Veri Madenciliği Yöntemlerine Genel Bakış

Overview to Data Mining Methods: Review

Füruzan KÖKTÜRK,[a]
Handan ANKARALI,[a]
Vildan SÜMBÜLOĞLU[a]

[a]Biyoistatistik AD,
Zonguldak Karaelmas Üniversitesi
Tıp Fakültesi, Zonguldak

Geliş Tarihi/*Received:* 02.07.2008
Kabul Tarihi/*Accepted:* 25.08.2008

Yazışma Adresi/*Correspondence:*
Handan ANKARALI
Zonguldak Karaelmas Üniversitesi
Tıp Fakültesi, Biyoistatistik AD,
Zonguldak,
TÜRKİYE/TURKEY
hankarali@yahoo.com

ÖZET Bilgi keşfi ve veri madenciliği, bir disiplinlerarası alandır ve veriden kullanışlı olan bilgiyi çıkarmak için gerekli olan metodlar üzerine odaklanmıştır. Son zamanlara kadar, veri madenciliği çalışmalarının çoğu, az sayıda bilgisayar bilimcisi, programcısı, veri tabanı yöneticisi ve yönetim bilgi uzmanlarının ilgi odağı idi. Fakat şimdi, modern deneysel ve gözleme dayalı metodlar yardımıyla elde edilen büyük veri setlerinden kullanışlı olan bilgiyi çıkarmak için iş dünyası, bankacılık ve tıbbi alan gibi çok çeşitli alanlarda hızla artan bir kullanım alanı bulmaktadır. Özellikle tıp alanındaki verinin büyüklüğü ve hayati önem taşıması bu alandaki uygulamaları daha da önemli kılmaktadır. Tıpta veri madenciliği, tıbbi verilerin heterojen yapıda olması, özel etik ve hukuki kurallar gerektirmesi ve hasta sırlarını temel alan deodeontolojik kurallar içermesi, istatistik metodların bu heterojenite ve sosyal konuları adres etmek zorunda olması ve tıbbın insan hayatında özel bir yerinin olması gibi nedenlerle diğer alanlardan farklılık gösterir. Veri madenciliğinde birinci ve en basit analitik adım, tanımlayıcı istatistikleri kullanarak, grafik ve şekillerle görsel inceleme yaparak ve değişkenler arası potansiyel anlamlı bağlantılara bakarak veriyi tanımlamaktır. Bilgi keşfi ve veri madenciliği ile istatistik disiplini sınıflama ve yapı tanımlama problemleri üzerine çalışmaktadır. Bilgisayarların gücündeki artış ve fiyatlarının düşmesi, mümkün olan çözümlerin incelenmesi yoluyla ortaya çıkan yeni tekniklerin gelişmesine imkân sağlamıştır. Yeni teknikler içinde yapay sinir ağlarının algoritmalarına benzer yeni algoritmalar, karar ağaçları ve diskriminant analizi gibi eski algoritmalara yeni yaklaşımların getirildiği teknikler yer alır. Bu yöntemlerin çoğunluğu tıpta tanı koyma ve sınıflama amaçlarıyla kullanılmaktadır.

Anahtar Kelimeler: Sınıflama; lojistik modeller; karar ağaçları; yapay sinir ağları; tıbbi veri madenciliği

ABSTRACT Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. Up until recently, much of the work of data mining has been the domain of a small number of computer scientists, programmers, data base administrators, and management information specialists. But now, it is increasingly being used in the all area such as business, financial, medical sciences to extract information from the large data set generated by modern experimental and observational methods. Data mining in medicine is distinct from that in other fields, because the data are heterogeneous; special ethical, legal, and social constraints apply to private medical information; statistical methods must address these heterogeneity and social issues; and because medicine itself has a special status in life. The first and simplest analytical step in data mining is to describe the data-summarize its descriptive statistics, visually review it using charts and graphs, and look for potentially meaningful links among variables. KDD and statistics disciplines have been working on problems of pattern recognition and classification. The increased power of computers and their lower cost have allowed the development of new techniques based on exploration of possible solutions. New techniques include relatively recent algorithms like neural nets and decision trees, and new approaches to older algorithms such as discriminant analysis. Much of these data mining methods are used for differential diagnosing and classification purposes in medical studies.

Key Words: Classification; logistic models; decision trees; artificial neural networks; medical data mining

Turkiye Klinikleri J Biostat 2009;1(1):20-5

Bilgisayar teknolojisindeki gelişmelerle birlikte üretilen sayısal bilgi miktarının arttığı, veri tabanlarının daha fazla veriyi saklaya- bilecek boyutlara ulaştığı ve veriye ulaşmanın kolaylaştığı görülmektedir. Öyle ki, dünyadaki bilgi miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir.[1] Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki bu olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, Veri Tabanlarında Bilgi Keşfi (VTBK) adı altında, sürekli ve yeni arayışlara neden olmaktadır. Büyük miktarlardaki verinin veri tabanlarında tutuldukları bilindiğine göre bu verilerin veri madenciliği teknikleriyle işlenmesine "Veri Tabanında Bilgi Keşfi" denir.[2] VTBK ve veri madenciliği terimleri farklı süreçleri işaret etseler de çoğunlukla birbirlerinin yerine kullanılırlar. VTBK, veriden faydalı bilgi keşfinin tüm aşamalarını ifade ederken veri madenciliği bu süreçte belirli bir adımdır. Tarihsel olarak, veri içerisindeki faydalı yapıların ortaya çıkarılması olayına pek çok isim karşılık gelmiştir. Bunlardan bazıları veri madenciliği, bilgi çıkarımı, bilgi keşfi, bilgi harmanlama, veri arkeolojisi ve veri modelleme sürecidir.[3]

VTBK süreci, birbirini etkileyen ve yinelemeli bir süreç olup,

- Veri seçimi
- Veri ön işleme
- Veri dönüştürme
- Veri madenciliği
- Veri yorumlama/değerlendirme aşamalarından oluşmaktadır.[3]

VERİ MADENCİLİĞİ NEDİR?

Veri madenciliği, veri ambarlarında yararlı olma potansiyeline sahip, aralarında beklenmedik/bilinmedik ilişkilerin olduğu verilerin keşfedilerek, hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur.[3] Veri madenciliğinin ilk uygulamaları pazarlama alanında olmuştur. Daha sonraki yıllarda karar alma ve bilgi yönetimi süreçlerinde yoğun bir şekilde kullanılmaya başlanmıştır.[3,4]

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. O dönemlerde bilgisayar yardımıyla yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenildi. Bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verildi. 1990'lı yıllara gelindiğinde veri madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atıldı. Bu camianın amacı, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktı. Bu noktadan sonra bilim adamları veri madenciliğine çeşitli yaklaşımlar getirmeye başladılar.[5]

Veri madenciliği büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri tabanlarının birleştirilmiş ve işlemeye daha uygun bir özetini saklamayı amaçlar. Günlük veri tabanlarından istenen özet bilgi seçilerek ve gerekli ön işlemeden sonra veri ambarında saklanır. Ardından amaç doğrultusunda gerekli veri ambardan alınarak veri madenciliği çalışması için standart bir forma çevrilir. Veri ambarlarının analizi için "Online Analytical Processing (OLAP)" programları kullanılır. OLAP, veriye çok boyutlu bakmayı ve incelemeyi sağlar.[3]

Veri madenciliği, veri tabanı teknolojisi, istatistik, makine öğrenimi, model tanımı, yapay sinir ağları, veri görselleştirme ve uzaysal veri analizi gibi farklı disiplinlerde yer alan tekniklerin bir birleşimini içerir. Bu disiplinlerin arasındaki kesin sınırları tanımlamak zor olduğu gibi, bu alanlar ile veri madenciliği arasındaki kesin sınırları tanımlamak da zordur.[6]

VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Veri madenciliğinin günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda uygulaması

mümkün olmakla birlikte belli başlı kullanım alanları olarak;

Pazarlama: Pazar segmentasyonu, müşteri değerleme ve çapraz satış analizleri,

Bankacılık: Risk analizi, usulsüzlük tespiti, müşteri kazanma ve mevcut müşterileri elde tutma analizleri, çapraz satış,

Sigortacılık: Müşteri kaybı sebeplerinin belirlenmesi, usulsüzlüklerin önlenmesi, ana giderlerin azaltılması, poliçe fiyatlarının belirlenmesi,

Perakendecilik: Satış noktası veri analizleri, alış veriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonları,

Borsa: Hisse senedi fiyat tahmini, genel piyasa analizleri, alım-satım stratejilerinin optimizasyonu,

Telekomünikasyon: Kalite iyileştirme, hile tespiti, hataların yoğunluk tahmini, müşteri kazanma ve elde tutma analizleri,

İlaç: Test sonuçlarının tahmini, ürün geliştirme,

Sağlık: Tıbbi teşhis ve tanı, uygun tedavi sürecinin belirlenmesi, sınıflandırma,

Endüstri: Kalite kontrol, lojistik, üretim süreçlerinin optimizasyonu,

Bilim ve Mühendislik: Ampirik veriler üzerinde modeller kurularak bilimsel ve teknik problemlerin çözümlenmesi sayılabilir.[7]

VERİ MADENCİLİĞİNİN İŞLEYİŞİ

Veri madenciliğinde veri, öğrenme ve deneme olmak üzere ikiye ayrılır. Her uygulamada kullanılabilecek birden çok teknik vardır ve önceden hangisinin en başarılı olacağını kestirmek olası değildir. Bu yüzden öğrenme kümesi üzerinde değişik teknik kullanılarak *L* tane model oluşturulur. Sonra bu *L* model deneme kümesi üzerinde denenerek en başarılı olanı, yani deneme kümesi üzerindeki tahmin başarısı en yüksek olan seçilir. Eğer bu en iyi model yeterince başarılıysa kullanılır, aksi takdirde başa dönerek çalışma tekrarlanır. Tekrar sırasında başarısız olan örnekler incelenerek bunlar üzerindeki başarının nasıl artırılabileceği araştırılır.[8]

VERİ MADENCİLİĞİ METOTLARI

Literatürde veri madenciliği metodları için pek çok sınıflama türü mevcuttur. Sınıflandırma biçimlerinden biri yöntemlerin parametrik olup olmamasına göredir. Parametrik modellerde çalışılan özelliklerle ilgili çeşitli varsayımların sağlanması koşulu aranır. Bu sebeple gerçek yaşamdaki kullanımı çok kısıtlı kalmaktadır. Parametrik olmayan modeller ise veri madenciliği için daha uygundur, çünkü bu modellerde veriye göre model oluşturulur. Parametrik olmayan teknikler arasında,

- Sinir ağları (neural networks)
- Karar ağaçları (decision trees)
- Genetik algoritmalar (genetic algorithms) sayılabilir.

Veri madenciliğinde kullanılan modeller bir başka sınıflandırma kuralına göre iki ana başlık altında incelenmektedir:

- Tahmin (Predictive) Modelleri: Bu modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Eğer tahmin edilecek değişken sayısal bir değişkense tahmin problemleri regresyon problemi, kategorikse sınıflama problemi adını alır.[2,9]

- Tanımlayıcı (Descriptive) Modeller: Bu tip modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki modellerin tanımlanması sağlanmaktadır.

- Veri madenciliği modellerini gördükleri işlevlere göre ise;

- Sınıflama *(classification*) ve regresyon *(regression)*

- Kümeleme *(Clustering),*

- Birliktelik kuralları *(association rules)* ve ardışık zamanlı modeller *(sequential patterns),* olarak sınıflandırmak mümkündür.[2]

Yapılan son sınıflama şekline göre ortaya çıkan yöntemlerin genel amaçları aşağıdaki gibi özetlenebilir.

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin

eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır:[2,10]

- 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),
- Naïve-Bayes
- Lojistik regresyon

Tıbbi araştırmalarda en sık kullanılan yöntemler bu gruba dahildir. Söz konusu yöntemler hasta tanısı koymada, hastalıkların sınıflandırılmasında, hasta özelliklerinin tahmin edilmesinde yaygın bir şekilde kullanılmaya başlanmıştır.

Kümeleme, veriyi sınıflara veya kümelere ayırma işlemidir. Kümeleme modelinde, sınıflama modelinde olan veri sınıfları yoktur. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir. Oysa kümeleme modelinde, sınıfları bulunmayan veriler gruplar halinde kümelere ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir ön işlemi gibi görev alabilmektedir.[11] Kümeleme analizi tıbbi araştırmalarda bilinmeyen hastalık kümelerinin tanımlanması, gen araştırmaları, biyolojik varyasyonun gruplandırılması, çeşitli hastalıklara özgü genetik yapı kümelerinin oluşturulması gibi amaçlarla kullanılabilir.

Birliktelik kuralları, büyük veri kümeleri arasındaki birliktelik ilişkilerini bulurlar. Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelik- leri bularak müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ışığında daha etkili satış stratejileri geliştirebilirler.[10] Örneğin bir müşterinin yoğurdun yanında ekmek alma olasılığı nedir? Bir market yöneticisi bu tip bir bilgi ışığında rafları yeniden düzenleyebilir. Birliktelik kurallarında eş zamanlı olarak gerçekleşen ilişkiler tanımlanmaktadır. Bu tip ilişkiler tıbbi araştırmalarda oldukça sık karşımıza çıkmaktadır. Mesela, HDL düşük ise kilo artışı gözlenmesi, tansiyon düşüklüğünde baş dönmesi görülmesi gibi çok sayıda örnek verilebilir.

Ardışık zamanlı modeller ise birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır.[2] Sigara içen bir kişide akciğer kanseri görülmesi, aşırı yağlı yiyeceklerle beslenen kişinin kolesterol düzeyinin yüksek çıkması veya doğum sonrası postpartum depresyon görülmesi gibi olaylar ardışık zamanlı modellere örnek olarak gösterilebilir.

TIPTA VERİ VE VERİ MADENCİLİĞİ

Sağlık sektöründeki gelişmeler ve insanların ortalama yaşam sürelerinin uzaması beraberinde bazı sorunları da getirmiştir. Örneğin birçok insan, kalp hastalıkları, diyabet ve astım hastalıkları gibi kronik hastalıklarla yaşamak zorundadır. Bu hastalıkların hem tıbbi açıdan hem de hastane kaynak ve maliyetleri açısından ele alınarak doğru yönetilmesi gerekmektedir. Bu noktada bilgi sistemleri üzerinde çalıştırılabilecek klasik sorgulama yöntemleri yetersiz kalmaktadır. Bu noktada yardıma veri madenciliği teknikleri yetişmektedir.

Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'li ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır.

Tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Sağlık alanında yapılan birçok veri madenciliği araştırmasında hastaların elektronik tıbbi kayıtları ve idari işleri belgeleyen veriler kullanılmaktadır. Bu verilerden yararlanılarak farklı tahminler yapılabilir. Bunlardan bazıları şunlardır:

- Belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesi,
- Tıbbi tedaviden sonra hastaların durumlarının tahmin edilmesi,
- Hastane maliyetlerinin tahmin edilmesi,
- Ölüm oranları ve salgın hastalıkların tahmin edilmesidir.

Tıbbi veri tabanlarında veri madenciliği ve bilginin bulunması da diğer türdeki veri tabanlarındakinden çok farklı değildir. Ancak, tıbbi veride diğer veri türlerinde olmayan bazı özellikler vardır. Tıp alanında belirli bir standardın olmayışı ve var olan standartlar arasında tam bir uyumun olmaması nedeniyle, bu alanda bir veri ambarının oluşturulması oldukça zor bir işlemdir. Bunun yanı sıra tıp alanındaki terimlerin hem karışık hem de birbirine yaklaşık olması da veri ambarı oluşumunu negatif yönde etkilemektedir. Tıbbi verinin heterojen yapıda olması, etik, yasal ve sosyal konular, hasta gizliliği-güvenliği gibi konular da bu tür verinin kısıtlılıkları arasındadır.[12]

Hastalıkların yönetimi ile ilgili veri madenciliği çalışmaları, hastalıkların ve durumlarının tanımlanması ve maliyetlerin modellenmesi gibi araştırmaları içerir. Örneğin, bir veri madenciliği algoritması kullanılarak, hastaların yaş, cinsiyet ve semptomlarının yoğunluğuna ait bilgiler yardımıyla hastalığın olup olmamasına dair kurallar çıkarabilir.[13]

Kore Tıbbi Sigorta Kurumu tarafından hazırlanan bir veri tabanı üzerinde yüksek tansiyon ile ilgili bir çalışma yapılmıştır. Bu çalışmada 1998 yılına ait 127.886 kayıt kullanılmıştır. Bu çalışmada kullanılan veri madenciliği tekniklerinden lojistik regresyon analizi sonuçlarına göre dört biyomedikal değişkenin (vücut kitle indeksi, idrar proteini, kan glukozu ve kolesterol), demografik faktörlerden ise sadece yaş faktörünün tansiyonun tahmininde önemli olduğu, bunun dışında yaşam tarzı faktörlerinin hiçbirinin yüksek tansiyonun tahmininde önemli olmadığı saptanmıştır.[14]

İlaçlar da tıbbın önemli araştırma konularından biridir. İlaçların onaylanmadan önce faydalarının risklerinden daha çok olması koşulu göz önünde bulundurulur. Bazı ilaçlar piyasaya sürüldükten sonra risklerinin çok fazla görülmesi nedeniyle kaldırılmaktadır. İlaç etkileri analizi, ilaç üretimi ve geliştirilmesi gibi konularda da veri madenciliği tekniklerinden sık sık faydalanılmaktadır. Yapılan bir çalışmada antipsikotik ilaçların kalp kası hastalıkları üzerine etkileri yapay sinir ağları yöntemi ile araştırılmış ve bulunan sonuçlara göre klozapin dışındaki antipsikotik ilaçların miyokardi ve kardiyomiyopati ile önemli derecede ilişkili olduğu belirlenmiştir.[15]

Sağlık uygulamaları ve tedaviler büyük oranda maliyet gerektirirler. Yapılan tetkikler ve tedavilerdeki hile tespit çalışmalarının saptanması için de veri madenciliği tekniklerinden yararlanılabilir. Bu tür araştırmalarda doktor, tetkik ve hasta bilgileri analiz edilerek sıra dışı veriler incelenir. Sağlık kurumlarının finansal performanslarının ölçümünde de veri madenciliği tekniklerinden faydalanılmaktadır. Bu konuya yönelik yapılan bir çalışmada Türkiye'de bulunan 645 halk hastanesi karar ağaçları yöntemlerinden CHAID analizi yardımıyla incelenmiştir. On iki farklı finansal profil grubunun belirlendiği bu çalışmada finansal performansı etkileyen 5 faktör arasından öz kaynakların aktiflere oranı finansal performansı en çok etkileyen değişken olarak bulunmuştur.[16] Bu uygulamaların yanı sıra hastanelerdeki hasta bakım kalitesini iyileştirebilmek amacıyla hasta bakım kalitesi ölçülerinin analizinde de veri madenciliği yöntemleri kullanılabilmektedir. Bu konu ile ilgili olarak 1 Aralık 2000 ve 31 Ocak 2001 tarihleri arasında 8405 hasta üzerinde yapılan bir araştırmada karar ağaçları yöntemi kullanılarak yatan hasta mortalitesine etki eden faktörler incelenmiş ve yatış süresi, hastalık türü, taburcu edilen bölüm ve yaş faktörlerinin yatan hasta mortalitesinde önemli olduğu belirlenmiştir.[17]

Tıbbi araştırmaların önemli konularından biri de genetik bozuklukların tespitine yönelik araştırmalardır. Genetik hastalıklar, nesilden nesile aktarılabilen, kimi zaman ise soyağacı içinde ailenin tek bir üyesini ilgilendiren özelliklerdir. Bazı genetik hastalıklar tek bir genin mutasyona uğramasıyla ortaya çıkabilirken bazı hastalıklar birçok küçük etkili genin çevreyle etkileşmeleri sonucu yaşamın herhangi bir döneminde ortaya çıkar. 2001 yılında yapılan bir araştırmada "rSNP_Guide" adı verilen

bir veri madenciliği sistemi kullanılarak DNA dizileri araştırılmış ve genetik hastalıklara sebep olabilecek mutasyonların DNA dizilerindeki düzenleyici bölgeleri bu yöntemle tespit edilmiştir.[18] Günümüzde kanser hastalığı ile genler arasındaki bağlantının ortaya çıkarılması ile kanser geni araştırmaları da hız kazanmış, bu konuda önemli gelişmelere imza atılmıştır. Shah ve Kusiak tarafından yapılan bir araştırmada yumurtalık, prostat ve akciğer kanseri gen ekspresyonları veri setleri, bütünleştirilmiş bir gen-araştırma algoritması (karar ağaçları ve destek vektör makine algoritması) vasıtasıyla incelenerek kanser oluşumunda en fazla etkiye sahip genler belirlenmiştir.[19]

Tıp, ilk olarak hasta sağlığı ile ilgili bir alan, daha sonra bir araştırma konusudur.[12] Yinede günümüzde en fazla bilgi birikiminin yaşandığı alanlardan biri olan tıp alanındaki bilgilerden faydalanılarak önemli bilgiler elde etmek mümkündür. Veri madenciliği, sağlık ve tıp alanındaki büyük veritabanlarından değerli bilgileri ortaya çıkartarak, hem tıp hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlamaktadır. Söz konusu insan sağlığı olduğu için bu alandaki veri madenciliği çalışmaları önemli bir uygulama alanı bulacaktır. Bu konudaki önemi son yıllarda giderek artan çalışmalar ile ortaya koyulmaya başlamıştır.

KAYNAKLAR

1. Raghavan VV, Deogun JS, Sever H. Introduction. J Am Society Information Sci 1998; 49(5):397-402.
2. Akpınar H. Veri tabanlarında bilgi keşfi ve veri madenciliği. İ.Ü İşletme Fakültesi Dergisi 2000;29:1-22.
3. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to discovery knowledge in databases. AI Magazine 1996;3(17):37-54.
4. Jing L. Data mining and its applications in higher education. New Directions For Institutional Research 2002;113:17-36.
5. Glymour C, Madigan D, Pregibon D, Smyth P. Statistical Themes and Lessons for Data Mining. Data Mining and Knowledge Discovery1 1997;11-28.
6. Oğuzlar A. Veri ön işleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 2003;21:67-76.
7. Alkan A, Falay E. Kamu uygulamalarında çözüm veri madenciliğinde. Strateji Bülteni 2007;5:7-8.
8. Martínez-Muñoz G, Suárez A. Using Boosting to prune bagging ensembles. Pattern Recognition Letters 2007;28:156-65.
9. Fayyad U, Stolorz P. Data mining and KDD: Promise and challenges. Future Generation Computer Systems 1997;13:99-115.
10. Özekes S. Veri madenciliği modelleri ve uygulama alanları. İstanbul Ticaret Üniversitesi Dergisi 2003;3:65-82.
11. Ramkumar GD, Swami A. Clustering Data Without Distance Functions. IEEE Bulletin of the Technical Committee on Data Engisneering 1998;21(1):9-14.
12. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26(1-2):1-24.
13. Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. J Computer Sci 2006;2(2):194- 200.
14. Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. Int J Med Inform 2001;62(2-3):103-11.
15. Coulter DM, Bate A, Meyboom RH, Lindquist M, Edwards IR. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. BMJ 2001;322 (7296):1207-9.
16. Özgulbas N, Koyuncugil AS. Financial profiling of public hospitals: an application by data mining. Int J Health Plann and Manage 2007; (inpress).
17. Chae YM, Kim HS, Tark KC, Park HJ, Ho SH. Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications 2003;24:167-72.
18. Ponomarenko J, Merkulova T, Orlava G, Fokin O, Gorshkov E, Ponomarenko M. Mining DNA sequences to predict sites which mutations cause genetic diseases. Knowledge-Based Systems 2002;15:225-33.
19. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. Comput Biol Med 2007;37(2):251-61.