AJIT-e: Online Academic Journal of Information Technology
2011 Fall/Güz – Cilt/Vol: 2 - Sayı/Num: 5
DOI: 10.5824/1309-1581.2011.4.005.x
http://www.ajit-e.org/?menu=pages&p=details_of_article&id=52

Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları: Kanser Teşhisine Yönelik Bir Ön Çalışma

Data Mining Studies in Medical and Healthcare: A Preliminary Study for Cancer Diagnosis

Sabri Serkan Güllüoğlu
İstanbul Arel Üniversitesi, Mühendislik Mimarlık Fakültesi
serkangulluoglu@arel.edu.tr

ÖZET:

Bilgiye sahip olmanın ve onu kullanmanın önemli olduğu günümüzde güçler dengesi bilgi üzerine yoğunlaşmaktadır. Çeşitli kaynaklardan ve yöntemlerle toplanan bilgilerin belirli bir disiplin ve sistem dâhilinde analiz edilmesi sonucunda ortaya çıkan sonuçlar, ekonomik, siyasi ve teknolojik alanlarda kullanılmaktadır. Bilgiyi zamanında ve doğru olarak kullananlar istedikleri sonuca kestirmeden ve süratli bir biçimde ulaşmaktadırlar.

Veri madenciliği ile eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilgi çıkarılabilir. Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır.

Bilgi kaynağının yanı sıra, bilginin doğruluğu da önemli bir sorundur. Bir bilginin veya daha somut ifadeyle mesela bir rakamın doğru olup olmadığı nasıl anlaşılmaktadır? Bilginin doğruluğu konusunda iki kriter vardır. Aynı sonucu işaret eden verilerin yoğun olması bilginin doğru olduğu yönündeki ilk kriterdir. Bir değer ne kadar yoğunsa o kadar inandırıcı olmaktadır. Ne kadar güçlü bir ilişki olduğu tespit edilirse, o kadar doğruluğuna hükmedilebilir. Hangi miktarda verinin toplanması gerektiği ayrı bir sorundur. Veri miktarı, kullanılan metoda bakılmaksızın çalışmanın amacına göre belirlenmektedir.

Gün geçtikçe çoğalan veri yığınlarından anlamlı ve faydalı bilgiye ulaşabilmek için "veri madenciliği" başlığı altında yöntemler geliştirilmeye başlanmıştır.

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Çalışmanın amacı Tıp alanında uygulanması düşünülen veri madenciliğe çalışmalarına örnek teşkil etmesi açısından bir plan çıkarmaktır.

Bu hususta bakıldığında çalışmanın amacı geliştirilecek yöntembilim ile saklı olan ve bilinmeyen bilgilere ulaşmaktır. Bunun için farklı tipteki veriler sınıflandırılacak, eğitilecek yeni veriler test edilecek ve yordama yapılacaktır. Böylece kaynaktan hedefe giden süreçte hedef karar vermede etkilenecektir. Bu şekilde çıkarılmak istenen bilgiye ulaşılmış olacaktır.

DOI: 10.5824/1309-1581.2011.4.005.x

1.GİRİŞ

Bu araştırmada amaç sağlam kişilerin ileride yakalanması ihtimal kanser tipini belirlemek, sağlıklı veya riskli teşhisi koymaktır. Buna istinaden öncelikli olarak daha önce kanser hastalığına yakalanmış kişilerin kişisel verilerine ihtiyaç duyulacaktır. Elde edilen bu verilerde yaş, cinsiyet, genetik yapı, meslek, yaşadığı çevre ve yakalandıkları kanser tipi gibi özellikleri olmasına özen gösterilecektir. Dağınık olan bu verilerin ilk olarak sınıflara ayrılması işlevi için veri madenciliği kullanılacaktır. Veri madenciliği yöntemlerinden biri olan yapay sinir ağları sınıflandırma yöntemi kullanılacaktır. Sınıflara ayrılan bu verilerden sağlıklı kişilerden alınacak bazı test sonuçları ve kişilik özellikleri baz alınarak yapay sinir ağları yöntemiyle önceden eğitilen veriler test edilip, bu kişilerin hasta, hasta değil veya riskli oldukları anlaşılacaktır.

Bu çalışma önerisi çerçevesinde, çalışmanın ana hatlarını belirlemek maksadıyla kanser alanındaki tıp uzmanlarıyla iletişim düşüncesi oluşmuştur. Kanser prognozunda kullanılması düşünülen tekniklerin uygulanabilmesi için öncelikli olarak işin uzmanları ile fikir alışverişi sağlanmalıdır. Verilerin elde edilmesi çalışma için en önemlisidir. Bu düşünceden yola çıkılarak hasta yoğunluğunun fazla olduğu bir Hastanesinin Radyasyon Onkoloji kliniği, patoloji klinik ve laboratuvarı doktorları ile randevular sağlanarak görüşmeler yapılması gerekir. Onkoloji kliniği doktorları ile görüşmeler sırasında doktorlara yöneltilen sorular aşağıda listelenmiştir.

1- Verilerin en fazla olduğu kanser tipleri hangisidir?
2- Hangi kanser tipleri büyük bir oranda sadece kan ve vs. testlerle anlaşılabilir?
3-Bu kanser tipleri arasında ilişkiler varmı?
4- Bu kanser tiplerine ait veriler hangi üniteden ve nasıl alınabilir?
5-Sizlerin Kanser teşhisinde sıkıntı çektiği konular nelerdir?
6-Kanser teşhisinde elinizde Türkiye kanser haritasının bulunması faydalı olurmu?
7-Hastalara ait ne tip veriler barındırılır?
8-Kanser hastalığında genetik faktörler ne derece etkilidir.

2. ÇALIŞMANIN ÖNEMİ

Bu uygulamadan elde edilecek sonuçların özellikle:

1. Kanser hastalığının erken teşhisinde veri madenciliği ilkelerinden hareketle hastalığa yakalanmış ya da yakalanması muhtemel kişiler için kullanılabilecek bir erken teşhis yaklaşımı olduğunu göstereceği;

2. Eğitim ve Araştırma hastanelerinde hekimler tarafından uygulanması planlanan erken teşhis yaklaşımlarının etkinliğini ölçmede yararlanılabilecek bir metodoloji oluşturacağı umulmaktadır.

3. Bu modelde giriş olarak özellikle hastane arşiv verileri ve internet ortamından alınan verilerin çıkış sonuçlarının geçerlilik ve güvenilirliği analiz edilmesi gerekir.

3.YÖNTEM

3.1. Araştırma

Veri madenciliği analizinde Matlab programı kullanılabilir. Programın özellikleri şunlardır.

- Matematik ve hesaplama işleri, algoritma geliştirme.
- Modelleme, benzetim ve prototipleme.
- Verilerin analizi, incelenmesi ve görüntülenmesi.
- Bilimsel ve mühendislik alanında grafik işlemleri.
- Grafiksel kullanıcı arayüz yapısını da içine alan uygulama geliştirme.

Program veri madenciliği uygulamaları gerçekleştirebilecek araç kutularını içerisinde barındırır. Araştırmada kullanacağımız yapay sinir ağları özelliği mevcuttur.

3.2. Analiz

Tıbbi kaynakların son derece kısıtlı olması, var olan kaynakların da etkin kullanılamaması sonucu dünyada her yıl yüz binlerce kişi hayatını kaybediyor. Tıpta ve sağlık sistemlerinde sayısal (kantitatif) tekniklerin kullanılması ile hasta kayıpları azaltılabiliyor. Karar verme probleminin olduğu hemen her yerde matematiksel modeller kullanılabilir. Kanser, DNA'nın hasarı ile hücrelerin programdan çıkması sonucu hücrelerin kontrolsüz bir şekilde veya anormal bir şekilde büyümesi ve çoğalması sonucu oluşan genetik bir hastalıktır. Kanser ne kadar erken teşhis edilirse, tedavisi de o düzeyde başarılı olur. Tıp, istatistik ve veri madenciliği gibi teknikleri kendi alanlarında kullanabilirse gelecekte kanser gibi birçok hastalık erken teşhis sayesinde ilaçla tedavi edilebilir. Böylece pahalı ameliyatlara gerek kalmayabilir. Günümüzde kansere yakalanan kişilerin çoğu hastalığın ilerlemiş safhalarında
hastanelere başvurmakta ve bu sebeple geç teşhis edilmektedir. Bunun sonucunda tedaviler çoğu zaman işe yaramamakta ve hasta kısa zamanda ölmektedir. Sağlam kişilerde ileriye yönelik kanser hastalığının teşhisi üzerinde durulması gereken en mühim konulardan biridir. Ülkemizde kanser hasta kayıtlarının düzenli bir ortamda tutulmadığı açıktır. Hâlbuki tutulacak kayıtlar sayesinde ileriye yönelik daha hızlı karar verme teknikleri oluşturulabilir. Yapılan araştırmalarda kanser hastalık türlerinin birbirleriyle bağıntılı olduğu anlaşılmaktadır. Değerlendirme aşamasına geçilmeden tek tek ele alınan hastalıkların teşhisi fazla zaman alabilmektedir. Öncelikle yapılması gereken pilot seçilecek kanser hastalık tiplerinin belirlenmesi, daha sonra disiplinler arası kullanımı yaygın, matematik altyapısına sahip veri madenciliği yöntemlerinden en uygun olanının seçimine karar verilmesidir.

1.Kanser tipleri belirlenir. Göğüs, akciğer ve bağırsak kanserleri.
2.Bu kanser tiplerine ilişkin hangi özellikteki verilerin alınması kararı verilir.
3.Belirlene kanser tiplerine ilişkin veriler büyük hasta potansiyeline sahip bir hastanenin onkoloji laboratuvarından ve internet ortamından temin edilir.

DOI: 10.5824/1309-1581.2011.4.005.x

4.Yapay sinir ağları yönteminde yer alan modelleri sırayla denenmek üzere ilk modelden başlanır.
5. 3 farklı kanser tipine ait veriler sınıflandırılmak ve eğitilmek üzere yapay sinir ağları modeline verilir.
6. Verilen girişler ile çıkış verileri eğitim verileri neticesinde karşılaştırılır.
7. Tüm modeller denendiğinde en fazla doğruluğa hâkim model seçilir.
8. Test verileri programa sunularak yeni hastalar için bilgiye ulaşılmış olur.

4. VARSAYIMLAR

Hipotezler

Büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kurallar, geçerli ve güvenilir bilgiye ulaşmada metot oluşturacaktır.

Bu araştırma hipotezinin testi için birtakım alt hipotezlerinde formüle edilmesi gerekmiştir. Bu hipotezler şunlardır:

Alt Hipotez1: Kanser hastalığının erken teşhisinde yapay sinir ağları yönteminin kullanılması Türkiye bazında kanser haritası oluşumuna destek olacaktır.

Alt Hipotez2: Kanser hastalığının erken teşhisinde yapay sinir ağları yönteminin kullanılması sağlıklı bireylerin gelecekte yakalanma ihtimallerine karşı önlem alınmasını sağlayacaktır.

Alt Hipotez3: Kanser hastalığının erken teşhisinde yapay sinir ağları yönteminin kullanılması hekimlerin hasta üzerinde teşhisini kolaylaştıracaktır.

Alt Hipotez4: Kanser hastalığının erken teşhisinde yapay sinir ağları yönteminin kullanılması bazı testlere gerek bırakmayacak şekilde hastane maliyetlerini düşürecektir.

Alt Hipotez5: Kanser hastalığının erken teşhisinde yapay sinir ağları yönteminin kullanılması hastalıktaki genetik faktörün ne kadar önemli olduğunu ortaya koyacaktır.

Alt Hipotez6: Kanser hastalığı riski yapay sinir ağları yönteminin kullanılması sayesinde ortaya çıkan kişi bu hastalık sınıfına dahil olmamak için gerekli önlemleri alacaktır.

Alt Hipotez7:Verilerde kullandığımız X tipi kanser teşhisi konulan hastaların hasta olmadan önceki durumları göz önüne alındığında bazı özellikler x tipi kanser hastalığı için belirleyici olabileceği çıkarımı yapılacaktır.

Alt Hipotez8: Yapay Sinir Ağları modeli için giriş oluşturacak Hastane arşiv verilerinin model sonucunda oluşacak ifadelerin güvenilirliği, internet ortamından alınacak verilerin geçerlilik ve güvenilirliğinde sonuçlarından fazladır.

Sınırlılıklar

1. Araştırma belirlenen Hastane arşivi ve internet ortamında google arama motoru ile sınırlıdır.
2. Teşhis için uygun görülen kanser tipleri göğüs, akciğer ve bağırsak kanserleri ile sınırlıdır.
3. Araştırmada ele alınan değişkenler, uygulanan ölçüm araçlarının güvenilirlik ve gerçeklik boyutlarıyla sınırlıdır.
4. Sosyal bilimler alanında yapılan çalışmaların tamamen deneyselliğe oturtulmamasından kaynaklanan sınırlılık, bu çalışma için de geçerlidir.

5. VERİ TOPLAMA TEKNİĞİ

Araştırma Modeli

Ulaşmaya çalıştığımız problemin çözümü için gerekli olan veriler belirlenen Hastanenin Radyoloji Onkoloji bölümü arşivinden elde edilecektir. Elde edilen veriler sonuçlarla ilişkilendirilerek yapay sinir ağımızın eğitiminde ve test setlerinde kullanılacaktır. Bu sayede yeni gelen verilerde ağımız sonucu tahmin yeteneği kazanacaktır. Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenenecektir.

Evren ve Örneklem

Amaç ve sınırlılıkta da belirtildiği üzere araştırma kişinin kanser hastalığının teşhisi ve kansere yakalanma riskinin ortaya çıkmasına yöneliktir. Elde edilecek sonuçların genellenmek isteneceği çalışma evreni Onkoloji birimleridir.

Veriler ve Toplanması

Araştırmada kişilerin çeşitli kanser tipleri ile ilgili hastalığı belirleyen kriterlerin yer aldığı bir hastanenin Onkoloji biriminden veriler alınacaktır. Teşhis ve risk durumu için incelenecek olan Kanser tipleri; bağırsak, göğüs ve akciğer kanseri olarak seçilmiştir. Bunun sebebi günümüzde en sık rastlanan kanser tiplerinden olmaları ve bu hastalık tiplerinde veri sıkıntısı diğer kanser tiplerine göre daha az olmasından ötürüdür. Veriler belgesel tarama yöntemiyle derlenmiştir. Toplanan bu veriler araştırmacı tarafından tasarlanacak ve uygulamadan elde edilen sonuçlarla teşhis sağlanacaktır.

DOI: 10.5824/1309-1581.2011.4.005.x

Verilerin Çözümü ve Yorumlanması:

Elde edilen hasta verileri yapay sinir ağlarına verilip eğitim tamamlandıktan sonra yeni veriler ağa sunulup ağın bulduğu sonuçlarla hedef değerlerinin karşılaştırılması yapılacaktır. Başarı oranı kabul edilebilir düzeye gelene kadar ağımızın eğitimine devam edilecektir.
Araştırmada çalışılan 3 kanser tipi, olan göğüs-akciğer ve bağırsak kanserleri araştırılan konu için pilot olarak seçilmesi uygun görülmüştür. Başarı oranı istenilen seviyeye ulaştığından dolayı bu tip bir uygulama diğer kanser tiplerinde de başarı ile uygulanabilir olması amaç olacaktır.

Çözümlemeler Matlab programından yararlanılarak gerçekleştirilecektir. Bu amaçla verilerin eğitimi tamamlanıp yeni veriler için ağımız sağlıklı sonuçlar çıkarabilecektir.

Yapay Sinir Ağları modelinde giriş olarak kullanacağım arşiv verileri ve İnternet ortamından alınan verilerin çıkış sonuçlarının geçerlilik ve güvenilirliği yorumlanacaktır.

6. SONUÇ

Veri madenciliği en basit tanımı ile çok büyük miktardaki ham veriler içinden amaca uygun modellerin ortaya çıkarılması işlemidir. Başka bir tabirle karmaşık ve düzensiz veriler içindeki modellerin ortaya çıkarıp bunları karar verme ve eylem planını gerçekleştirmek için kullanma sürecidir.

Bu makalede sağlıkta uzman kişilere sağlık sektöründe Veri Madenciliği'nin kullanımı ile ilgili bir çalışmanın nasıl olması gerektiği hususunda bilgi sunarak karar verme süreçleri açısından yeni bir bakış açısı kazandırmak amaçlanmıştır.

Veri Madenciliği, sağlık profesyonellerinin en doğru ve güncel bilgiye ulaşmasını, en objektif ve optimum çözümleri kullanmasını sağlayacak bir karar destek aracıdır. Geleceğin sayısal karar verme ve iş zekası yöntemi olan Veri Madenciliğinin konunun uzmanı kişiler tarafından sağlık sektöründe kullanımı, sağlık hizmetlerinin daha etkin sunumu, kaynakların daha verimli kullanımı ve bilimsel, karşılaştırılabilir, şeffaf bilgi erişimi açısından önerilmektedir.

7. KAYNAKLAR

[1] Cabena P., Hadjinia P. n, Stadler R., Verhees J., Zanasi A.,(1997) "Discovering Data Mining: From Concept To Implemantation", Prentice Hall PTR, Upper Saddle River, New Jersey, 195, USA.

[2] Carino., C., Jia., Y., Lambert., B., West., P., Yu., C., (2005)"Mining Officially Unrecognized Side Effects of Drugs by Combining Web Search and Machine Learning", CIKM'05, Bremen, Germany.

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

[4]Nagadevara., V., (2006)"Application of Neural Prediction Models in Healthcare".

[5] Obenshain K. M.,(2004) "Application of Data Mining Techniques To Healthcare", Data Infect Control Hosp Epidemiol, 25, 690–695, DOI: 10.1086/502460.

[6] Prather J. C., Lobach D. F., Goodwin L. K., Hales J. W., Hage M. L., Hammond W. E., (1997)."Medical Data Mining: Knowledge Discovery In A Clinical Data Warehouse", Proc AMIA Annu Fall Symp. 101–105.

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

[8] Zhong N.. – Zhou L., (1999). "Methodologies for Knowledge Discovery and Data Mining" : Third Pacific-Asia Conference, Pakdd-99, Beijing, China, Proceedings,