyeceği anlamına gelmemelidir. Fakat bu durumda tahminlerin güvenilirliği azalacaktır.

C. VERİLER GÜRÜLTÜLÜ MÜ?

Veriler genellikle bir hata içermektedir ve bu hatalar toplu olarak gürültü olarak adlandırılmaktadır. Tipik olarak verilerde ne kadar çok oranda gürültü varsa, o derece de güvenilir sonuçlara ulaşmak güçleşecektir. Bununla beraber makine öğrenim yöntemleri gürültü içeren verilerle çalışmak için uygun olacaktır.

D. YETERİNCE VERİ VAR MI?

Bu cevaplaması zor bir sorudur. Veri madenciliğinde, veri kümesinin büyüklüğünden ziyade veri kümesinin temsil edebilirliği önemlidir. Ele alınan ne kadar çok nitelik varsa, o kadar çok sayıda kayıta gereksinim duyulacaktır.

E.VAR OLAN VERİLER İÇİN BİLİR KİŞİ RAPORU VAR MI?

Çoğu zaman elimizdeki mevcut veriler üzerinde çalışır ve onun içeriği ile anlamını biliriz. Bununla beraber örneğin başka bir bölümün verileri ile çalışıldığında, bu verileri bilen kişilerin yardımına gereksinim duyarız. Bu kişiler veriler ile

ilgili bilgi sağlayarak, veri madenciliği sonuçlarının özetlenmesi için yardım ederler.

Klasik istatistiksel uygulamalar ve veri madenciliği arasındaki en temel farklılık, veri kümesinin büyüklüğüdür. Bir istatistikçi için 'büyük' veri kümesi birkaç yüz veya bin veri içerir. Veri madenciliği ile uğraşan birileri için ise milyon veya milyarlık veri beklenmeyen bir sayı değildir. Bu tip büyük veri tabanları gerçek hayatta sıkça ortaya çıkmaktadır.

Çok fazla sayıda değişken olduğunda daha ileri düzeyde zorluklar meydana gelmektedir. Değişken sayısı arttığında uzaydaki birim hücrelerin sayısında üstel oranda artış olacaktır. Örneğin bir tek ikili değişken ele alalım. Bu değişkenin hücrelerinin her birimde 10 gözlem olduğunu varsayarsak, toplam hücre sayısı 20 olacaktır. İki tane ikili değişken olduğunda gözlem sayısı 40 olurken, 10 tane ikili değişken için gözlem sayısı 10240 ve 20 değişken için sayı 10485760 olacaktır. Ayrıca yüksek boyut sayısına sahip uzayda en yakın noktalar birbirlerinden çok uzakta olabilmektedir. Bazı durumlarda modelin ön seçimiyle ek kısıtlamalar yapılabilir (örneğin doğrusal modellerin varsayılması gibi).

Büyük veri kümelerine erişilmesinin zorlukları açıktır. İstatistikçinin veri kütüğüne bakış açısı satırların objeleri ve sütunların da değişkenleri temsil ettiği bir veri kümesidir. Pek çok durumda veriler karışık bir haldedir ve farklı bilgisayarlara yüklenilmiş olabilir. Bu durumda verilerden bir rassal örneklem elde edilmesi için çerçevenin tanımlanması ve verilere erişimin ne kadar süreceği önemli konular olacaktır.

Veri kümesinin büyüklüğü zorluklara yol açarken, standart istatistiksel uygulamalarda sık karşılaşılmayan bir takım özellikler ortaya çıkabilir. Veri madenciliğinde veriler, veri madenciliği uygulamak üzere değil diğer bazı amaçlar için toplanmaktadır. Tersi bir biçimde, pek çok istatistiksel çalışmada veriler akıldaki belirli sorular için toplanır ve bu sorulara yanıt bulmak için analiz edilir. İstatistik, deney tasarımı ve alan araştırması gibi alt disiplinleri içermektedir. Bu disiplinler, veri toplamak için en iyi yollarla ilgili ipucu sağlarlar.

Verilerin toplanmasında ortaya çıkan problemlerin yanında, büyük veri kümeleri ile çalışılırken başka bir takım problemler de oluşabilir. Büyük veri kümeleri çoğunlukla eksik, kirli ve hatalı veri noktalarını içerecektir. Bu tip hatalara sahip olmayan veri kümeleri az rastlanılan veri kümeleridir. EM algoritması gibi bir tahmin yöntemi veya bir yerine koyma (imputasyon) yöntemi, eksik veriler için kullanılacak aynı genel dağılım özelliklerine sahip yapay veri üretilmesine yardımcı olur. Bu problemlerle standart istatistiksel uygulamalarda da karşılaşılmakla beraber, veri kümesi nispeten küçük olduğundan veri madenciliği uygulamalarındaki gibi büyük problemlere yol açmamaktadır. Özetlemek gerekirse veri madenciliğinde pek çoğu büyüklük ve ele alınan veri kümesinin doğasından kaynaklanan yeni problemler ortaya çıkmaktadır (Hand vd., 2001:19-21).

Veri madenciliğinde veri kümesinin büyüklüğünden kaynaklanan en fazla zaman alıcı aşama, verilerin ön işlemden geçirilmesi aşamasıdır. Veri madenciliği uygulamalarında kaynakların %80' i verilerin ön işlemden geçirilmesi ve temiz-

lenmesi süreçleri için harcanmaktadır (Piramuthu, 2003:1). Veri ön işleme iki farklı türde işlemi gerektirir. Bu farklı işlemlerin ilki veri kümesinin seçilmesi ve birleştirilmesi, ikincisi ise veri madenciliği için verilerin daha yararlı bir hale getirilmesi amacıyla verilerin işlenmesidir (Pyle, 1999:125). Verilerin seçimi, verilerin özellikleri ve veri büyüklüğü ile veri türü gibi teknik sınırlamaların elverdiği ölçüde gerçekleştirilmektedir (DMS Tutorial, 2001:1). Verilerin daha yararlı olması için verilerin işlenmesine ilişkin yapılması gerekenler ise asıl inceleme konumuz olan veri ön işleme teknikleri başlığı altında kısaca açıklanmaya çalışılmıştır.

II. VERİ ÖN İŞLEME TEKNİKLERİ

Veri kalitesi, veri madenciliğinde anahtar bir konudur. Veri madenciliğinde güvenilirliğin artırılması için, veri ön işleme yapılmalıdır. Aksi halde hatalı girdi verileri bizi hatalı çıktıya götürecektir. Veri ön işleme, çoğu durumlarda yarı otomatik olan ve yukarıda da belirtildiği gibi zaman isteyen bir veri madenciliği aşamasıdır. Verilerin sayısındaki artış ve buna bağlı olarak çok büyük sayıda verilerin ön işlemeden geçirilmesinin gerekliliği, otomatik veri ön işleme için etkin teknikleri önemli hale getirmiştir.

Veri ön işleme aşağıdaki sebeplerden dolayı verilere uygulanmaktadır:

1. Veriler üzerinde herhangi bir analiz türünün uygulanmasını engelleyecek veri problemlerinin çözümü
2. Verilerin doğasının anlaşılması ve anlamlı veri analizinin başarılması
3. Verilen bir veri kümesinden daha anlamlı bilginin çıkarılması.

Çok sayıda uygulamada, veri ön işlemenin bir türünden daha fazlasına ihtiyaç duyulmaktadır. Veri ön işlemenin türünün belirlenmesi bu sebeple önemli bir iştir (Famili vd., 1997:4-5).

Çok sayıda veri ön işleme tekniği mevcuttur. Bunlardan biri olan ***veri temizleme*** (data cleaning) verilerdeki gürültünün giderilmesi ve tutarsızlıkların düzeltilmesi için uygulanır. Bir diğer teknik olan ***veri birleştirme*** (data integration) ise farklı kaynaklı verileri uygun bir veri tabanında birleştirir. Normalleştirme gibi ***veri dönüştürme*** yöntemleri (data transformations) uygulanabilir ***Veri indirgeme*** (data reduction) de ise fazla olan bazı değişkenlerin atılması ve birleştirilmesi veya kümeleme yolu ile veri büyüklüğünün azaltılması amaçlanır. Bu sözü edilen veri ön işleme teknikleri, veri madenciliğinden önce uygulanarak elde edilen sonuçların kalitesi ve/veya veri madenciliği için harcanacak zaman artırılmış olur (Han and Kamber, 2001:105).

Veri ön işleme teknikleri yukarıdaki paragraftan da anlaşılabileceği gibi şu şekilde sıralanabilir:

1. Veri Temizleme
2. Veri Birleştirme
3. Veri Dönüştürme
4. Veri İndirgeme

Eksik, tutarsız ve gürültülü veriler gerçek veri tabanlarındaki ve veri ambarlarındaki yaygın özelliklerdir. Bu tür veriler literatürde kirli veriler olarak isimlendirilmektedir. Kirli verilerin geniş bir taksonomisi hazırlanmıştır (Kim, Choi, Hong, Kim and Lee, 2003:3). Eksik veriler çok sayıda sebepten kaynaklanabilir. İlgilenilen değişkenler veri tabanında bulunmayabilir. Örneğin satışlara ilişkin bir veri tabanında müşteri bilgileri yer almayabilir. Bir kısım bilgiler, verilerin girildiği zaman diliminde önemsiz bulunarak kayıt edilmemiş olabilir. İlgili veriler yanlış anlamadan dolayı kayıt edilmemiş olabilir. Tüm bu nedenlerle de tam olmayan bir veri tabanına sahip olunabilir. Veriler, bir kısım verinin silinmesinden dolayı tutarsız olabilir. Başka bir neden olarak da zaman içinde kayıt ve düzenlemelerdeki farklılıklar gösterilebilir. Verilerin gürültü içermesinin nedenleri de farklılık gösterecektir. Bir neden olarak veri toplama araçlarının yanlış kullanımı gösterilebilir. Diğer bir neden olarak veri girişindeki insan veya bilgisayar hataları sayılabilir. Veri naklinde karşılaşılan hatalar da gürültüye neden olabilecektir. Hatalı ve gürültülü veriler neden kaynaklanmış olursa olsun, verilerdeki hata ve gürültü doğru bir biçimde teşhis edilerek uygun çözümler getirilmelidir (Han and Kamber, 2001:106).

Yukarıda sayılan veri ön işleme teknikleri verilerin kalitesini artıracaktır. Kaliteli kararların kaliteli veriler gerektirdiği bir gerçektir. Bu sebeple, veri ön işleme süreci veri madenciliğinin önemli bir adımıdır. Bundan sonraki kısımlarda bu veri ön işleme teknikleri açıklanmaya çalışılacaktır.

A. VERİ TEMİZLEME

Veri temizleme, eksik verilerin tamamlanması, aykırı değerlerin teşhis edilmesi amacıyla gürültünün düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemleri gerektirmektedir. Bu başlık altında veri temizleme için temel yöntemlere kısaca değinilecektir.

Herhangi bir değişkene ilişkin eksik değerlerin doldurulması için farklı yollar vardır. Bunlardan bazıları aşağıda kısaca açıklanmaktadır (Roiger and Geatz, 2003:155):

1. Eksik değer içeren kayıt veya kayıtlar atılabilir.

2. Değişkenin ortalaması eksik değerlerin yerine kullanılabilir.

3. Aynı sınıfa ait tüm örneklemler için değişkenin ortalaması kullanılabilir. Örneğin aynı kredi risk kategorisine giren müşteriler için ortalama gelir değeri eksik değerler yerine kullanılabilir.

4. Var olan verilere dayalı olarak en uygun değer kullanılabilir. Burada sözü edilen en uygun değerin belirlenmesi için regresyon veya karar ağacı gibi teknikler kullanılabilir. Örneğin yaşı x, eğitim düzeyi y olan bir kişi için ücret durumu, mevcut verilerden yukarıdaki tekniklerden birinin kullanılmasıyla tahmin edilebilir.

Veri temizleme tekniğinin kullanılması gereken bir diğer problem ise gürültülü verilerdir. Gürültü, ölçülen değişkendeki varyans veya rassal bir hatadır. Gürültülü verilerin teşhis edilmesi amacıyla histogram, kümeleme analizi ve reg-

resyon gibi teknikler kullanılabilir. Eğer ele alınan değişken fiyat gibi sürekli bir değişken ise, gürültülü verilerin düzeltilmesi (smoothing) gerekmektedir. Aşağıda veri düzeltme tekniklerinden bazıları açıklanmaya çalışılmıştır:

1. Binning: Binning yöntemleri, küçükten büyüğe veya büyükten küçüğe sıralanmış verileri düzeltmek için kullanılır. Binning yönteminde öncelikle sıralanmış veriler eşit büyüklükteki bin'lere ayrılır. Daha sonra bin'ler, bin ortalamaları, bin medyanları veya bin sınırları yardımıyla düzeltilir (Gohorian and Grossman, 2003:11). Örneğin değerler 4, 8, 15, 21, 21, 24, 25, 28 ve 34 olsun.

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

Örneğin bin ortalaması ile düzeltme yapılırsa,

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29 olacaktır.

2. Kümeleme (Clustering): Aykırı değerler kümeler ile belirlenebilir. Benzer değerler aynı grup veya küme içinde yer alırken, aykırı değerler kümelerin dışında yer alacaktır.

3. Regresyon (Regression): Veriler regresyon ile verilere bir fonksiyon uydurularak düzeltilebilir. Uydurulan fonksiyona uymayan noktalar aykırı değerlerdir.

Veri temizlemeyi gerektiren bir diğer durum ise verilerdeki tutarsızlıklardır. Bazı veri tutarsızlıkları dışsal referansların kullanımıyla düzeltilebilir. Örneğin kodların kullanımındaki tutarsızlıklar düzeltilebilir. Verilerin birleştirilmesinden kaynaklanan tutarsızlıklar da olabilir. Bir değişken ismi farklı veri tabanlarında farklı şekillerde olabilir.

B. VERİ BİRLEŞTİRME

Veri madenciliğinde genellikle farklı veri tabanlarındaki verilerin birleştirilmesi gerekmektedir. Bu farklı veri tabanlarındaki veriler tek bir çatı altında - ki bu genellikle bir veri ambarıdır- birleştirilirler. Farklı veri tabanlarındaki verilerin tek bir veri tabanında birleştirilmesiyle şema birleştirme hataları (schema integration errors) oluşur. Örneğin, bir veri tabanında girişler "tüketici-ID" şeklinde yapılmışken, bir diğerinde "tüketici-numarası" şeklinde olabilir. Bu tip şema birleştirme hatalarından kaçınmak için meta veriler kullanılır. Veri tabanları ve veri ambarları genellikle meta veriye sahiptirler. Meta veri, veriye ilişkin veridir.

Veri birleştirmede önemli bir konu da indirgemedir. Bir değişken, başka bir tablodan türetilmişse fazlalık olabilir. Değişkendeki tutarsızlıklar da, sonuçta elde edilen veri kümesinde fazlalıklara neden olabilir. Bu fazlalıklar korelasyon analizi ile araştırılabilir. Örneğin yukarıda da bahsedilen "tüketici-ID" ile "tüketici-numarası" korelasyon katsayısı bulunabilir. Eğer bulunan korelasyon katsayısı yüksek bulunuyorsa, değişkenlerden biri veri tabanından çıkarılarak indirgeme yapılır.

C. VERİ DÖNÜŞTÜRME

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir. Veri normalleştirme tekniklerinden bazıları aşağıdaki biçimde sıralanabilir (Roiger and Geatz, 2003:156):

1. Min-Max
2. Z Skor
3. Ondalık Ölçekleme

Min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.

Z Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir:

Ondalık ölçekleme ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:

Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir.

D. VERİ İNDİRGEME

Veri indirgeme teknikleri, daha küçük hacimli olarak ve veri kümesinin indirgenmiş bir örneğinin elde edilmesi amacıyla uygulanır. Bu sayede elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir.

Veri indirgeme yöntemleri aşağıdaki biçimde özetlenebilir:

1. Veri Birleştirme veya Veri Küpü (Data Aggregation or Data Cube)
2. Boyut indirgeme (Dimension Reduction)
3. Veri Sıkıştırma (Data Compression)
4. Kesikli hale getirme (Discretization)

Veri birleştirme veya veri küpü yapılacak 2000-2003 yılları için çeyrek dönemlik satış tutarlarından oluşan bir veri kümesinin bulunduğunu varsayalım. Bu yıllar için yıllık satış tutarları tek bir tabloda toplandığında veri birleştirmesi yapılmış olur. Sonuç olarak elde edilen veri kümesinin hacmi daha küçüktür fakat yapılacak analiz için bir bilgi kaybı söz konusu değildir. Veri küpleri ise çok değişkenli birleştirilmiş bilginin saklandığı küplerdir. Örneğin bir firmanın satış tutarları yıllar, satışı yapılan ürünler ve firmanın farklı satış yerleri için aynı küp üzerinde gösterilebilir. Veri küpleri özet bilgiye herhangi bir hesaplama yapmadan hızlı bir biçimde erişilmesini sağlarlar.

Veri madenciliği yapılacak veri kümesi bazen gereksiz olarak yüzlerce değişken içerebilir. Örneğin bir ürünün satışına ilişkin olarak düzenlenen bir veri kümesinde, tüketicilerin telefon numaraları gereksiz bir değişken olarak yer alabilir. Bu tür gereksiz değişkenler elde edilecek örüntüleri kalitesizleştirebileceği gibi veri madenciliği sürecinin yavaşlamasına da yol açacaktır. Gereksiz değişkenlerin elenmesi amacıyla ileri veya geri yönlü olarak sezgisel seçimler yapılabilir. İleri yönlü sezgisel seçimde orijinal değişkenleri en iyi temsil edecek değişkenler belirlenir. Ardından her bir değişken veya değişkenler grubunun, bu kümeye dahil edilip edilmeyeceği sezgisel olarak belirlenir. Geri yönlü sezgisel seçimde ise öncelikle değişkenlerin tüm kümesi ele alınır. Daha sonra gereksiz bulunan değişkenler kümeden dışlanarak, en iyi değişken kümesi elde edilmeye çalışılır. Boyut indirgeme amacıyla kullanılacak bir diğer yöntem ise karar ağaçlarıdır. Karar ağaçları ele alınacak çıktı değişkenini en iyi temsil edecek değişken kümesini verecektir.

Veri sıkıştırmada ise orijinal verileri temsil edebilecek indirgenmiş veya sıkıştırılmış veriler, veri şifreleme veya dönüşümü ile elde edilirler. Bu şekilde indirgenmiş veri kümesi, orijinal veri kümesini bir bilgi kaybı olacak biçimde temsil edebilecektir. Bununla beraber bilgi kaybı olmaksızın indirgenmiş veri kümesi elde edilmesine yarayacak bir takım algoritmalar da mevcuttur. Bu algoritmalar bir takım sınırlamalara sahip olduklarından sıkça kullanılamamaktadır. Bununla beraber temel bileşenler analizi gibi yöntemler, bir bilgi kaybına göz yumularak sıkıştırılmış veri kümesi elde edilmesinde kullanışlıdır.

Kesikleştirme, bazı veri madenciliği algoritmaları yalnızca kategorik değerleri ele aldığından, sürekli verilerin kesikli değerlere dönüştürülmesini içerir. Bu şekilde sürekli verilerin kesikli değer aralıklarına dönüştürülmesiyle elde edilen kategorik değerler, orijinal veri değerlerinin yerine kullanılırlar. Bir kavram hiyerarşisi (concept hierarchy), verilen sürekli değişken için, değişkenin ayrıştırılması olarak tanımlanabilir. Kavram hiyerarşileri, düşük düzeyli kavramların yüksek düzeyli kavramlarla değiştirilmesiyle verilerin indirgenmesinde kullanılır. Örneğin yaş değişkeni 1-15, 16-40, 40+ olacak biçimde daha yüksek kavram düzeyinde ifade edilebilir. Bu şekilde veri indirgemede detay bilgiler kayboluyorsa da, genelleştirilmiş veriler daha anlamlı olacak, daha kolay yorumlanabilecek ve orijinal verilerden daha düşük hacim kaplayacaktır.

Kullanılan veri madenciliği programları sayılan veri ön işleme tekniklerinden bir çoğunu gerçekleştirmektedir. Bununla beraber veri işleme ile ilgili özel programlar veya veri ön işleme açısından güçlü bir takım özel programlar vardır.

Özellikle veri ön işleme tekniklerini içeren bu açıdan güçlü programlar arasında; BioComp i-Suite, Data Digest Business Navigator 5, Data Detective, IBM Intelligent Miner for Data, KXEN, Magnify PATTERN, Quadstone DecisionHouse, Salford Systems Data Mining Suite ve Xpertrule Miner 4.0 sayılabilir.

SONUÇ

Veri madenciliğinin uygulanabilmesi için yığın halinde verilerin elimizde bulunması ön koşuldur. Veri madenciliği farklı formatlarda çok sayıda kütükte yığın halindeki veriler arasında gizli bir şekilde bulunan mesajları çekip çıkarmamıza yarayan bir araçtır. Veri madenciliği çeşitli açılardan geleneksel istatistiksel yöntemlerle önemli farklılıklar gösterir. Özellikle zaman içinde verinin azlığının değil, çokluğunun bir sorun olması ve bilgisayarların veri saklama ve işleme hızlarındaki inanılmaz artışların sonucunda veri madenciliğinin güncelliği her geçen gün artmış ve artmaktadır. Veri madenciliğinde kullanılmak üzere verilerin ön işlemeden geçirilmesi aynı anlama gelmek üzere verilerin veri madenciliğinde için hazır duruma getirilmesi veri madenciliğinin en önemli aşamalarındandır. Bu çalışmada veri ön işleme tekniklerine açıklık kazandırılmaya çalışılmıştır.

KAYNAKÇA

DMS Tutorial, "Data Preparation", İnternet Adresi; http://dms.irb.hr/tutorial/tut_data_prepare.php. Erişim Tarihi: 10.02.2003.

FAMILI, A., SHEN W, WEBER R. and E. SIMOUDIS (1997), 'Data Preprocessing and Intelligent Data Analysis', **Intelligent Data Analysis**, 1, USA, pp.3-23.

GOEBEL, M. and L. GRUENWALD (1999), "A Survey of Data Mining and Knowledge Discovery Software Tools", **SIGKDD Explorations**, 1(1), USA, pp.20-33.

GOHARIAN, N. and D. GROSSMAN (2003), "Data Preprocessing", İnternet Adresi;<http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-reprocessing.pdf> .Erişim Tarihi:08.05.2003.

GÜRSAKAL, N. (2001), **Sosyal Bilimlerde Araştırma Yöntemleri**, VİPAŞ, Bursa, 189s.

HAN, J. and M. KAMBER (2001), **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, USA, 550p.

HAND, D., MANNILA H. and P. SMYTH (2001), **Principles of Data Mining**, MIT, USA, 546p.

KIM, W., CHOI B., HONG E., KIM S. and D. LEE (2003), "A Taxonomy of Dirty Data", **Data Mining and Knowledge Discovery**, 7, pp.81-99.

PYLE, D. (1999), **Data Preparation For Data Mining**, Morgan Kaufmann Publishers, USA, 540p.

PIRAMUTHU, S. (2003), "Evaluating Feature Selection Methods for Learning in Data Mining Applications" **European Journal of Operational Research**, Article In Press, pp.1-11.

ROIGER, R. J. and M. W. GEATZ (2003), **Data Mining A Tutorial-Based Primer**, Addison Wesley, USA, 350p.

ZHOU, Z. (2002), "Three Perspectives of Data Mining" **Artificial Intelligence**, 143 (2003), pp.139-146.