Veri madenciliği

Murat Bütün

1980 yılında Ordu'da doğdu. 2002'de Marmara Üniversitesi Sağlık Eğitim Fakültesi Sağlık Eğitim Bölümü'nden mezun oldu. 2006'da Marmara Üniversitesi Sağlık Bilimleri Enstitüsü Sağlık Kurumları Yöneticiliği Bölümü'nde yüksek lisans eğitimini bitirdi. Üniversite yıllarının başlarında Sağlık Enformasyon Sistemleri ile ilgilenmeye başladı, ayrıca sağlık alanında çeşitli web sitelerinin ve portallarının kurulumunda ve işletilmesinde aktif görev aldı. Halen www.saglikplatformu.com web sitesinin Genel Koordinatörlüğü'nü yürütmekte ve İstanbul Sağlık A.Ş. Bilgi İşlem Biriminde yöneticisi olarak çalışmaktadır.

Günümüzde bilgisayar sistemleri her geçen gün ucuzluyor ve aynı zamanda güçleri de artıyor. Bilgisayar sistemlerindeki bu gelişmeyle birlikte kullanımı da bu ölçüde yaygınlaşmaktadır. Bu gelişmeyle birlikte işletmelerde üretilen sayısal bilgi miktarının arttığını, buna paralel veri tabanlarının daha fazla veriyi saklayabilecek boyutlara ulaştığını ve bilgisayar sistemlerindeki gelişme ile veriye ulaşmanın kolaylaştığını görmekteyiz. Bu sayede doğru ve daha detaylı bilgiye ulaşmamız mümkün hale gelmiş fakat başka bir sorun ortaya çıkmıştır. Bu sorun; oluşan bu büyük sayısal veri yığınlarının yönetilmesi ve anlamlı hale getirilmesi sorunudur.

Şirketlerin bilgi sitemleri üzerinden ürettiği bilgi miktarının büyük artış gösterdiğini ve bazı firmaların veri tabanlarının boyutlarının 1 milyon gigabyte'a (GB) ulaştığını görmekteyiz. İşte veri tabanlarındaki teknolojik gelişme ve hacimlerindeki bu olağanüstü artış, firmaları elde toplanan bu verilerden nasıl faydalanacağını ve bu verilerin nasıl anlamlı hale getirileceği sorunuyla karşı karşıya bırakmıştır.

Bilgisayar sistemleri ile üretilen bu veriler tek başlarına değersizdirler. (Özellikle veri tabanlarının bilgiyi sadece saklamak için dizayn edildiği düşünüldüğünde...) Çünkü çıplak gözle bakıldığında verilerin bir anlam ifade etmediğini söyleyebiliriz. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. İşte ham veriyi bilgiye veya anlamlı hale dönüştürme işini veri madenciliği ile yapabiliriz.

Veri tabanlarındaki bu veriler üzerinde analiz yapmak ve karar destek aşamasında faydalanmak, herhangi bir araç kullanmaksızın imkânsız hale gelmiştir. Çoğu zaman iyi kullanılmamaları durumunda veri tabanlarında tutulan veri, insanlar için külfet haline de gelebilmektedir. Bu noktada karşımıza "Veri Ma-

denciliği" (data mining) bir çözüm olarak çıkmaktadır. Fakat madenciliği yapılacak olan verinin de bazı vasıflara sahip olması gerekmektedir. Bu vasıflar veri ambarı (data warehouse) ile sağlanmaktadır. Veri ambarları basit olarak veri madenciliği işleminin yapılacağı verilerin oluşturulduğu özel veri tabanlarıdır. Veri ambarlarının oluşturulması işlemi, verinin çeşitli kaynaklardan toplanarak, veriler içerisindeki uyuşmazlıklar ve hatalardan arındırılmasından ibarettir.

Veri madenciliği, belirli bir alanda ve belirli bir amaç için toplanan veriler arasındaki gizli kalmış ilişkilerin ortaya konulmasıdır. Bunun yanında, geleceğe dönük kararlar almamızda bize fikir de verir. Veri madenciliği, disiplinler arası doğasından dolayı istatistik, veri tabanları, makine öğrenmesi, bilgi toplama, görselleştirme, paralel ve dağıtık hesaplama gibi birçok disiplinden yardım alır. Aynı zamanda veri madenciliği birçok farklı alanda da kullanılmaktadır.

Bir süpermarket örneğinde veri analizi yaparak her mal için bir sonraki ayın satış tahminleri çıkarılabilir; müşteriler satın aldıkları mallara bağlı olarak gruplanabilir; yeni bir ürün için potansiyel müşteriler belirlenebilir; müşterilerin zaman içindeki hareketleri incelenerek davranışları ile ilgili tahminler yapılabilir. Binlerce malın ve müşterinin olabileceği düşünülürse bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkar. Veri madenciliği burada devreye girer:

Veri madenciliği büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programlarıyla aranmasıdır.

Geleceğin, en azından yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapmamızı sağlayacaktır.

Veri madenciliği

Veri madenciliği nedir? Öncelikle bu soruyu cevaplamaya çalışalım. Veri madenciliği, büyük veri yığınları arasında gizli kalmış anlamlı kuralların zekî olarak ortaya çıkarılmasıdır. Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı birçok teknik kullanılmaktadır.

Veri madenciliği ile büyük veri yığınlarından oluşan database sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, database teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır.

Veri madenciliği kendi başına bir çözüm değil, çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analistine, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir.

Günümüzde veri madenciliğinin başlıca ilgi alanları olarak aşağıdakiler sayılabilir;

Pazarlama

• Müşteri segmentasyonunda,

• Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,

• Çeşitli pazarlama kampanyalarında,

• Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,

• Pazar sepeti analizinde,

• Çapraz satış analizlerinde,

• Müşteri değerlemede,

• Müşteri ilişkileri yönetiminde,

• Çeşitli müşteri analizlerinde,

• Satış tahminlerinde.

Bankacılık

• Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,

• Kredi kartı dolandırıcılıklarının tespitinde,

• Müşteri segmentasyonunda,

• Kredi taleplerinin değerlendirilmesinde,

• Usulsüzlük tespitinde,

• Risk analizlerinde,

• Risk yönetiminde.

Sigortacılık

• Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,

• Sigorta dolandırıcılıklarının tespitinde,

• Riskli müşteri tipinin belirlenmesinde.

Perakendecilik

• Satış noktası veri analizlerinde,

Bilgisayar sistemleri ile üretilen veriler tek başlarına değersizdirler. Çünkü çıplak gözle bakıldığında veriler bir anlam ifade etmez. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. Ham veriyi bilgiye veya anlamlı hale dönüştürme işini veri madenciliği ile yapabiliriz.

• Alışveriş sepeti analizlerinde,

• Tedarik ve mağaza yerleşim optimizasyonunda.

Borsa

• Hisse senedi fiyat tahmininde,

• Genel piyasa analizlerinde,

• Borsada alım-satım stratejilerinin optimizasyonunda.

Telekomünikasyon

• Kalite ve iyileştirme analizlerinde,

• Hisse tespitlerinde,

• Hatların yoğunluk tahminlerinde.

Sağlık ve ilaç

• Test sonuçlarının tahmininde,

• Ürün geliştirmede,

• Tıbbi teşhiste,

• Tedavi sürecinin belirlenmesinde.

Endüstri

• Kalite kontrol analizlerinde,

• Lojistikte,

• Üretim süreçlerinin optimizisyonunda.

Veri madenciliği; disiplinler arası doğasından dolayı istatistik, veri tabanları, makine öğrenmesi, bilgi toplama, görselleştirme, paralel ve dağıtık hesaplama gibi birçok disiplinden yardım alır. Aynı zamanda veri madenciliği birçok farklı alanda da kullanılmaktadır.

Bilim ve mühendislik

• Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi.

Veri madenciliği, büyük veri yığınları arasında gizli kalmış anlamlı kuralların zekî olarak ortaya çıkarılmasıdır. Bu çıkarım işlemi tek başına basit bir işlem değildir. Çok yoğun işleyen alt işlemlerden oluşmaktadır. Bu alt işlemler şunlardır:

Uygulama alanının ortaya konulması

Bu adım veri madenciliğinin hangi yönde ve hangi amaçlar için yapılacağının belirlendiği adımdır. Bu aşamada belirlenen alan ile ilgili uzmanlar ile görüşmeler yapılarak bundan sonraki aşamalar için stratejilerin ortaya konulması gerekir. Örneğin kullanılacak olan model için bu aşamadan faydalanılmaktadır.

Veri ambarının oluşturulması

Veri ambarı aşaması, veri madenciliği sürecinde önemli bir aşamadır. Bu süreç toplam maliyet ve zamanın önemli bir kısmını almaktadır. Madenciliğini yapacağımız veri tek bir yapı içerisinde bulunmayabilir. Bu nedenle bilginin tek bir çatı altında toplanması gerekir. Fakat veri ambarı oluşturma aşamasında, amaç sadece verinin tek bir çatı altında toplanması değildir. Aynı zamanda toplanan veriler içerisinde var olan hataların ve belirsizliklerinde temizlenmesi aşamasıdır. Bu aşamada veri bazı alt işlemlere tabi tutulmaktadır. Bu işlemler Veri Toplama, Uyumlandırma, Birleştirme ve Temizlenme, Seçme ve Dönüştürmedir.

• **Toplama:** Bilginin keşfi için gerekli veriler faklı kaynaklarda olabilir. Toplama işlemi; verinin farklı kaynaklardan alınarak bir kaynakta birleştirilmesidir. Hastanın tıbbi verileri yanında, yaşadığı yer veya gelir düzeyi gibi bilgilere de ihtiyaç duyulabilir. Bu bilgilerin farklı kaynaklarda bulunması durumunda toplama işlemi gerekmektedir.

• **Uyumlandırma:** Veri ambarındaki verilerin farklı kaynaklardan toplanabileceğini söylemiştik. Fakat bu durumda karşımıza uyumsuz veri tipleri çıkacaktır. Bunun en yaygın örneği cinsiyette görülmektedir. Çok fazla tipte tutulabilen bir veri olup bir veri tabanında 0/1 olarak tutulurken diğer veri tabanında E/K veya Erkek/Kadın şeklinde tutulabilir. Bilginin keşfinde başarı, verinin uyumuna da bağlı olmaktadır. Özellikle tıp alanındaki verilerin çeşitliliği oldukça fazladır. Bu nedenle uyumlandırma aşaması, tıbbi verilerde daha da önem kazanmaktadır.

• **Birleştirme ve temizleme:** Yukarıda bahsedilen uyumlandırma işlemi sırasında, farklı veri kaynaklarından gelen verilerin birleştirilmesi veya fazlalıkların temizlenmesi de gerekmektedir.

• **Seçme:** Bu adım bizim ilerde kuracağımız model için uygun verinin seçilmesi işlemidir. Bir sınıflandırma işleminde öznitelikleri içine alan bir verinin seçilmesi anlamını taşımaktadır. Veri tabanlarındaki işlem hızları artmasına rağmen büyük veritabanları üzerinde birden fazla modelin denenmesi oldukça zaman ve maliyet gerektirmektedir. Bunun yerine verinin bütününü temsil edecek şekilde bir parça üzerinde işlemler yapılabilir. Fakat seçilecek parçanın verinin tamamını temsil etmesi açısından önemi büyüktür.

• **Dönüştürme:** Verinin kullanılacak modele göre içeriğini koruyarak şeklinin dönüştürülmesi işlemidir. Dönüştürme işlemi kullanılacak modele uygun biçimde yapılmalıdır. Çünkü verinin gösterilmesinde kullanılacak model ve algoritma önemli bir rol oynamaktadır.

Yukarıdaki bilgiler ışığında veri tabanları ile veri ambarları arasındaki bazı farklar göze çarpmaktadır. İçerik olarak veri tabanları bütün detayları kapsamakta, veri ambarlarında ise daha çok özet ve ilgili bilgileri tutulmaktadır. Veri tabanı işlemleri, bir kısım veri üzerinde yapılırken veri ambarları, daha fazla veri üzerinde işlem yapmaktadır. Veri tabanlarında veriler iki boyutlu tutulurken veri ambarları çok boyutlu veri saklama imkânı sunmaktadır. Bu sebeple verinin analizi kolaylaşmaktadır. Veri tabanları sürekli güncellenirken veri ambarları belirli periyotlar ile güncellenirler.

Modelin kurulması ve değerlendirilmesi

Bilginin keşfi sürecinde hazırlanan verilerin, ortaya konulan probleme uygun modelin ortaya konulması ve bu modele ait algoritmanın seçilmesidir. Genel olarak Sınıflama ve Tahmin (Classification and Prediction), Küme Analizi (Cluster Analysis), Birliktelik Kuralları (Associaton Rules) şeklinde modelleri sınıflandırabiliriz.

• **Sınıflama ve tahmin:** Bu iki modeli aslında birbirlerinden farklı amaçları olsa da, aynı tekniklerle kullanıldığı için tek başlık içerisinde alabiliriz. Bu iki model arasındaki bağlantı, tahmin edilen değerlerin sınıflanmış bir yapıya sahip olmasıdır. Sınıflanma modeli iki adımdan oluşmaktadır. İlk adımda gözlenmiş veriler sınıflandırma algoritması kullanılarak sınıflandırma kuralları oluşturulur. İkinci adımda ise oluşturulan sınıflandırma kuralları kullanılarak veriler sınıflandırılır. Tahmin modelinde sürekli veriler alınarak oluşturulan kurallara göre sonuçlandırılır. Bu modellerde kullanılan algoritmalardan bazıları Karar Ağaçları (Decision Tree), Hatayı Geri Yayma (Backpropagation), Bayes Sınıflandırması (Bayesian Classification)'dır.

• **Küme analizi:** Kümeleme işlemi, birbirine benzeyen nesnelerin aynı grupta toplanmasıdır. Bu modelde en büyük etken hangi kriterlere göre kümeleme yapılacağıdır. Bu işlem, konu ile ilgili uzman tarafından tahmin edilir. Veriler kümeleme işleminde aynı sınıfta yer almalarına rağmen farklı gruplarda da yer alabilir. Nüfus bilimi ve astronomi alanında kullanımları yaygındır.

• **Birliktelik kuralları:** Bu model veri nesneleri arasındaki ilginç ilişkileri araştırır; gerek birbirini izleyen gerekse eş zamanlı durumlarda araştırma yapar. Bu model yaygın olarak Market Sepet Analizi uygulamalarında kullanılmaktadır. Bunun yanında finans ve tıp alanında da kullanılmaktadır.

Bilgi keşfi için modelin kurulması çok zahmetli bir işlemdir. Çünkü hangi model ve algoritmanın bize daha iyi performans vereceğini önceden kestirmemiz imkansız olmasa da çok zordur. Bu nedenle mevcut olan bütün modelleri kurarak bunlar arasında mukayese edilmelidir. Modelin öğrenmesi Deneticili (Supervised) ve Deneticisiz (Unsupervised) olmak üzere ikiye ayrılır. Deneti-

cili yöntemde verinin bir kısmı seçilen algoritmanın eğitimi için diğer kısmı da eğitimin testi için kullanılmaktadır. Test işlemindeki başarı, o modelin kalitesini ortaya koymaktadır.

Öğreticisiz yöntemde ise ilgili özellikler arasındaki benzerlikten ortaya çıkarak eğitim yapılmaktadır. Veri tabanı işlemlerindeki hızlanmaya rağmen verilerin çok fazla olması eğitim ve test süresini olumsuz yönde etkilemektedir. Bu nedenle verinin tamamı yerine onu en iyi şekilde temsil edecek daha küçük bir veri topluluğu üzerinde modelin seçilmesi ve sonra tüm veriye uygulanması iyi bir çözüm olmaktadır.

Şablonların ve ilişkilerin yorumlanması

Yapılan çalışmalar sonucunda elde edilen ilişkilerin ve kuralların uzman tarafından incelenerek yorumlanması aşamasıdır. Bu aşamada modelin bize sunduğu ilişkiler incelenmektedir. Biz model üzerinde verinin bir kısmını kullandığımız için karşımıza gelen bütün ilişkiler anlamlı olmayabilir. Bu nedenle uzmanların bu aşamada yaptığı inceleme ve yorumlar ışığında model üzerinde değişiklikler yapılarak işlemlere faydalı yeni boyutlar kazandırılabilir.

Örnek Uygulamalar

Bağıntı: "Çocuk bezi alan müşterilerin % 30'u bira da satın alır."

Sepet analizinde (basket analysis) müşterilerin beraber satın aldığı malların analizi yapılır. Buradaki amaç mallar arasındaki pozitif veya negatif korelasyonları bulmaktır. Çocuk bezi alan müşterilerin mama da satın alacağını veya bira satın alanların cips de alacağını tahmin edebiliriz. Ancak otomatik bir analiz, bütün olasılıkları göz önüne alır ve kolay düşünülemeyecek; örneğin çocuk bezi ve bira arasındaki bağıntıları da bulur.

Sınıflandırma: "Genç kadınlar küçük araba satın alır; yaşlı, zengin erkekler büyük, lüks araba satın alır."

Amaç bir malın özellikleri ile müşteri özelliklerini eşlemektir. Böylece bir müşteri için ideal ürün veya bir ürün için ideal müşteri profili çıkarılabilir. Örneğin bir otomobil satıcısı, şirketin geçmiş müşteri hareketlerinin analizi ile yukarıdaki gibi iki kural bulursa; genç kadınların okuduğu bir dergiye reklâm verirken küçük modelinin reklâmını verir.

Regresyon: "Ev sahibi olan, evli, aynı iş yerinde beş yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 825'dir."

Başvuru skorlamada (application scoring) bir finans kurumuna kredi için başvuran kişi ile ilgili finansal güvenilirliğini notlayan örneğin 0 ile 1.000 arasında bir skor hesaplanır. Bu skor kişinin özellikleri ve geçmiş kredi hareketlerine dayanılarak hesaplanır.

Zaman içinde sıralı örüntüler: "İlk üç taksitinden iki veya daha fazlasını geç ödemiş olan müşteriler, % 60 olasılıkla kanuni takibe gidiyor."

Davranış skoru (behavioral score), başvuru skorundan farklı olarak kredi almış ve taksitleri ödeyen bir kişinin sonraki taksitlerini ödeme/geciktirme davranışını notlamayı amaçlar.

Benzer zaman sıraları: "X şirketinin hisse fiyatları ile Y şirketinin hisse fiyatları benzer hareket ediyor."

Amaç zaman içindeki iki hareket serisi arasında bağıntı kurmaktır. Bunlar örneğin iki malın zaman içindeki satış miktarları olabilir. Örneğin dondurma satışları ile kola satışları arasında pozitif, dondurma satışları ile sahlep satışları arasında negatif bir bağıntı beklenebilir.

İstisnalar (Fark saptanması): "Normalden farklı davranış gösteren müşterilerim var mı?"

Amaç önceki uygulamaların aksine kural bulmak değil, kurala uymayan istisnai hareketleri bulmaktır. Bu da örneğin olası sahtekârlıkların saptanmasını (fraud detection) sağlar.

Örneğin 'Visa' kredi kartı için yapılan CRIS sisteminde bir yapay sinir ağı, kredi kartı hareketlerini takip ederek müşterinin normal davranışına uymayan hareketler için müşterinin bankası ile temasa geçerek müşteri onayı istenmesini sağlar.

Doküman Madenciliği: "Arşivimde (veya internet üzerinde) bu dokümana benzer hangi dokümanlar var?"

Amaç dokümanlar arasında ayrıca elle bir tasnif gerekmeden benzerlik hesaplayabilmektir (text mining). Bu genelde otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı sayesinde yapılır.

Sağlık alanında veri madenciliği

Sağlık alanında bilginin kullanım şeklinde meydana gelen değişiklikler sağlık bakım hizmetini verenleri etkilemiştir. Sağlık bakım hizmetinin verilmesinde bilgisayar kullanımı, bilginin paylaşımekip yaklaşımını, veri ve bilgi temelli uygulama gibi kavramlar yaygınlaşmaya başlamıştır. Bilgisayarlar, hasta bakım hizmetlerinin destekleme, sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanılmasının yanı sıra, karar verme, yönetim, planlama ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanılmıştır.

Sağlık alanında bulunan mevcut veri oldukça hayati öneme sahiptir. Hastane bilgi sistemleri sayesinde bu veriler düzenli olarak tutulmaktadır. Hayati öneme sahip olan bu verilerden daha fazla yaralanmak mümkündür. Hastane bilgi sistemlerinden veya diğer tıbbi veri toplayan sistemlerden alınan veriler üzerinde yapılan veri madenciliği çalışmaları, hem uzmanlar için hem hastane yönetimi için hem de hastaların daha kaliteli bir hizmet almalarında etkin rol alabilir.

Bilgisayarlar, hasta bakım hizmetlerinin desteklenmesi, sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanılmasının yanı sıra, karar verme, yönetim, planlama ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanılmıştır.