Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences
International Indexed & Refereed
ISSN: 2667-6702

EN YAKIN K KOMŞU VE DESTEK VEKTÖR MAKİNELERİ YÖNTEMLERİ KULLANILARAK MEME KANSERİ TAHMİNİNE YÖNELİK ÖZELLİKLER BULMA

FINDING FEATURES FOR THE PREDICTION OF BREAST CANCER USING THE K NEAREST NEIGHBORHOOD AND SUPPORT MACHINES METHODS

Serdar GÜNDOĞDU

Öğr. Gör. Dr., Dokuz Eylül Üniversitesi, Bergama MYO, Bilgisayar Teknolojileri Bölümü

Geliş Tarihi / Received: 18.05.2020
Kabul Tarihi / Accepted: 14.07.2020
Araştırma Makalesi/Research Article
DOI: 10.38065/euroasiaorg.151

ÖZET

Kadınlar arasında en sık görülen kanser türü olan meme kanseri, her yıl milyonlarca kadını etkilemektedir. Bu çalışmanın amacı iki temel sınıflandırıcı yöntemi kullanılarak meme kanseri tahminine yönelik özellikler bulmak ve sınıflandırıcı performanslarını karşılaştırmaktır. Çalışmada, 52 sağlıklı ve 64 hasta kişiye ait yaş, vücut kitle indeksi (VKİ), glikoz, resistin, insülin, insülin direnci için homeostaz model değerlendirmesi (HOMA-IR), monosit kemo-çekici protein-1 (MCP1), leptin hormonu ve adiponektin hormonu verilerini içeren hazır veri setinden yararlanılmıştır. Eğitim ve test verilerini seçerken 10 katlamalı çapraz doğrulama (10 fold) yöntemi kullanılmıştır. Sınıflandırıcılardan DVM'nin sınıflandırma performansı, KNN'den daha iyi olmuştur. 5 öznitelikli veri setinin, %85.3 sınıflandırma doğruluğu ve % 89.1 özgüllük (kanser olanları tespit etme oranı) değerleri ile en iyi sınıflandırma performansı sergilediği görülmüştür. Meme kanser tahmini için en uygun özelliklerin yaş, VKİ, resistin, glikoz ve adiponektin parametreleri olduğu tespit edilmiştir.

Anahtar Kelimeler: Meme kanser teşhis, Çapraz doğrulama, Öznitelik seçimi, Sınıflandırma KNN, DVM

ABSTRACT

Breast cancer, the most common type of cancer among women, affects millions of women every year. The aim of this study is to find properties for estimating breast cancer using two basic classifier methods and to compare the classifier performances. In the study, a ready data set including age, body mass index (BMI), glucose, resistin, insulin, homeostasis model assessment for insulin resistance (HOMA-IR), monocyte chemoattractant protein 1 (MCP1), leptin hormone and adiponectin hormone were used. 10 fold cross verification (10 fold) method was used while selecting the training and test data. Classification performance of SVM from classifiers was better than KNN. It was seen that the data set with 5 clinical features showed the best classification performance with 85.3% classification accuracy and 89.1% specificity (the rate of detecting those with cancer) values. It has been determined that the most suitable features for breast cancer prediction are age, BMI, resistin, glucose and adiponectin parameters.

Keywords: Breast cancer diagnosis, Cross-validation, Feature selection, Classification, KNN, SVM

1. GİRİŞ

Kadınlar arasında en sık görülen kanser türü olan meme kanseri, her yıl 2,1 milyon kadını etkilemekte ve kadınlar arasında kansere bağlı en fazla ölüme neden olmaktadır. 2018'de 627.000 kadının meme kanserinden öldüğü tahmin edilmekte olup; bu kadınlar arasındaki tüm kanser ölümlerinin yaklaşık % 15'i oluşturmaktadır. Hastalık oranları, küresel olarak hemen hemen her bölgede artmaktadır [1]. Bazı risklere rağmen, önleme ile ölüm sayıları azaltılabilir. Erken teşhis ve doğru tanı koyma, meme kanseri tedavisinin en önemli adımlarındandır. Otomatik teşhis sistemleri kullanılarak, teşhis sırasında olası doktor hataları ortadan kaldırılabilir ve tıbbi istatistikler daha kısa sürede daha ayrıntılı olarak analiz edilebilir. Bu görevi çözmek için farklı veri madenciliği teknikleri uygulanmış ve performansları karşılaştırılarak değerlendirilmiştir. Bu teknikler arasında Bayes Ağı, Rastgele

Orman, Lojistik Regresyon, Radyal Temelli Fonksiyon Ağları, Karar Ağaçları, Destek Vektör Makinesi ve Çok Katmanlı Algılayıcı bulunmaktadır [2].

KNN (K En Yakın Komşular) sınıflandırması, ilk 10 veri madenciliği algoritmasından birisidir [3]. KNN etkili bir tembel öğrenme algoritmasıdır ve gerçek uygulamalarda başarıyla geliştirilmiştir. Klasik kNN yöntemi önce bir test numunesi için en yakın k eğitim örneklerini seçer ve daha sonra en yakın k eğitim örnekleri arasında ana sınıfı içeren test örneğini tahmin eder [4]. Basitlik, kolay anlaşılır ve nispeten yüksek kNN performansı nedeniyle meme kanserinin varlığını tahmin etmede kullanılmıştır [5].

Makine öğreniminde DVM'ler, sınıflandırma ve regresyon analizi için kullanılan denetimli öğrenme modelleridir. DVM'ye atıfta bulunurken, genellikle doğrusal DVM değil, tanıtılan çekirdek yöntemleri anlamına gelir. DVM eğitim algoritması, bir kategoriye veya diğerine yeni örnekler atayan bir model oluşturur ve bu da olasılık dışı olmayan bir ikili doğrusal sınıflandırıcı yapar. DVM modeli, örneklerin uzaydaki noktalar olarak temsil edilmesidir, böylece ayrı kategorilerin örnekleri mümkün olduğunca geniş bir açıklığa bölünür [6].

Patricio vd. (2018), son zamanlarda obezite ile ilişkili meme kanseri profillerinde bir deregülasyonu doğruladıklarından dolayı [7], meme kanserinin varlığını tahmin etmek için rutin kan analizleri, özellikle glikoz, insülin, HOMA-IR, leptin, adiponektin, resistin, MCP1, yaş ve vücut kitle indeksi parametrelerinin iyi bir aday seti olduklarına inanmaktadırlar [8].

Bu çalışmanın amacı, iki farklı temel sınıflandırıcının meme kanser tahmininde performanslarını karşılaştırmak ve algoritma girişlerine uygulanan farklı öznitelik seçimlerin tahminlere etkisini araştırmaktır.

2. MATERYAL VE METOT

Meme kanseri, meme hücrelerinden kaynaklanan kötücül bir tümördür. Genetik yapı, yaşlanma, aile öyküsü, çocuk sahibi olmama, adet dönemleri, obezite gibi bazı risk faktörler, meme kanseri geliştirme olasılığını artırdığı bilinmektedir. Bu çalışmada, UCI kütüphanesinden **[9]** elde edilen Glikoz, Resistin, yaş, insülün direnci için Homeostaz Model Değerlendirmesi (HOMA-IR), Vücut Kitle İndeksi (VKİ), İnsülin, Leptin hormonu, monosit kemo-çekici protein 1 (MCP-1) ve Adiponektin hormonu gibi 9 farklı öznitelik ve 116 örnekten oluşan meme kanseri Coimbra veri seti kullanılmıştır. Özniteliklerle ilgili istatiksel bilgiler Tablo 1'de gösterilmiştir. Veri seti, 52 sağlıklı ve 64 hastalıklı olmak üzere toplam 116 kişinin örneklerinden oluşmaktadır.

Tablo 1. Özniteliklerle ilgili istatiksel bilgiler

No	Öznitelikler	Minimum	Maximum	Ortalama	St. Sapma
Ö1	Glukoz (mg/dL)	60.00	201.00	97.79	22.53
Ö2	Resistin (ng/mL)	3.21	82.10	14.73	12.39
Ö3	Yaş (yıl)	24.00	89.00	57.30	16.11
Ö4	VKİ (kg/m2)	18.37	38.58	27.58	5.02
Ö5	HOMA-IR	0.47	25.05	2.69	3.64
Ö6	Leptin (ng/mL)	4.31	90.28	26.62	19.18
Ö7	İnsulin (µU/mL)	2.43	58.46	10.01	10.07
Ö8	Adiponectin (µg/mL)	1.66	38.04	10.18	6.84
Ö9	MCP1 (pg/dL)	45.84	1698.44	534.65	345.91

Model geçerliliğini doğrulamak için 10 katlamalı çapraz doğrulama (10 fold) ve temel sınıflandırıcılar olarak K en yakın komşu (KNN) ve destek vektör makineleri (DVM) yöntemi kullanılmıştır.

Modellerin performansını çıkarmak için karışıklık matrisinden (confusion matrix) yararlanılmıştır. Karışıklık matrisi, sınıflandırma tahminin doğruluğunu değerlendirmek için kullanılan önemli bir ölçümdür. Matrisin Gerçek Negatif (TN), Gerçek Pozitif (TP), Yanlış Negatif (FN) ve Yanlış Pozitif (FP) olarak adlandırılan dört elemanı Tablo 2'de gösterilmiştir.

Tablo 2.İkili sınıflandırma için karışıklık matrisi

KNN Karışıklık Matrisi				
		Tahmin Sınıfı		
		Sağlıklı	Hasta	Toplam
Gerçek Sınıfı	Sağlıklı	**TP**	FP	TP+FP
	Hasta	FN	**TN**	FN+TN
	Toplam	TP+FN	FP+TN	

Sınıflandırıcı modellerinin performanslarını değerlendirmek karışıklık matrisinin yardımıyla doğruluk, duyarlılık, özgüllük, hassasiyet ve F1 skorları (hassasiyet ve duyarlılığın harmonik ortalaması) kullanılmıştır.

3. BULGULAR

Meme kanser tahmini için yapılan sınıflandırma işleminin ilk basamağında, 9 öznitelik girişli KNN ve DVM yöntemleri denenmiştir. KNN için komşu sayısı 10, minkowski mesafe ölçülü kübik tip ve DVM için ise çekirdek fonksiyonu gaussian, skalası 3 olan Medium Gaussian tip eğitim kullanılmıştır. İkinci basamakta ise ilkinde 9 girişli olarak daha başarılı olan DVM tekniğine ayrı ayrı 4 öznitelikli (Glikoz, Resistin, yaş ve VKİ) ve 5 öznitelikli (Glikoz, Resistin, yaş, VKİ ve Adiponektin) girişler uygulanmıştır. Sınıflandırma işlemlerinde farklı öznitelik ve sınıflandırıcı algoritmalardan elde edilen karışıklık matrisleri Tablo 3'de gösterilmiştir.

Tablo 3. Öznitelik ve sınıflandırıcı algoritma seçimlerine karışıklık matrisleri

KNN Karışıklık Matrisi				
		Tahmin Sınıfı		
		1	2	Toplam
Gerçek	1	**42**	10	52
	2	21	**43**	64
	Toplam	63	53	**116**

a) 9 öznitelikli KNN

DVM Karışıklık Matrisi				
		Tahmin Sınıfı		
		1	2	Toplam
Gerçek	1	**40**	12	52
	2	15	**49**	64
	Toplam	55	61	**116**

b) 9 öznitelikli DVM

DVM Karışıklık Matrisi				
		Tahmin Sınıfı		
		1	2	Toplam
Gerçek	1	**43**	9	52
	2	12	**52**	64
	Toplam	55	61	**116**

c) 4 öznitelikli DVM

DVM Karışıklık Matrisi				
		Tahmin Sınıfı		
		1	2	Toplam
Gerçek	1	**41**	11	52
	2	7	**57**	64
	Toplam	48	68	**116**

e) 5 öznitelikli DVM

DVM* Karışıklık Matrisi				
		Tahmin Sınıfı		
		1	2	Toplam
Gerçek	1	**42**	10	52
	2	7	**57**	64
	Toplam	49	67	**116**

f) 5 öznitelikli DVM

1-Sağlıklı kontrol grubu
2-Hasta grubu

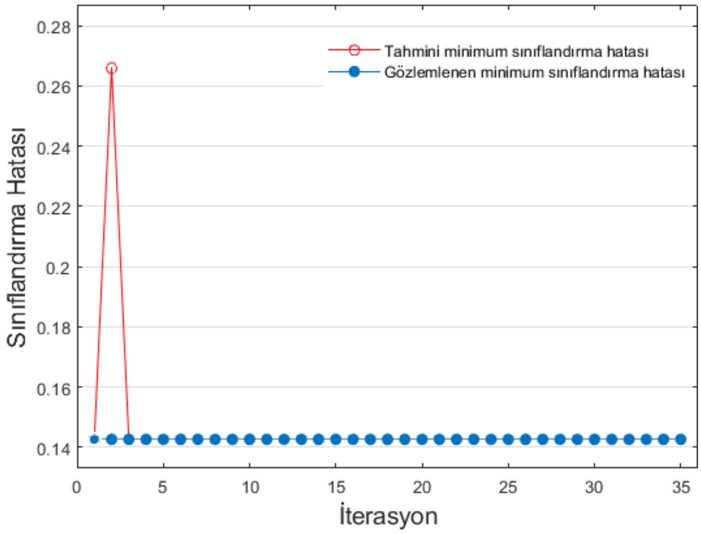
Farklı öznitelik ve sınıflandırıcı algoritmalardan elde edilen karışıklık matrisleri yardımıyla elde edilen doğruluk, duyarlılık, özgüllük, hassasiyet ve F1 skoru gibi performans göstergeleri Tablo 4'de gösterilmiştir.

Tablo 4. Öznitelik ve sınıflandırıcı seçimine göre sınıflandırıcı algoritmaların performansları

Öznitelik sayısı	Sınıf. Türü	Doğruluk	Duyarlılık	Özgüllük	Hassasiyet	F1 skoru
9 öznitelikli(Ö1-Ö9)	KNN	73.3	80.8	67.2	66.7	73.0
9 öznitelikli(Ö1-Ö9)	DVM	76.7	76.9	76.6	72.7	74.8
4 öznitelikli (Ö1-Ö4)	DVM	81.9	82.7	81.3	78.2	80.4
5 öznitelikli (Ö1-Ö4, Ö8)	DVM	84.5	78.8	89.1	85.4	82.0
5 öznitelikli (Ö1-Ö4, Ö8)	DVM*	**85.3**	**80.8**	**89.1**	**85.7**	**83.2**

Doğruluk skorları kullanılarak yapılan 9 öznitelik girişli iki sınıflandırıcı yöntemlerin karşılaştırmada %76.7 ile DVM, %73.3 yakalayan KNN'den daha başarılı olmuştur. İlkinde 9 girişli de daha başarılı olan DVM tekniğine uygulanan 4 öznitelikli ve 5 öznitelikli girişler için yapılan kıyaslamada; 4 öznitelikli ile %81.9 ile başarı elde edilirken; 5 öznitelikli de ise % 84.9 başarı elde edilmiştir. 5 öznitelikli veri setinin, sınıflandırmada en iyi performansı sergilediği görülmüştür.

DVM parametrelerinden kutu kısıtlama seviyesi 1'den 2.412'ye ve çekirdek skalası 3'den 3.532 'ye değiştirilerek ***DVM**** elde edilmiş ve sınıflandırma doğruluğu %85.3 ile çalışmanın en başarılı algoritması olmuştur. İterasyon sayısına göre minimum sınıflandırma hatasını gösteren grafik Şekil 1'de gösterilmiştir. En iyi sınıflandırma doğruluğu, birinci iterasyonda yakalanmıştır. Ayrıca normal ve hastaları tahmin için AUC (ROC eğrisi altında kalan alan) değeri %89 bulunmuştur.



Şekil 1. İterasyon sayısına göre minimum sınıflandırma hatası

4. SONUÇ

Dünya nüfusunda kadınlar üzerinde bu kadar etki yapan ve ölümlere neden olan bu kanser türü için erken teşhise katkı sağlayacak araştırma yapmanın önemli olduğu düşünülmüştür. Amaç iki sınıflandırıcıdan hangisinin ve bağımsız değişkenlerden hangilerinin meme kanser tahminde etkili olduğunun tespitini yapmaktır. Veri setinde bulunan 9 bağımsız değişkenin giriş olarak kullanıldığı sınıflandırıcı performansları karşılaştırıldığında destek vektör makinelerinin K en yakın komşu algoritmasına göre daha iyi olduğu anlaşılmıştır. Çalışmada meme kanser tahmini için yapılan en başarılı sınıflandırma, algoritma girişlerine glikoz, resistin, yaş, VKİ ve adiponektin verileri uygulandığında görülmüştür. Bu 5 değişkenin giriş yapıldığı ve destek vektör makinelerinin kullanıldığı bir sınıflandırma işleminde, en yüksek sınıflandırma doğruluğu, duyarlılık, özgüllük, hassasiyet, F1 skoru ve AUC değerleri elde edilmiştir. Çalışmanın sonuçları meme kanser teşhisine yardımcı olup, doğru teşhis koymayı destekleyebilir. Bir sonraki çalışmada performans değerlerini artırabilecek çalışmalar devam edecektir.

KAYNAKÇA

1. WHO, Breast Cancer, Son Erişim tarihi: 10.05.2020, World Health Organization. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.

2. Aličković, E. & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput & Applic 28:753–763. https://doi.org/10.1007/s00521-015-2103-9

3. Zhang, S. 2020. Cost-sensitive KNN classification, Neurocomputing, 391:234-242, https://doi.org/10.1016/j.neucom.2018.11.101.

4. Deng, Z., Zhu, X., Cheng, D., Zong, M. & Zhang, S. (2016). Efficient kNN classification algorithm for big data, Neurocomputing, 195:143-148. https://doi.org/10.1016/j.neucom.2015.08.112.

5. Wu, X., Zhang, C. & Zhang, S. (2004). Efficient mining of both positive and negative association rules, ACM Trans. Inform. Syst., 22 (3):381–405. https://doi.org/10.1145/1010614.1010616

6. Liu, P., Choo, K.R., Wang, L. et al. (2017). SVM or deep learning? A comparative study on remote sensing image classification. Soft Comput., 21:7053–7065 https://doi.org/10.1007/s00500-016-2247-2.

7. Crisóstomo, J., Matafome, P., Santos-Silva, D. et al. (2016). Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. Endocrine, 53(2):433-42. https://doi.org/10.1007/s12020-016-0893-x.

8. Patrício, M., Pereira, J., Crisóstomo, J. et al. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18- 29 https://doi.org/10.1186/s12885-017-3877-1

9. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra. Son Erişim Tarihi: 30.04.2020