Hastane Bilgi Sistemlerinde Veri Madenciliği

Pınar YILDIRIM*, Mahmut ULUDAĞ**, Abdülkadir GÖRÜR*

(*) *Çankaya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara*
(**) *European Bioinformatics Institute, Cambridge, UK*
pınar@cankaya.edu.tr, uludag@ebi.ac.uk, agorur@cankaya.edu.tr

ÖZET

Günümüz Hastane Bilgi Sistemleri hastalara ve onların tıbbi durumlarına ait birçok veriyi barındırırlar. Bu verilerin çokluğuna ve zenginliğine rağmen, değerli bilgiler hastane ve klinik veritabanlarında gizlidir. Hastanelerin ve sağlık merkezlerinin verimliliğini artırmak ve gelecek planları yapabilmek için yararlı bilgilere ihtiyaçları vardır. Veri madenciliği teknikleri tıbbi verilerden gizli kalmış önemli bilgileri ortaya çıkarır ve böylece bu teknikler hastaneler ve klinik araştırmalar için değerli bilgiler sağlarlar. Bu çalışmada tıbbi veri madenciliği üzerine yapılan çalışmalar anlatılmış ve Hacettepe Üniversitesi Hastanelerinde yapılacak bir veri madenciliği çalısması ile ilgli kısa bilgi verilmiştir.

Anahtar Kelimeler: *Hastane Bilgi Sistemleri, Veri Madenciliği, Bilgi Keşfi*

Today's Hospital Information Systems have large quantities of information about patients and their medical conditions. Despite the richness of health-care data, useful information is often invisible in hospital and clinical databases. Hospitals and health centers need useful information to improve productivity and make strategic plans. Data mining techniques extract hidden and useful information from health-care data and therefore, these techniques provide valuable knowledge for hospitals and clinical research. This study presents a survey on medical data mining applications in the literature and briefly introduces a data mining study which will be done in Hacettepe University Hospitals.

Keywords: *Hospital Information Systems, Data Mining, Knowledge Discovery*

1. GİRİŞ

Günümüzde bilgi sistemleri ve iletişim teknolojilerindeki gelişmeler sayesinde tıp ve sağlık alanındaki birçok veri sayısal ortamda saklanabilmekte ve kolaylıkla erişilebilmektedirler. Hastane bilgi sistemleri hastalara ait demografik bilgiler, hastalık ve tedavi durumları, yapılan tetkikler, faturalama ve idari işlere ait bilgileri içerirler. Sağlık ve tıp, çağımızın en önemli bilimsel araştırma alanları olduğu için bu alandaki bilgi sistemleri de araştırmalar için en büyük veri kaynaklarıdır. Son otuz yılda dünyada sağlık bilgi sistemlerinde büyük gelişmeler yaşanmıştır. Sağlık Bilişiminin yeni bir alan olmasına rağmen özellikle bilgi modelleme ve tanı araçlarında hızlı yenilikler yapılmıştır [1].

Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına rağmen, bu alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamıştırlar. Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır.

Günümüzde tıbbın gelişmesi ve insanların ortalama yaşam sürelerinin uzaması beraberinde bazı sorunları da getirmiştir. Örneğin birçok insan, kalp hastalıkları, diabet 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 yeterli kalmamaktadır. Veri madenciliği yöntemleri kullanılarak bu sistemlerdeki gizli ve önemli bilgiler keşfedilmelidir. Keşfedilen bu bilgiler hem tıbbi araştırmalar hem de yönetim planları için değerlendirilmelidir.

2. VERİ MADENCİLİĞİ TEKNİKLERİ

Veri Madenciliği (Data Mining) ve Bilgi Keşfi (Knowledge Discovery), verilerde daha önceden bilinmeyen, anlamlı ve değerli bilgiler elde etme işlemidir. Temel olarak beş aşamadan oluşur:

- Veri seçimi
- Önişleme
- Dönüştürme
- Veri Madenciliği
- Yorumlama/Değerlendirme.

Veri seçme işlemi, üzerinde çalışılacak veritabanından veya diğer veri kaynaklarından verilerin seçilerek veri dosyası oluşturulmasıdır. Önişleme, anormal verilerin kaldırılması, eksik veya hatalı verilerin düzeltilmesi gibi işlemleri içerir. Dönüştürme, verilerin kategorize edilmesi, ilgili özelliklerin seçilmesi veya boyut azaltma işlemleridir. Veri madenciliği aşamasında ise, uygun bir algoritma seçilerek hazırlanmış verilere uygulanır. En son aşamada keşfedilen bilgiler ve özellikler yorumlanır ve değerlendirilir. Veri Madenciliği bilgi keşfi araştırmalarının bir parçasıdır ve istatistik, makine öğrenmesi ve örüntü tanıma gibi araştırma alanları ile ortak çalışır [2].

Genellikle veri madenciliği modelleri tahmin edici ve tanımlayıcı olmak üzere ikiye ayrılırlar. Tahmin edici modeller bilinen verilerden yararlanarak, bilinmeyen bir değeri tahmin etmeye çalışırlar. Tanımlayıcı algoritmalar ise verilerdeki gizli ortak özellikleri ve ilişkileri araştırırlar.

Sınıflandırma (Classification), regresyon (regression) ve zaman serileri(Time Series) analizi gibi yöntemler tahmin edici algortimalardır. Kümeleme (Clustering), özetleme(Summarization) ve ilişkisel kural madenciliği(Association Rule Mining) gibi algoritmalar da tanımlayıcı yöntemlerdir.

Geleneksel veri madenciliği algoritmaları çoğunlukla tek bir tablo veya düz bir dosya üzerinde çalışırlar. Oysaki gerçek yaşamda çoğu veriler ilişkisel veritabanlarında saklanırlar. Bu sistemlerde birden çok tablo kullanılır ve bunların birbirleriyle karmaşık ilişkileri vardır. Geleneksel yaklaşımda veritabanındaki ilişkiler kaybolabilir veya bazı veriler tekrar edebilir. Bu nedenle çoklu ilişkilere sahip veritabanları üzerinde veri madenciliği çalışmaları yapabilmek için hem tahmin edici hem de tanımlayıcı veri madenciliği algoritmalarını içeren İlişkisel Veri Madenciliği (Relational Data Mining) algoritmaları geliştirilmiştir. Bu algoritmalar tahmin ve tanımlarını Çoklu İlişkisel Karar Ağacları (Multi Relational Decision Trees), ve Çoklu İlişkisel Kurallar (Multi Association Rules) ile tarif ederler.

Metin madenciliği de bir veri madenciliği yöntemidir ve yapısal olmayan metinlerden bilgi keşfi yapılmasını sağlar. Yaygın olarak aynı konuda yazılmış belgeleri bulmak, birbiriyle ilişkili belgeleri bulmak, kavramlar arası ilişkileri keşfetmek için kullanılır. Doğal Dil İşleme (Natural Language Processing), Bilişsel Bilimler (Cognitive Sciences) ve Makine Öğrenmesi (Machine Learning) gibi bilimlerle ortak çalışan bir araştırma alanıdır.

3.SAĞLIK ALANINDA VERİ MADENCİLİĞİ UYGULAMALARI

Sağlık alanında yapılan birçok veri madenciliği araştırmalarında hastaların elektronik tıbbi kayıtları ve idari işleri belgeleyen veriler kullanılmaktadır. Bu verilerden yararlanılarak farklı tahminler yapılabilir. Örneğin 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 edilmesi [1].

Örneğin Tablo 1. de hastalığın olup olmaması durumuna göre hastaların yaş, tansiyon ve sigara kullanımı gibi bilgileri verilmiştir. Bir veri madenciliği algoritması bu verilerden yararlanarak hastalığın olup olmamasına dair kurallar çıkarabilir.

Tablo 1. Hastalık Sınıflandırma Veri Seti

Örnek No	Yaş	Tansiyon	Sigara Kullanımı	Hastalık (Class)
1	25	Yüksek	Kullanmıyor	var
2	37	Normal	Bazen	yok
3	37	Normal	Sıklıkla	var
4	53	Normal	Bazen	yok
5	53	Düşük	Herzaman	var
6	53	Normal	Herzaman	yok
7	53	Düşük	Genellikle	var

Hastalıkların yönetimi ile ilgili veri madenciliği çalışmaları hastalıkların ve durumlarının tanımlanmasını ve maliyetlerin modellenmesi gibi araştırmaları içerir. Bu çalışmalarda amaç pozitif sonuç elde etmektir. Örneğin Harleen Kaur ve arkadaşları hastaların yaş ve cinsiyet gibi verilerini karar ağacı yöntemleri ile analiz ederek göğüs kanseri olup olmadığını tahmin etmeye çalışmışlardır [3].

Hastane bilgi sistemlerindeki verilerle yapılmış diğer bir çalışmada, hastaların sık sık farklı doktorları ziyaret etmeleri araştırılmış ve hasta demografik bilgileri ve işlemsel veriler analiz edilmiştir. İlişkisel kural analizi (Association rules analysis) kullanılarak yapılan veri madenciliği çalışmasında, yaşın, cinsiyetin, hastanelerin özelliklerinin, kronik ve akıl hastalıklarının sürekli doktor ziyaret etme davranışında etkili oldukları ortaya çıkartılmıştır[4].

Hastanelerde maliyetleri etkileyen en önemli konulardan birisi de hastaların kalış süreleridir. Kalış sürelerinin etkileyen faktörler de günümüzde veri madenciliği çalışmalarının araştırma konusudur ve birçok çalışma yapılmıştır. Örneğin, yapılan bir çalışmada hastaların demografik ve çevresel bilgileri, sinir ağları ile analiz edilmiş ve bazı önemli bilgiler elde edilmiştir. Bu bilgilere göre 40 yaşından büyük hastalar, şehirlerde yaşayan hastalar, alkol ve sigara bağımlılığı olan hastalar daha uzun süre hastanede kalmaktadırlar. Ayrıca özel hastanelerdeki kalış süreleri devlet hastanalerinden daha kısadır[5].

Sağlık uygulamaları ve tedaviler büyük oranda maliyet gerektirirler.Yapılan tetkikler veya tedavilerden hile yapılarak çıkar sağlanmaya çalışılabilir. Özellikle Avrupa ve Amerika'da sağlık sigorta şirketleri bu konuları araştırmaktadırlar. Hile tespiti için veri madenciliği yöntemlerinden yararlanılır. Bu tür araştırmalarda hasta, tetkik ve doktor bilgileri analiz edilir ve anormal veriler incelenir. Örneğin ortalama maliyetin üzerindeki tetkikler veya tedaviler şüphe kaynağıdır. Bu çalışmalarda genellikle kümeleme(clustering) algoritmaları kullanılır.

İlaçlar da tıbbın önemli araştırma konularından birisidir. Amerika Birleşik Devletleri'nde yeni bir ilaç geliştirildiğinde, klinik denemelerden sonra FDA (Food and Drug Administrators) kurumu tarafından onaylanarak piyasaya sürülür. Onaylanmadan önce ilacın faydalarının risklerinden daha çok olması gözönünde bulundurulur. Bazı ilaçlar piyasaya sürüldükten sonra risklerinin çok fazla görülmesi nedeniyle kaldırılmışlardır. İlaçların önceden tanımlanmamış yan etkilerinin bulunabileceği olasılığı, web üzerinden tıbbi yayınlar analiz edilerek veri madenciliği çalışmaları da yapılmaktadır [6].

4.TIP VE BİYOİNFORMATİK ALANLARINDA VERİ MADENCİLİĞİ ÇALIŞMALARI

Tıp ve sağlık alanındaki verilerin birçoğu yapısal olmayan metinlerde saklanmaktadırlar. Örneğin, hastaların tıbbi durumları, tanı, tedavi bilgileri ve klinik dokumanlar metin olarak saklanmaktadırlar. Ayrıca, uygulanan işlemlere ait faturalar ve işakışını belgeleyen raporlar da metin formatındadır. Tıp alanındaki bilimsel makaleler de sağlık alanında yapılan araştırmalar ve yenilikler için değerli bilgi kaynaklarıdır ve metinsel yapılarda saklanırlar. Bu yapılar üzerinde bilgi keşfi yapmak için metin madenciliği yöntemleri kullanılmaktadır. Metin madenciliği Doğal Dil İşleme (Natural Language Processing), Tıp Bilişimi (Medical Informatics) ve İstatistik gibi alanlarla ortak çalışılan bir araştırma alanıdır. Tıp alanında yapılan metin madenciliği çalışmalarında qzellikle Medline gibi tıbbi alanda yapılmış bilimsel yayınların saklandığı büyük veritabanları için bilgi

keşfetme yöntemleri geliştirilir ve bu çalşımaların amacı metin yapısındaki verilerin analiz edilip bilgi keşfi yapmak ve bilgi yönetimi sağlamaktır. Tıp alanındaki makalelerden tedavi ve tanı ile ilgili yeni yaklaşımlar, kavramlar arasındaki gizli ilişkiler ortaya çıkartılabilir. Elde edilen önemli bilgiler hem araştırmalara büyük destek sağlar hem de sağlık kurumlarının başarısını artırır.

Avrupa Biyoenformatik Enstitüsü (European Bioinformatics Institute) Metin Madenciliği araştırma grubu biyomedikal alandaki bilimsel makalelerden bilgi keşfi yapmak için araştırmalar yapan bir gruptur. Bu grup, Tıbbi Ontoloji (Medical Ontology), Anlamsal Ağlar (Semantic Networks), Doğal Dil İşleme (Natural Language Processing) ve Birleştirilmiş Tıbbi Dil Sistemi (Unified Medical Language System-UMLS) gibi teknikler kullanarak tıbbi makaleler üzerinde metin madenciliği çalışmaları yapmaktadır. Bu araştırma grubu aynı zamanda birçok yazılım sistemi ve aracı geliştirmiştir. Bunlardan EBIMed sisteminde girilen anahtar kelime için, Medline veritabanındaki özetler bulunur ve biyoinformatik kaynaklardaki biyomedikal terminolojileri içeren cümleler seçilir. Seçilen bu cümle ve terminolojilerle aynı kavram için tanımlanan proteinler, gen ontoloji (Gene Ontology) açıklamaları, ilaçlar ve canlı türleri bulunarak genel bir tabloda görüntülenir. Bulunan özetler ve seçilen cümleler diğer biyomedikal veritabanlarındaki varlıklarla ilişkilidir [7].

Tıbbi gelişmelerde biyolojik araştırmaların da büyük katkısı vardır. Günümüzde biyolojik yapılar ve özellikle genlerle ilgili büyük uluslararası veritabanları ve bunlara erişimi kolaylaştıran yazılım araçları kullanılmaktadır. Örneğin BIOMART, Avrupa Biyoinformatik Enstitüsü ve Cold Spring Harbor Laboratuarı (CSHL) tarafından geliştirilmiş, ilişkisel veriler için biyoinformatik dünyasında yaygın olarak kullanılan bir veri madenciliği sistemidir [8]. Bu sistem karmaşık ilişkilere sahip biyolojik veriler için gelişmiş sorgulama arayüzlerinin kolayca üretilmesini ve bu arayüzlerden gelecek sorguların etkin bir şekilde cevaplanmasını sağlar. OMIM (Online Mendelian Inheritance in Man) projesi ise National Center for Biotechnology Information (NCBI) tarafından geliştirilmiş genetik bozukluklarla ilgili bilinen hastalıkların saklandığı bir veritabanıdır. Bu veritabanı metin bilgiler, resimler ve referans bilgilerinden oluşmuştur ve ayrıca Medline veritabanına da bağlantısı vardır. Büyük bir bilgi kaynağına sahip olan veritabanı genlerle ilgili araştırmalarda önemli katkılar sağlar [9].

İnsan Genom projesinin tamamlanması ile birlikte elde edilen bilgilerin daha iyi anlaşılması ve kullanıma dönüştürülmesi için başlatılan projelerin önemli bir hedefi genom bilgileri ile klinik bilgiler arasında bağ kurabilmek ve bu bağı hastalıklar ve genetik diziler arasındaki ilişkinin daha iyi anlaşılmasında kullanmaktı. Örneğin HapMap projesi ile genomun insan bireyleri arasında değişimini daha iyi anlamak ve bu değişimle hastalıklar ve hastalık hassasiyetleri arasındaki ilişkileri belirlemek hedef alındı [10]. Daha önce İnsan Genom Projesinin Avrupa ayağı olan Sanger Enstitüsünde yürütülmekte olan DECIPHER projesini de bu konuda çalışılan projelere örnek olarak verebiliriz [11]. Birçok ülkeden hastanelerin bilgilerini paylaştığı DECIPHER projesinde insan genomunda kromozom düzeyindeki değişim ile hastalıklar arasındaki ilişkinin anlaşılması için kapsamlı bir veritabanı oluşturulmaktadır.

5. HACETTEPE ÜNİVERSİTESİ RADYOLOJİ BİLGİ SİSTEMİNDE YAPILACAK VERİ MADENCİLİĞİ ÇALIŞMALARI

Hacettepe Üniversitesi Hastaneleri ülkemizin en büyük üniversite hastanelerinden birisi olup birçok alanda öncü ve yenilikçi bir sağlık kurumu olarak bilgi sistemlerinin kullanımı açısından da diğer sağlık kurumlarına örnek teşkil etmiştir. Hastane genel bilgi sistemiyle tümleşik olarak çalışan Radyoloji Bilgi Sistemi, hastalar, tetkikler ve işakışlarına ait verilerden oluşan büyük bir ilişkisel veritabanına sahiptir. Radyoloji bölümünde kaynakların daha doğru planlanması, gelecek planları yapılabilmesi ve tıbbi açıdan yapılan

çalışmalara katkıda bulunabilmesi için bu verilerden çıkartılacak değerli bilgilere ihtiyaç vardır. Bu nedenle radyoloji veritabanı üzerinde ilişkisel veri madenciliği (Relational Data Mining) algoritmaları kullanılarak veri madenciliği çalışmaları yapılacak ve değerli bilgiler keşfedilmeye çalışılacaktır.

6. SONUÇ

Sağlık ve tıp, günümüzün en çok bilgi ihtiyacı olan araştırma alanlarıdır. Son yıllarda özellikle sağlık veri modelleri, standartlar ve kodlama sistemlerindeki yenilikler sayesinde hastanelerde ve sağlık merkezlerinde kullanılan bilgi sistemlerinde önemli gelişmeler yaşanmıştır. Bu gelişmeler daha çok ve çeşitli verinin saklanabilmesini sağlamış ve beraberinde bilgi keşfi ihtiyacını ortaya çıkarmıştır. Veri Madenciliği, sağlık ve tıp alanındaki büyük veritabanlarından değerli bilgileri ortaya çıkartarak, hem tıp açısından hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlar. Günümüzde uluslarası ortak projeler kapsamında geliştirilen ve biyoloji verilerinin saklandığı veritabanları, bu veritabanlarına erişim ve veri madenciliği sistemleri de klinik araştırmaların önemli bir parçası haline gelmişlerdir.

Bu çalışmada dünyadaki sağlık ve tıp alanındaki yapılan veri madenciliği çalışmaları tanıtılmış ve Hacettepe Üniversitesi Hastanelerinde yapılacak çalışma hakkında bilgi verilmiştir.

9. KAYNAKLAR

[1] Kudyba, S., "Managing Data Mining", CyberTech Publishing, 2004, 146-163.

[2] Han, J., ve Kamber, M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[3] Kaur., H., ve Wasan., S., "Empirical Study on applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 2006.

[4] Chen., Y., ve Wu., S., "Exploring Out-Patient Behaviors in Claim Database: A Case Study Using Association Rules", AMIA Symposium Proceedings, 2003.

[5] Nagadevara., V., "Application of Neural Prediction Models in Healthcare".

[6] Carino., C., Jia., Y., Lambert., B., West., P., Yu., C., "Mining Officially Unrecognized Side Effects of Drugs by Combining Web Search and Machine Learning", CIKM'05 Oct 31-Nov-5, 2005 Bremen, Germany.

[7] Rebholz-Schuhmann,D. , Kirsch,H. , Arregui,M. , Gaudan,S. , Riethoven,M. , Stoehr,P. EBIMed--text crunching to gather facts for proteins from Medline, *Bioinformatics* 2007, 23 (2):e237-44

[8] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439-40.

[9] Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM, National Center for Biotechnology Information (NCBI).

[10] International HapMap Project, http://www.hapmap.org, International HapMap Project Organization.

[11] Decipher Project, http://decipher.sanger.ac.uk., Sanger Institute.