Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi

Machine Learning Based Performance Development for Diagnosis of Breast Cancer

Burcu Bektaş[1], Sebahattin Babur[2]
[1]Bilgisayar Teknolojileri Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye
burcu.bektas@istanbul.edu.tr
[2]Tıbbi Görüntüleme Teknikleri Bölümü, İstanbul Gedik Üniversitesi, İstanbul, Türkiye
sebahattin.babur@gedik.edu.tr

***Özetçe*—Meme kanseri, sıklıkla kadınlar arasında görülen ve meme hücrelerinde başlayan bir kanser türüdür. Erken teşhis ve doğru tedavi meme kanseri hastalarının hayatta kalma oranını arttırabilme açısından son derece önemlidir. Mikrodizi teknolojisi ile hastalığın genetik faktörlerinin belirlenmesi teşhis ve tedavi sürecine önemli katkı sağlayabilir. Bu çalışmada, meme kanserinin teşhisi için birçok makine öğrenmesi algoritmaları kullanılmıştır ve bu algoritmaların birbirlerine göre sınıflandırma performansları karşılaştırılmıştır. Bunun yanı sıra öznitelik seçme yöntemleri ile meme kanserinde etkin genler belirlenip yapılan çalışma sonucunda 139 öznitelik ile % 90,72 başarı elde edilmiştir.**

Anahtar Kelimeler — makine öğrenmesi; meme kanseri teşhisi; mikrodizi; öznitelik seçme.

***Abstract*— Breast cancer is prevalent among women and develops from breast tissue. Early diagnosis and accurate treatment is vital to increase the rate of survival. Identification of genetic factors with microarray technology can make significant contributions to diagnosis and treatment process. In this study, several machine learning algorithms are used for Diagnosis of Breast Cancer and their classification performances are compared with each other. In addition, the active genes in breast cancer are identified by attribute selection methods and the conducted study show success rate 90,72 % with 139 feature.**

Keywords — machine learning; breast cancer diagnosis; microarray; feature selection.

I. GİRİŞ

Meme kanseri gün geçtikçe ağırlığı artan ve kadınlarda en sık rastlanan bir sağlık sorunudur. Batı ülkelerinde her 8-9 kadından birinin yaşamı boyunca meme kanserine yakalandığı görülmektedir. Bu oran %10'dan daha fazladır. Sık görülmesi, erken evrelerde tedavi edilebilir olması, günümüz koşullarında tanınmasının olanaklı olması, meme kanserinin önemini artırmaktadır [1]. Meme kanseri kompleks ve tek sebebe bağlı olmayan genetik bir hastalıktır. Kalıtsal, ailesel ve sporadik (tesadüfi / rastlantısal) olarak 3 grupta incelenmektedir. Ailede bir veya daha fazla kişide meme kanseri görülmesi halinde buna ailesel meme kanseri denilmektedir. Meme kanseri ile birebir ilişkisi tespit edilmiş olan BRCA1 ve BRCA2 genlerinde mutasyon olması, ailedeki kanserlerin erken yaşta görülmesi, ailede erkeklerde de meme kanseri görülmesi hastalığın kalıtsal olduğunu işaret etmektedir. Kalıtsal meme kanseri durumundaki kişilerin yaşamları boyunca meme kanserine yakalanma riskleri %85 gibi çok yüksek oranlardadır. Bu iki grubun dışında kalanlar ise ailesel ve kalıtımsal bir bağlantısı olmayan sporadik meme kanserleridir [2, 3].

Meme kanserinin temelinde yatan genetik faktörlerin belirlenmesi bu hastalığın teşhis ve tedavisine çok önemli katkılar sağlayabilir. Son yıllarda DNA analizlerinde çok büyük gelişmeler görülmektedir. Bunlardan biri aynı anda on binlerce genin birbirleriyle olan etkileşiminin ölçülebilmesine imkân sağlayan gen çipleri, diğer bir adıyla mikrodizi (microarray) teknolojisidir. Mikrodizi teknolojisi sayesinde hastalıkların temelinde yatan genetik faktörler belirlenerek, hastalığın erken teşhisi gerçekleştirilebilir. Mikrodizi ile on binlerce gen ifade bilgisi aynı anda ölçülebilmektedir. Ancak gen ifade verileri az örneklem ve çoğu gürültü olarak adlandırılan ilgisiz on binlerce gen bilgisini içermektedir. Ayrıca öznitelik sayısının çok olması, veri boyutunu arttırdığı gibi analiz ve sınıflandırma problemini de beraberinde getirmektedir. Bu durum, mikro diziler ile çalışılabilmesi için öncelikle boyut indirgeme ve gen seçimi gibi işlemlerin yapılmasını zorunlu kılmaktadır. Bu sebeple gen analizlerinde nitelik seçme işlemi kritik bir öneme sahiptir [4 - 9].

M. Bilen ve arkadaşları yaptığı çalışmada mikrodizi verileri üzerinde boyut azaltma işlemi uygulamıştır. Elde edilen yeni veri kümesinin Genetik Algoritma ve Yapay Sinir Ağı ile sınıflandırılma neticesinde doğruluk oranı %75 olarak bulunmuştur [8].

978-1-5090-2386-8/16/$31.00 ©2016 IEEE

O. Yıldız ve arkadaşlarının yaptığı çalışmada nitelik seçme yöntemleri ile meme kanserinde rol alan etkin genler belirlenmiş ve DVM sınıflandırıcısı kullanılarak % 82,69 doğruluk oranı elde edilmiştir [6].

K. Polat ve arkadaşlarının yaptığı çalışmada en küçük kareler destek vektör makinesi sınıflandırıcı algoritması kullanılarak 98.53% doğruluk oranı elde edilmiştir. [10].

II. MATERYAL VE METOD

A. *Veri Seti*

Bu çalışmada *Kent Ridge 2* mikrodizi veri seti kullanılmıştır [11]. 97 meme kanseri hastasına ait 24482 öznitelik bulunmaktadır. Bu hastaların 46'sında metastaz (kötü prognoz) görülürken kalan 51 kişide metastaz görülmemektedir (iyi prognoz).

B. *Öznitelik Seçme*

Nitelik seçme, ilgisiz niteliklerin atılması ya da verinin gürültüden temizlenmesi işlemidir. Bu işlem sınıflandırma başarısını ve performansını doğrudan etkiler. Veriler içerisindeki etkin genleri seçme işlemi WEKA programı yardımıyla gerçekleştirilmiştir. Öznitelik seçim yöntemi olarak Correlation-based Feature Subset Selection (CfsSubsetEval), arama metodu olarak bestFirst tercih edilmiştir. Bu seçimler neticesinde yüksek öznitelik vektörlerine sahip yeni veri seti oluşturulup 139 adet etkin gen belirlenmiştir.

C. *Sınıflandırma*

LibSVM, Destek Vektör Makinesi (SVM) için geliştirilmiş ve en yaygın kullanılan SVM kütüphanelerinden biridir. İki sınıflı verilerin tahmininde güçlü bir makine öğrenme tekniği olan geleneksel SVM' nin aksine çok sınıflı verilerin kullanılmasına olanak sağlamaktadır. LibSVM çalışması iki aşamadan oluşmaktadır: İlk aşamada model oluşturmak için veri seti eğitilir, ikinci aşamada oluşturulan model, test veri setine ait bilgilerin tahmini için kullanılır [12].

Rastgele Orman (Random Forest) karar ağacı, veri setinde en iyi niteliklerden seçilen düğümleri dallara ayırmak yerine, her bir düğümde rastgele alınan niteliklerin en iyisini seçerek tüm düğümleri dallara ayırır. Her veri kümesi asıl veri setinden yer değiştirmeli olarak üretilir. Rastgele özellik seçimi kullanılarak ağaçlar geliştirilir ve budama işlemi yoktur. Rastgele orman algoritmasının diğer algoritmalara göre daha hızlı ve doğru olmasının sebebi bu yöntemdir [13].

K-star algoritması iki özelliği birbirine bağlayan en kısa uzaklık olarak Kolmogorov mesafesini dikkate almaktadır. Bu durumda K-star uzaklığı, iki özellik arasındaki tüm olası dönüşümlerin toplamı olmaktadır [14]. Olasılık fonksiyonu P^*, t dönüşümleri a özelliğinden b özelliğine olan tüm yolların olasılığı olarak,

$$P^*(b/a) = \sum_{t\in P:t(a)=b} \overrightarrow{p(t)} \quad \text{belirlenir.} \qquad (1)$$

Bu durumda K-star (K^*)fonksiyonu,

$$K^*(b/a) = -\log_2 P^*(b/a) \quad \text{şeklinde ifade edilir.} \qquad (2)$$

Seçimli Algılayıcı (Voted Perceptron), Rosenblatt ve Frank tarafından sinir ağı tabanlı oylama yöntemidir. Biyolojik sinir sisteminin çalışma şekli simüle edilerek tasarlanmış, nöronlar içeren ve bu nöronların çeşitli şekillerde birbirlerine bağlanarak oluşturduğu bir öğrenme sistemidir. Algoritma doğrusal olarak ayrılabilen verilere dayanır. DVM ile karşılaştırıldığında uygulama açısından basit ve hesaplama süresi verimli olmaktadır. Ayrıca algoritma çekirdek fonksiyonları kullanılarak yüksek boyutlu uzaylarda kullanılabilir [15].

D. *Performans Değerlendirme*

Alıcı İşletim Karakteristiği (Receiver Operating Characteristics - ROC), sınıflandırıcı performansını test etmek için biyoinformatikte sıklıkla kullanılan bir yöntemdir [16]. Bir veri kümesinde, dört muhtemel sonuç vardır; pozitif örnek doğru sınıflandırıldığında Doğru Pozitif (DP), yanlış sınıflandırıldığında Yanlış Negatif (YN) olarak sayılırken, Negatif örnek doğru sınıflandırıldığında Doğru Negatif (DN) ve yanlış sınıflandırıldığında Yanlış Pozitif (YP) olarak sayılır. Doğru Pozitif Oranı ve Yanlış Pozitif Oranı Eşitlik (3) ve (4) ile hesaplanabilir. Buradan elde edilen değerlere göre hata matrisi Şekil 1'de görüldüğü gibi olacaktır [9].

$$\text{Doğru Pozitif Oranı} = \frac{DP}{DP+YN} \qquad (3)$$

$$\text{Yanlış Pozitif Oranı} = \frac{YP}{YP+DN} \qquad (4)$$

Tahmin edilen sınıf	Gerçek sınıf: Pozitif	Gerçek sınıf: Negatif
Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

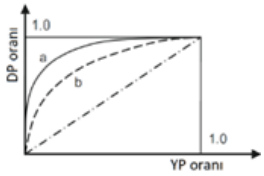
Şekil 1. Hata matrisi (Confusion Matrix)

Eşitlik (5) ile sınıflandırma doğruluğu hesaplanırken, eşitlik (6) ve (7) kullanılarak, ROC eğrisi Şekil 2'de görüldüğü gibi elde edilebilir. X ekseni YP oranı, y ekseni DP oranı olarak çizildiğinde, diyagonal eğrinin üstünde ve sol üst köşeye yaklaşan ROC eğrisine sahip sınıflandırıcı performansı iyi kabul edilir. Şekil 2'de a sınıflandırıcısının b sınıflandırıcısına göre daha başarılı olduğu söylenebilir [9].

$$\text{Sınıflandırma Doğruluğu} = \frac{DP+DN}{DP+YP+DN+YN} \qquad (5)$$

$$\text{Duyarlılık (Sensitivity)} = DP\ Oranı \qquad (6)$$

$$\text{Kesinlik (Specificity)} = 1 - YP\ Oranı \qquad (7)$$



Şekil 2. ROC eğrisi (ROC Curve)

III. DENEYSEL SONUÇLAR

Göğüs kanseri hastalığının teşhisi amacıyla elde edilen veri setleri üzerinde yeni öznitelik vektörleri oluşturularak 10-kat çapraz doğrulama test tekniği gereğince sınıflandırma yapılmıştır. Bu test tekniğine göre veri seti 10 eşit kümeye bölünür. 9 küme eğitim için 1 küme test için kullanılır. Böylece test sonunda 10 adet performans metriği elde edilir. Elde edilen her bir metriğin aritmetik ortalaması alınır. Bu çalışmada performans değerlendirmesi için sınıf doğruluğu, duyarlılığı ve kesinliği istatistik metrik değerleri kullanılmıştır [17].

Weka ortamında gerçekleştirilen çalışmalarda oluşturulan yeni vektörler; DVM, k-Yıldız, Rastgele Orman Algoritması ve Seçimli Algılayıcı Sinir Ağı algoritmalarına göre sınıflandırılmıştır. Yapılan deneyler sonucunda, Tablo 1 ve 2'de görüldüğü gibi Rastgele Orman yöntemi 139 öznitelik ile seçilmiş yeni veri seti üzerinde % 90,72 sınıf doğruluğu, % 91,30 duyarlık, % 90,19 kesinlik değerleri ile en iyi sonucu vermiştir.

Bu sonuçlar Rastgele Orman Algoritmasının, diğer sınıflandırma algoritmalarına göre bu problem için daha iyi performans sergilediğini göstermektedir.

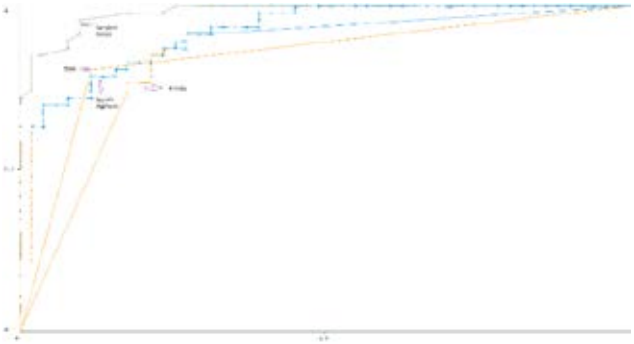
Sınıflandırıcı	Performans Sonuçları (24482 öznitelik - 97 hasta için)				
	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	MCC	AUC
DVM (Lineer)	67,01	60,89	72,54	0,337	0,667
K-Yıldız	47,42	100	0	0	0,5
Rastgele Orman Algoritması	61,85	54,34	68,62	0,232	0,7
Seçimli Algılayıcı Sinir Ağı Yöntemi	59,79	56,52	62,74	0,193	0,656

Tablo 1. 24482 öznitelik ve 97 hasta için sınıflandırıcı performans sonuçları

Şekil 3'de görülen ROC eğrisine göre Rastgele Orman algoritması en iyi doğruluk sonucunu verirken onu sırasıyla DVM, Seçimli Algılayıcı ve K-Yıldız algoritması takip etmektedir.

Sınıflandırıcı	Performans Sonuçları (139 öznitelik - 97 hasta için)				
	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	MCC	AUC
DVM (Lineer)	84,53	80,43	88,23	0,69	0,843
K-Yıldız	80,41	60,86	98,03	0,643	0,891
Rastgele Orman Algoritması	**90,72**	**91,30**	**90,19**	**0,814**	**0,98**
Seçimli Algılayıcı Sinir Ağı Yöntemi	81,44	86,95	76,47	0,635	0,848

Tablo 2. 139 öznitelik ve 97 hasta için sınıflandırıcı performans sonuçları



Şekil 3. Sınıflandırıcı performanslarının ROC eğrisinde gösterimi

IV. SONUÇ

Göğüs kanseri hastalığının yüksek doğruluk oranları ile kestirimi, kanserin teşhisi ve tedavisi açısından önemli bir eşiktir. Bu çalışma *Kent Ridge 2* mikrodizi veri seti kullanılarak, oluşturulan 139 öznitelik vektörü, DVM, k-Yıldız, Rastgele Orman, Seçimli Algılayıcı Sinir Ağı yöntemi algoritmalarına göre sınıflandırılmıştır. Problemin çözümünde 139 yeni öznitelik ile Rastgele Orman Algoritmasının, 24482 özniteliğe göre daha iyi performans sergilediği görülmüştür. İleriye yönelik yapılacak çalışmalar da mikrodizi veri setinden yeni öznitelik seçim yöntemleri geliştirerek, elde edilecek olan özniteliklere göre yeni sınıflandırma yöntemlerinin problemin çözümünde kullanılması planlanmaktadır.

KAYNAKÇA

[1] Aydıntuğ S., "Meme Kanserinde Erken Tanı", *Sürekli Tıp Eğitimi Dergisi*, Cilt. 13, Sayı. 6, ss 228, 2014.

[2] Öztemur Y, Aydos A, Gür-Dedeoğlu B., "Meme kanseri mikrodizin verilerinin biyoinformatik yöntemler ile bir araya getirilmesi - Meta-analiz yaklaşımları", *Turk Hijyen Deneysel Biyoloji Dergisi*, 72(2), 155-162, 2015.

[3] www.cancer.org [S.E.T: 21.7.2016]

[4] E. Segal, H. Wang ve D. Koller, "Discovering molecular pathways from protein interaction and gene expression data", *Bioinformatics*, Cilt 19, 264-272, 2003.

[5] C.P. Lee ve Y. Leu, "A novel hybrid feature selection method for microarray data analysis", *Applied Soft Computing*, Cilt 11, 208–213, 2011.

[6] Yıldız, O., Tez, M., Bilge, H. Ş., Akcayol, M. A., & Güler, İ., Gene selection for breast cancer", *In 2012 20th Signal Processing*

and Communications Applications Conference (SIU) (pp. 1-4). IEEE,2012.

[7] Y. Peng, Z. Wu ve J. Jiang, "A novel feature selection approach for biomedical data classification", *Journal of Biomedical Informatics*, Cilt 43, 15-23, 2010.

[8] Bilen, Mehmet, Ali Hakan Işik, and Tuncay Yiğit. "A hybrid Artificial Neural Network-Genetic Algorithm approach for classification of microarray data", *2015 23nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2015.

[9] Yıldız, O., Tez, M., Bilge, H. Ş., Akcayol, M. A., & Güler, İ., " Meme Kanseri Sınıflandırması İçin Veri Füzyonu Ve Genetik Algoritma Tabanlı Gen Seçimi", *Journal of the Faculty of Engineering & Architecture of Gazi University*, 27(3), 2012.

[10] Polat, Kemal, and Salih Güneş. "Breast cancer diagnosis using least square support vector machine", *Digital Signal Processing* 17.4 (2007): 694-701.

[11] http://mldata.org/repository/data/viewslug/breast-cancer-kent-ridge-2/ [S.E.T: 21.07.2016].

[12] Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

[13] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi: 10.1023/A:1010933404324.

[14] Clear, J.G., Trigg L.E., 1995. K*: An instance-based learner using an entropic distance measure. *Proceedings of the 12th International Conference on Machine Leaning*, Tahoe City, California, USA, pp.108-114.

[15] Y. Freund, R. E. Schapire: Large margin classification using the perceptron algorithm. In: *11th Annual Conference on Computational Learning Theory*, New York, NY, 209-217, 1998.

[16] Lasko, T.,A., Bhagwat, J., G., Zou, K., H. Ve Ohno-Machado, L., "The Use Of Receiver Operating Characteristic Curves In Biomedical Informatics", *Journal of Biomedical Informatics*, Cilt 38, 404–415, 2005.

[17] Witten I H, Frank E. (2005) Data Mining: *Practical Machine Learning Tools and Techniques*. 2nd ed. CA: Morgan Kaufmann, San Francisco.