Laboratuvar Test İstemleri Üzerinde Veri Madenciliği: Destek Vektör Makineleri ile Kadın Genital Bölge Kanserleri Tanı Kontrolü

Kemal TURHAN[a], Yasemin Zeynep ENGİN[a], Burçin KURT[a]

[a] *Biyoistatistik ve Tıp Bilişimi AD, Karadeniz Teknik Üniversitesi, Trabzon*

Data Mining on Laboratory Test Orders: Female Genital Cancer Diagnosis Control by Using Support Vector Machines

Abstract: *Laboratory tests orders contains very intense and equally important data for Data Mining. In this study, we have been developed and tested a model for diagnostic accuracy and consistency of "Malignant neoplasms of female genital organs". Model based on two important data mining techniques; such as Association Rule Mining and Support Vector Machines.*
Key Words: Data Mining; Support Vector Machine; laboratory tests; ICD

Özet: *Laboratuvar testleri, veri madenciliği açısından incelenmesi gereken çok yoğun ve son derece önemli veriler içermektedir. Bu çalışmada Birliktelik Kuralları (Association Rules) ve Destek Vektör Makineleri(Support Vector Machines) gibi iki önemli veri madenciliği tekniği kullanılarak kadın genital bölge kanser teşhislerinin doğruluğunun ve tutarlılığının test edildiği bir model geliştirilmiştir.*

Anahtar Kelimeler: Veri Madenciliği; Destek Vektör Makineleri; laboratuvar testleri; ICD

1. Giriş

Laboratuvar test istem sonuçları hekimlerin tanı koymalarına yardımcı olacak çok önemli ipuçları verebilmektedir. Yapılan bazı çalışmalara göre tanıların %70'inin bu sonuçlardan yararlanılarak konulduğu ileri sürülmektedir [1]. Bu istemler bir hastane bilgi sistemi düşünüldüğünde çok yoğun data birikimine neden olan kaynaklardan en önemlisidir. Bilgi teknolojilerindeki gelişmelerin sonucunda veri madenciliği doğal bir devrim olmuştur. Veri madenciliği, tek boyutlu yararlanılan çok miktardaki datanın yararlı bilgi birikimine (knowledge) dönüştürülmesini ve özellikle öngörülemeyen boyutlarıyla kullanılmasını sağlayan popüler bir çalışma alanıdır. Bu bağlamda, sağlık alanında veri madenciliği çalışmaları için laboratuvar test istemleri ilk çalışılması gereken kaynaklardan birisidir.

Bu çalışmada KTÜ Farabi Hastanesi 2010 yılı laboratuvar istemleri temel veri kaynağı

olmuştur. Çalışma için gerekli olan verileri içeren değişik veritabanı tablolarından bir SQL view oluşturularak, kayıtlar veri ambarına bu view üzerinden çekilmiştir.

2. Gereç ve Yöntem

Veritabanlarında bilgi keşfi süreci genellikle aşağıdaki sırayı izlemektedir:

1. Problemin tanımlanması:
2. Verilerin Hazırlanması: Değişik veri kaynaklarından çeşitli yöntemlerle alınan dataların çalışmanın amacına uygun olarak temizleme, birleştirme, seçme ve dönüştürme gibi işlemlerle hazır hale getirilmesi.
3. Veri Madenciliği: Modelin kurulması ve değerlendirilmesi
4. Modelin Kullanılması
5. Modelin İzlenmesi.

Ön İşlemler

Bu çalışma için tüm veri madenciliği çalışmalarında kullanmak üzere anabilim dalımızca Dönüşüm Aracı Yazılımı(DAY) geliştirilmiştir. Farabi Hastanesi 2010 yılına ait, hastaların kişisel olmayan demografik verilerini içeren 227.228 farklı klinik laboratuvar test istem kaydı DAY kullanılarak çalışma kapsamına alınmıştır. Bu kayıtların 86.588'i yatan, 140640'ı ayaktan hastalara aittir. Çalışma için önemli olduğu düşünülen hasta vizit numarası, vizit tarihi, poliklinik adı ve kodu, cinsiyeti, doğum yeri ve tarihi, son tanı kodu (ICD10), test kodu ve adı, örnek numarası, sonuç ve sonuç birimi alanları, bu çalışma için tasarlanmış Oracle 10gR2 veritabanına aktarılmıştır. Sırasal yapıda olan datalar veri analizi için yatay tabloya dönüştürülmüş, böylece her bir test isteminin tüm özellikleri tek bir tablo satırında toplanmıştır. Böylece WEKA, Statistica gibi veri madenciliği araçlarında veri analizi için en uygun yapı oluşturulmuştur(Bkz. Şekil 1-2).

NO	VIZIT_TAR	POLIKLINIK_ADI	CI	DOGUM_TAF	DOGUM_YE	ICDKOD	TETKIK_KI	TETKIK_ADI	ORNEKNO	SONUC
0167773	27.05.2010	Radyasyon Onkolojis	K	12.12.1944	GÖLKÖY	C50.4	90126044	Fosfor (P)	7232525	3.1
0167897	27.05.2010	Üroloji Polikliniği	K	02.03.1955	SÜRMENE	R30.0	90399044	Trigliserid	7232670	105
0168332	27.05.2010	Üroloji Polikliniği	E	23.01.1982	TRABZON	N23	90025044	Alfa- feto protein (AF	7232675	2.90
0168683	27.05.2010	Cildiye Polikliniği	E	05.05.1989	ÜSKÜDAR	L70.0	90399044	Trigliserid	7232849	127
0170840	31.05.2010	Nöroloji Polikliniği	K	05.09.1957	VAKFIKEBİR	R42	90347044	Serbest T3	7232939	2.99
0170840	31.05.2010	Nöroloji Polikliniği	K	05.09.1957	VAKFIKEBİR	R42	90399044	Trigliserid	7232939	148
0066477	02.03.2010	Göğüs Hastalıkları Pc	E	19.07.1967	TRABZON	A15.0	90576171	Boyalı mikroskopik ir	7232457	NEGATİF

Şekil 1- Hastane Bilgi Sistemi view görünümü

IKLINIK_ADI	CINSİ	DOGUM_TAF	DOGUM_YERI	ICDKOD	ORNEKNO	AC_705321	AC_900131	AC_900210	AC_900250	AC
üs Hastalıkları Polik	E	19.07.1967	TRABZON	A15.0	7232457					
yasyon Onkolojisi F	K	12.12.1944	GÖLKÖY	C50.4	7232525					
oji Polikliniği	K	02.03.1955	SÜRMENE	R30.0	7232670					
oji Polikliniği	E	23.01.1982	TRABZON	N23	7232675				1	
ye Polikliniği	E	05.05.1989	ÜSKÜDAR	L70.0	7232849					
loji Polikliniği	K	05.09.1957	VAKFIKEBİR	R42	7232939					

Şekil 2- DAY ile dönüştürülmüş veri tablosu.

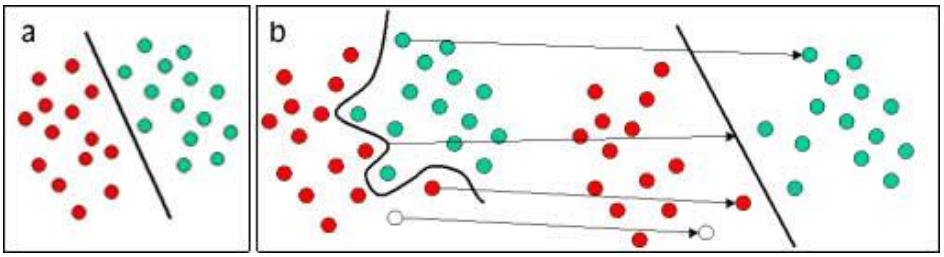
Dönüştürme işlemlerinde birliktelik kuralları analizi için testlerin varlığı 1, yokluğu boş değer olarak işlenmiştir. Destek Vektör Makinaları(DVM , Support Vector Machine-

SVM) için ise aynı yapıda, çalışılmış testler 1 olarak değil; doğrudan sonuçları kaydedilmiştir.

Modelin Kurulması

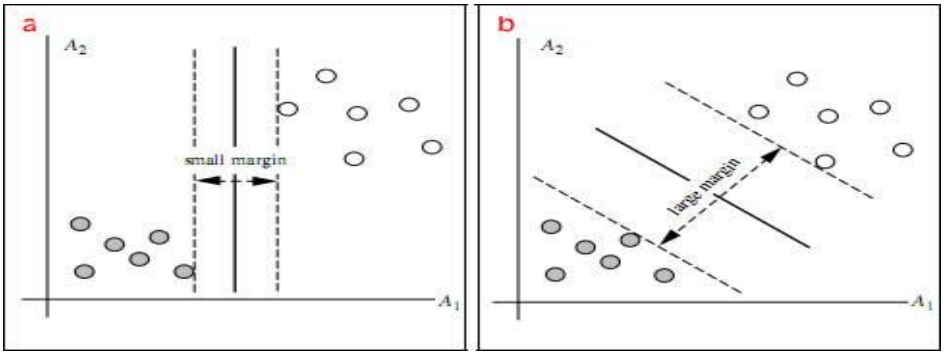
Öncelikle veri tablosu genel kümeleme(clustering) analizine tabi tutulmuş ve analizler sonucunda kanser şüphesi ve tanısı konusunda önemli ipucu sağlayan test sonuçları ve hasta özellikleri erkekler için Total ve Free PSA, kadınlar için ise CEA, AFP, CA 125, CA 15-3, CA 19-9 olarak belirlenmiştir.

Birliktelik kuralları sonucu belirlenen test istemlerinden yararlanılarak hastaların ICD10 kodlarının doğruluğunun sınanması amacı ile doğrusal ya da doğrusal olmayan sınıflandırma problemlerinde oldukça başarılı sonuçlar veren DVM kullanılmasına karar verilmiştir. DVM'nin temel prensibi bir karar düzleminde sınıflar arası optimum karar sınırlarını belirlemektir. Bu sınırlara göre istenen nesnenin hangi sınıfa ait olduğu belirlenebilmektedir.



Şekil 3- DVM Doğrusal(a) ve doğrusal olmayan(b) sınıflandırma[1]

Gerçek yaşamda sınıflama problemleri genellikle doğrusal olmamaktadır(Şekil 3-b). Bu nedenle DVM'de doğrusal olmayan sınıflama problemleri kernel adı verilen matematiksel fonksiyonlar kullanılarak doğrusal ayırımın yapılabilmesi için dönüşüm işlemine tabi tutulur(Şekil 3-b). Bunun sonucu sınıflar arası özellikleri açısından en çok benzeyen, başka bir deyişle en yakın iki nesne arasındaki en uzak mesafeyi bulmayı sağlayan iteratif bir algoritma ile optimum karar doğrusu çizilebilmektedir(Şekil 4-b).



Şekil 4- Optimal Doğru [2]

[1] Statistica Programı yardım dokümanından adapte edilmiştir.

Modelin Çalıştırılması

Çalışma kapsamına CEA, AFP, CA 125, CA 15–3, CA 19–9 testlerinden en az birisi istenmiş olan 6.628 istem dahil edilmiştir. Bunlar içerisinde ön çalışma olarak ICD10 kodu C51-C58 aralığında olan Kadın Genital Bölge Kanseri(KGBK) tanısı konmuş 642 hasta (Y) ile geri kalan hastalar (N) ile işaretlenerek gruplandırılmıştır.

DVM'de kullanılacak sürekli değişkenlerin seçimi için CA 125 testi ile birlikte en çok istenen testleri ya da testi belirlemek amacıyla birliktelik kuralı çıkarımı(Association Rule Mining) yapılmıştır. CA 125'in seçilme nedeni KGBK tanısı ile ilgili önemli veri içermesidir. Çıkarılan birliktelik kuralı aşağıdaki gibidir;

CA 125=1 1401 ==> CA 19- 9=1 1172 conf (0.84), supp (0,23)

Birliktelik kuralları WEKA kullanılarak hesaplanmıştır. Buna göre çalışmada CEA, AFP, CA 125, CA 15-3, CA 19-9 testlerinden herhangi birinin istendiği toplam 5.105 ayaktan hasta kaydı kullanılmış, bunlar içinde CA 125 testi 1.401 kez istenmiş ve bu istemlerin 1.172'sinde aynı zamanda CA 19- 9 da istenmiştir (conf = 1172/1401 = 0.84).

Birliktelik kuralına göre DVM'ye kesiksiz değişken olarak CA 125 ve CA 19-9 sonuçları ve kategorik veriler olarak poliklinik kodu, cinsiyet, yaş ve ICD10 tanı kodları verilmiştir. DVM sonuçları aşağıda listelenmiştir.

Dataset CANCER_DATA:
Dependent: KADINGK
Independents: CA125_90081044, CA19-9_90083044, POLIKLINIK_KODU, CINSIYET, AGE, ICD10
Sample size = 867 (Train), 304 (Test), 1171 (Overall)

Support Vector machine results:
SVM type: Classification type 1 (capacity=10,000)
Kernel type: Radial Basis Function (gamma=0,167)
Number of support vectors = 320 (0 bounded)
Support vectors per class: 208 (N), 112 (Y)

Class. accuracy (%) = 100,000(Train), 99,342(Test), 99,829(Overall)

Sonuçlara göre toplam 1.171 test istemi içerisinde 867'si DVM eğitimi için kullanılmış, 304 istem ise DVM'nin testi için kullanılmıştır. Eğitim başarısı %100, test başarısı %99,3 olduğu görülmektedir. Test setinde yer alan 304 hastanın 166'sı KGBK tanısı (N) ve 138'i ise(Y) şeklinde iken DVM'nin sınıflamasına göre ise 168 (N), 136(Y) şeklinde oluşmuştur. Buna göre 2 hasta KGMK sonucu (Y) iken DVM'ye göre (N) olmuştur. Bu iki test istemi incelendiğinde CA 125 ve CA 19-9 test sonuçlarının normal

---

[2] Han J., Sayfa 337 den alınmıştır.

aralıkta olduğu görülmektedir. Dikkat çeken diğer özellik ise bir hastanın kardiyoloji polikliniğinde kayıtlı ve 30 yaşında olduğudur. DVM'in sınıflandırmayı başarılı yaptığı, yanlış sınıflandırmaların ise araştırılması gerektiği düşünülmektedir. Bu araştırmanın amacına uygun bir durumdur.

3. Tartışma ve Sonuç

Bu çalışmada "Çıkarılan birliktelik kurallarına göre çalışma kapsamına alınan özelliklerin KGBK tanılarının tutarlılığını DVM ile sınamak mümkün müdür?" sorusuna cevap aranmıştır. Elde edilen sonuçlar DVM'nin buna benzer sınıflama ve tahmin için ümit verici bir teknik olduğunu göstermektedir. Laboratuvar testleri ve tanılar arasındaki ilişkileri analiz eden birçok veri madenciliği çalışmaları görülmektedir. Birlikte istenen testlerden yararlanılarak yapılan çalışmada, test rutinlerinin birlikteliği araştırılmış ve yeni rutin grupları önerilmiştir[2]. Başka bir çalışmada total kolestrol, LDL, trigliserit, HDL ve VLDL gibi biyokimya test parametreleri analiz edilerek hiperlipidemi hastalığının teşhisi için karar destek sistemi geliştirilmiştir[3]. Göğüs ağrısının daha hızlı ve doğru bir şekilde sınıflandırılması için yapılan çalışmada, birliktelik kuralları ve sınıflandırma veri madenciliği tekniklerini kullanan hibrit bir model geliştirilmiştir[4]. Glukoz, total kolesterol, kreatinin gibi biyokimya testleri ve alkol ve sigara kullanımı, aile öyküsü, kardiyak veriler ve anjiyografi sonuçları gibi verilerden yararlanılarak koroner arter hastalıklarının teşhisi için Bayesian uzman sistem önerilmiştir[5]. Sağlık alanında yapılan veri madenciliği çalışmaları dikkate alındığında bu alanında çalışmaya değer yığınla veri olduğu görülmektedir.

Hastanelerde ICD kodları özellikle SGK ödemeleri için zorunlu olduğundan girilmektedir. Bu kodlar hızlı çalışma temposu ve belirsizlikler nedeniyle hatalı girilebilmektedir. Bu tür çalışmalar girilen tanıların tutarlı ve doğru olması için yararlı sonuçlar verebilir. Gerek ödeyici kurumlar ve gerekse hizmet veren kurumlarda yapılacak veri madenciliği çalışmaları için geniş bir uygulama alanı olduğu gözükmektedir.

Çok kısıtlı veri seti kullanılarak alınan bu sonuçlar daha çok veri seti kullanılarak, kapsamlı ve uzun süreli araştırmalarla çok daha iyi noktalara getirilebilir.

4. Kaynakça

[1] Göktaş P. Laboratuvar testleri fazla mı isteniyor ?. http://www.saglikaktuel.com/yazi/laboratuvar-testleri-fazla-mi-isteniyor--6306.htm, Son erişim: 29.09.2011.

[2] Santangelo J, Rogers P, Buskirk J, Mekhjian HS, Liu J, Kamal J. Using Data Mining Tools to Discover Novel Clinical Laboratory Test Batteries, *AMIA* 2007: New York, USA.

[3] Dogan S and Turkoglu I. Diagnosing Hyperlipidemia Using Association Rules, *Mathematical and Computational Applications* 2008: Vol. 13, No. 3, pp. 193-202.

[4] Ha SH and Joo SH. A Hybrid Data Mining Method for the Medical Classification of Chest Pain, *International Journal of Computer and Information Engineering* 2010: 4:1.

[5] Chu CM, Chien WC, Lai CH, Bludau HB, Tschai HJ, Pai L, Hsieh SM, Chu NF, Klar A, Haux R and Wetter T. A Bayesian Expert System for Clinical Detecting Coronary Artery Disease, *J Med Sci* 2009:

29(4):187-194.
[6] Agrawal R, Imielinski T, Swami A, Mining Association Rules Between Sets of Items in Large Databases, *ACM SIGMOD Conference* 1993, Washington, USA.
[7] E. Delibas, Birliktelik Analizi İle Reçeteli İlaç Satişlari Üzerinde Bir Uygulama. [Yüksek Lisans Tezi], Cumhuriyet Uni., Sivas, Türkiye, 2010.
[8] Han J, Kamber M. *Data Mining: Concepts and Techniques*. 2nd ed. New York: Morgan Kaufmann Publishers, 2006.
[9] Statistica yazılımı yardım dokümanları.

5. Sorumlu Yazarın Adresi

Kemal TURHAN
kturhan.tr@gmail.com