VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini

Hasibe CANDAN[1]*, Arzu DURMUŞ[1], Güneş HARMAN[2]

[1]*Yalova Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yalova*
[2]*Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Yalova*

Makale Bilgisi

Başvuru:
06/12/2018
Kabul:
11/07/2019

Özet

Kanser vücutta hücrelerin kontrolsüz çoğalması ile oluşan bir hastalık grubudur. Çok çeşitli kanser türleri vardır. Hastalık erken teşhis edilip tedavi edilmezse ciddi rahatsızlıklara ve ölüme sebebiyet verebilir. Bu sebeple hastalığın erken teşhisi oldukça önemlidir. Bu çalışmada, akciğer ve beyin kanseri veri setleri üzerinde, Naive Bayes, Bayes NET, k-En Yakın Komşu, Rastgele Orman ve Destek Vektör Makineleri (DVM) makine öğrenmesi sınıflandırma yöntemleri kullanıldı. Genetik Algoritma (GA) ile öznitelik seçimi yapılarak ve tam veriler kullanılarak performans değerleri elde edildi. Hastalık tahmininde tam veriler ve öznitelik seçimi yapılmış verilerin performans değerleri karşılaştırıldı ve yorumlandı. En başarılı sonuçları akciğer kanseri verileri üzerinde %94.09, beyin kanseri verileri üzerinde ise %90 değerleriyle DVM vermiştir. Bu model, GA ile öznitelik seçimi uygulanmasının başarım oranını arttırdığını göstermektedir.

Anahtar Kelimeler: *Beyin Kanseri, Akciğer Kanseri, Genetik Algoritma, Özellik Seçimi, Naive Bayes, Bayes Ağları, k-En Yakın Komşuluk, Rastgele Orman, Destek Vektör Makineleri*

Cancer Prediction with Classification Methods and Genetic Algorithm

Abstract

Cancer is a group of diseases with uncontrolled proliferation of cells in the body. There are many types of cancer.if the disease is not diagnosed and treated early. Therefore, early diagnosis of the disease is quite ahead. In this study, Naive Bayes, Bayes NET, k-Nearest Neighbor (kNN), Rotasyon Ormanı and Support Vector Machines (SVM) machine learning classification methods were used on lung and brain cancer data sets. The system performance evaluation is based on original data and selected data. Complete data and attribute selected data were compiled and interpreted in disease prediction. SVM yielded 94.09%, 95.09%, at least results, and 90.09%, at least one result, 90 %. This model demonstrates that the success of the application of attribute selection with genetic algorithm increases the performance rate.

Keywords: *Brain Cancer, Lung Cancer, Genetic Algorithm, Feature Selection, Naive Bayes, Bayes Network, k-NN, Rotasyon Ormanı, Support Vector Machines*

* İletişim e-posta: ms.candan16@gmail.com

1 Giriş

Kanser organlarda doku ve hücrelerin kontrolsüz bir biçimde çoğalması ve büyümesi ile oluşan, tedavisi ve yaklaşımı birbirinden farklı olan bir hastalık grubudur. Kanser hastalığı çok hızlı metastaz göstermektedir [1]. Bu yüzden erken teşhis edilip tedavi edilmesi oldukça önemlidir.

Bilgisayar ve elektronik alandaki teknolojik gelişmeler, kanser hastalığının erken safhalarda doğru tanısını artırmıştır. Özellikle son yıllarda yapılan makine öğrenmesi temelli çalışmalar ile kanserin tanısında daha iyi sonuçlar alınmaktadır. Cheng Lung Huang ve arkadaşlarının yapmış olduğu çalışmada, Destek Vektör Makineleri (DVM) yöntemi ile % 89,6 başarı oranı elde etmişlerdir [2]. Diğer bir çalışmada Statnikov ve arkadaşları, kanser tanısı için DVM ve Rastgele Orman algoritmalarını kıyaslamışlardır. İlgili çalışmada DVM'nin daha iyi sonuç verdiğini tespit etmişlerdir [3]. Korkem ve arkadaşlarının, çalışmasında 9 farklı kanser türünden oluşan mikro dizi veri setleri üzerinde Rastgele Orman ve Naive Bayes yöntemleriyle sınıflandırma yapılmıştır. Her iki sınıflama yönteminin performanslarını incelenmişlerdir [4].

Cai Z. ve arkadaşları, akciğer kanseri türlerinden oluşan veriler üzerinde Rastgele Orman ve DVM algoritmalarını kullanılmıştır. DVM ile % 86.54 sınıf doğruluğu değeri elde etmişlerdir [5].

Bu çalışmada, akciğer ve beyin kanseri veri setleri üzerinde Naive Bayes, Bayes Ağları, k-En Yakın Komşu(k-EYK), Rastgele Orman ve DVM kullanılarak sınıflandırma yöntemlerinin kanser tespitindeki başarım oranları kıyaslanmıştır. Öznitelik seçimi yapılmamış performans değerleri ile Genetik Algoritmalar (GA) kullanılarak seçilmiş öznitelikler üzerinden kanser hastalığı tahmini yapılmıştır. En başarılı sonuçları akciğer kanseri verileri üzerinde % 94.09, beyin kanseri verileri üzerinde ise % 90 değerleriyle DVM yöntemi vermiştir. Bu model, genetik algoritma ile öznitelik seçimi uygulanmasının başarım oranını arttırdığını göstermektedir.

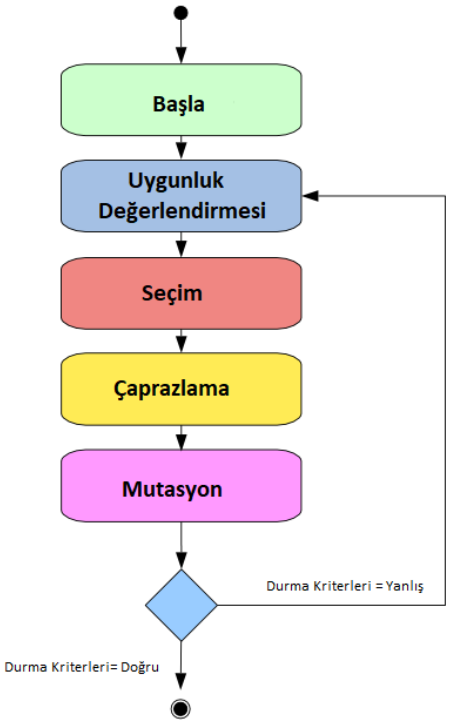
2 Materyal ve metotlar

2.1 Veriseti

Bu çalışmada akciğer ve beyin kanseri verileri kullanılmıştır. Akciğer verileri 12601 öznitelik ve 203 örnek içermektedir. Beyin verileri ise 5921 öznitelik ve 90 örnek içermektedir [6].

2.2 Öznitelik seçimi

Özellik Seçimi, veri kümesi içerisinde veri sınıfının hangi özelliklerinin sonuç üzerinde ne kadar etkili olduğunu belirlemek için özellikleri parça parça değerlendirmeye ve seçmeyi sağlayan yöntemlere denir. Özellik seçimin orijinal veri kümesi içerisinden (m < n) minimum m özellik alt kümesini seçme işlemidir [7]. Özellik seçimi makine öğrenmesi algoritmalının daha hızlı çalışmasını sağladı gibi oluşturulan modelin karmaşıklığını en basit düzeye indirgemiş olur. Sinyal işleme görüntü işleme vb. alanlarda kullan farklı öznitelik seçme algoritmaları bulunmaktadır. Bu çalışmada özellik seçimi için en gelişmiş algoritmalardan biri genetik algoritma kullanılmıştır. GA, doğal genetik ve biyolojik evrimin mekaniklerine dayalı fonksiyon optimizasyonu için stokastik bir yöntemdir [8]. GA'nın temel amacı fazla sayıda sınırlama içeren ve karmaşık optimizasyon sorunlarının çözümlerini, yazılımlar yardımıyla araştırmaktadır. GA evrimsel hesaplamanın bir parçasıdır. Problemler evrimsel bir süreç kullanılarak bu süreç sonunda en iyi sonucu veren çözüme erişmeye çalışmaktadır [9].
Şekil 1'de Genetik algoritmaların akış diyagramı verilmiştir.



Şekil 1 Genetik algoritmaların akış diyagramı

Özellik seçiminde Genetik algoritmalar 5 temel adımdan oluşmaktadır. Bu adımlar kısaca;

- ✓ Başla: Belirli kriterlere göre rastgele bireylerden oluşan bir popülasyon

oluşturulur. Rastgele toplum oluşturulur. Bu aşamada problemin çözümleri belirlenir

- ✓ Uygunluk: Popülasyondaki her x kromozomu için f(x) uygunluk değeri özellikler arasında bulunan karşılıklı bilgiyi hesaplamak için kullanılır.
- ✓ Seçim: Popülasyondan uygunluklarına göre iki ata seçilmektedir. Mevcut popülasyonda bulunan en iyi diziyi seçmek için kullanılır. Daha uygun olanın seçilme şansı daha yüksektir.
- ✓ Çaprazlama: Çaprazlama olasılığı ile ataları yeni yavru oluşturmak için birbirleriyle eşleştirilir. Eğer çaprazlama yapılmazsa, yavru ataların birebir aynısı olacaktır.
- ✓ Mutasyon: Mutasyon olasılığı ile yeni yavru üzerinde her yörünge için mutasyon işlemi yapılacaktır.

2.3 Sınıflandırma algoritmaları

Sınıflandırma, verileri benzerliklerine göre gruplama işlemidir. Sınıflandırıcı algoritmalar, verilen eğitim kümesi ile dağılım şeklini öğrenip, sınıfının belirli olmadığı test verileri üzerinde sınıflandırma yapmaktadır.

Bu çalışmada, Naive Bayes, Bayes NET, k-EYK, Rastgele Orman ve DVM makine öğrenmesi sınıflandırma yöntemleri kullanıldı. WEKA ile sınıflandırıcı yöntemlerin performans metrikleri elde edildi.

Naive Bayes, istatistiksel sınıflandırıcılardır ve sınıfı tahmin edebilir. Naive Bayes sınıflandırıcılar, belirli bir sınıfta bir özellik değerinin etkisinin olduğunu varsayar ve diğer özelliklerin değerlerinden bağımsızdır. Bu varsayım basitleştirmek için yapılır ve sınıf şartlı bağımsızlık denir [10].

Bayes NET, bazı koşulları veya durumu temsil eden modelleri içeren olasılıksal analiz yöntemleridir. Olasılıksal akıl yürütme yöntemleri, mevcut gözlemin depolanmış hipotezlerden birine karşılık geldiğine dair bir inancı belirlemek için kullanılır [11].

k-EYK, çok etiketli öğrenmede, eğitim seti, her biri bir etiket kümesiyle ilişkilendirilmiş örneklerden oluşur ve görev, bilinen etiket kümeleriyle eğitim örneklerini analiz ederek, sınıfı belli olmayan örneklerin etiket kümelerini tahmin etmektir [12].

Rastgele Orman, büyük veri kümeleri üzerinde hızlı bir şekilde çalışabilen, hesaplama açısından verimli bir tekniktir. Çok sayıda araştırma projesinde ve farklı alanlarda gerçek dünya uygulamalarında kullanılmıştır [13].

DVM, değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki sınıflandırma problemleri için önerilmiş bir makine öğrenmesi yöntemidir. İstatistiksel öğrenme teorisine ve yapısal risk minimizasyonuna dayanmaktadır [14].

3 Bulgular

Kanser tahmininde gerçekleştirilen deneylerde 10-kat çapraz doğrulama test tekniğini kullanılmıştır [15]. Bu teknikte veri seti 10 eşit kümeye ayrılır. Dokuzu eğitim, birisi test olmak üzere toplam on deney gerçekleştirilir. Daha sonra elde edilen performans değerlerinin ortalaması alınır. Makine öğrenmesinde kullanılan sınıflandırma algoritmaların performans değerlendirmeleri Şekil 2'de görülen karmaşıklık matrisi kullanılarak elde edilir. Karmaşıklık matrisi gerçek değer ve tahmin edilen değer arasındaki ilişkiyi gösterir.

		Tahmin	
		Doğru	**Yanlış**
Gerçek	**Doğru**	*DP*	*YN*
	Yanlış	*YP*	*DN*

Şekil 2 Karmaşıklık matrisi

DP: Gerçek doğrudur ve doğru olarak tahmin edilmiştir.
YN: Gerçek doğrudur, ancak yanlış olarak tahmin edilmiştir.
YP: Gerçek yanlıştır ve yanlış olduğu tahmin edilmiştir.
DN: Gerçek yanlıştır ancak yanlış olarak tahmin edilmiştir.

Performans değerlendirmesi için sınıf doğruluğu (SD), ROC eğrisi altında kalan alan (REKA) ve Matthew's Correlation Coefficient (MCC) değerleri hesaplanmıştır. SD, modelin hedef ve ikincil sınıfı ne kadar başarıyla tahmin ettiğini verir. Denklem 1'de SD değerinin formülü görülmektedir. REKA, duyarlığın 1-özgünlüğe oranını veren ROC eğrisi grafiği altında kalan alandır [16]. REKA değeri, 1'e ne kadar yakınsa modelin performansı o kadar iyi demektir. MCC değeri de Denklem 2'de görüldüğü gibi karmaşıklık matrisinden hesaplanır ve -1 ile 1 arasında değer alır. MCC değeri 1'e ne kadar yakınsa model o oranda her iki sınıfı dengeli, iyi tahmin ediyor demektir.

$$ACC = (DP + DN) / (P + N) \quad (1)$$

$$MCC = DP * DN - YP * YN/1/2((DP + YP)(DP + YN)(DN + YP)(DN + YN)) \quad (2)$$

Sınıflandırma algoritmalarının akciğer ve beyin kanseri için performans metrikleri ilgili tablolarda verilmiştir.

Tablo 1. Akciğer kanseri verileri üzerinde sınıflandırıcıların başarımları

	SD (%)	REKA	MCC
Naive Bayes	80,79	0,853	0,655
Bayes Ağları	73,89	0,838	0,581
k-EYK	89,66	0,881	0,777
Rotasyon Ormanı	89,16	0,983	0,753
DVM	**93,59**	**0,909**	**0,861**

Tablo 2. Genetik algoritma ile öznitelik seçimi uygulanmış akciğer kanseri verileri üzerinde sınıflandırıcıların başarımları

	SD (%)	REKA	MCC
Naive Bayes	90,15	0,914	0,803
Bayes Ağları	74,88	0,878	0,588
k-EYK	88,67	0,865	0,747
Rotasyon Ormanı	90,64	0,981	0,784
DVM	**94,09**	**0,916**	**0,872**

Tablo 1 ve Tablo 2'den görüldüğü üzere, akciğer kanseri verileri üzerinde yapılan çalışmada, GA ile öznitelik seçimi yapılmış olan veri seti üzerinde elde edilen performansların daha başarılı olduğu görülmektedir. Tam öznitelikler üzerinde DVM'nin SD değeri % 93,59 iken öznitelik seçimi sonrası % 94,09'a çıkmıştır. Hatta Naive Bayes için % 10'u geçen dramatik bir artış meydana gelmiştir. Kısaca, GA ile öznitelik seçimi, akciğer kanseri verileri üzerinde tüm başarım metrikleri açısından daha başarılı sonuç vermiştir.

Tablo 3. Beyin kanseri verileri üzerinde sınıflandırıcıların performansları

	SD (%)	REKA	MCC
Naive Bayes	**90,00**	0,85	**0,767**
Bayes Ağları	81,11	0,909	0,705
k-EYK	86,67	0,831	0,716
Rotasyon Ormanı	82,22	**0,912**	0,602
DVM	85,56	0,847	0,708

Tablo 3'de görüldüğü gibi beyin kanseri verileri üzerinde öznitelik seçimi yapmadan önce Naive Bayes yöntemi % 90 SD ve 0,767 MCC değerleri ile en iyi performansı sergilemiştir.

Tablo 4. Genetik algoritma ile öznitelik seçimi uygulanmış beyin kanseri verileri üzerinde sınıflandırıcıların performansları

	SD (%)	REKA	MCC
Naive Bayes	83,33	0,806	0,642
Bayes Ağları	81,11	0,888	0,673
k-EYK	84,44	0,829	0,681
Rotasyon Ormanı	77,78	0,902	0,5
DVM	**90**	**0,882**	**0,79**

Tablo 4'de görüldüğü gibi GA ile öznitelik seçimi yöntemi, Naive Bayes sınıflandırıcısının başarımını yaklaşık olarak % 7 oranında düşürmüştür. Öznitelik seçimi sonuçlarına göre en iyi performansı DVM sınıflandırıcısı göstermiştir. GA ile öznitelik seçimi DVM sınıflandırıcısı için % 4,44 değerinde performansı artırmıştır.

4 Sonuçlar

Bu çalışmada kanser verileri üzerinde GA ile öznitelik seçimi yapılarak Naive Bayes, Bayes Ağları, k-EYK, Rotasyon Ormanı ve DVM algoritmaları ile sınıflandırma yapılmıştır. GA ile öznitelik seçimi işleminin DVM sınıflandırıcısının başarımını artırmıştır. Bu çalışma, GA ile öznitelik seçimi işleminin makine öğrenmesi ile yapılacak olan kanser teşhisi çalışmalarında başarım açısından önemli olduğunu göstermiştir. İlerde yapılacak çalışmalarda, farklı öznitelik seçimi yöntemlerinin kanserin teşhisi açısında kıyaslanması ve kanserin makine öğrenmesi yöntemleri ile teşhisini gerçekleştiren bir web sunucu geliştirilmesi planlanmaktadır.

Kaynaklar

[1] Wang, K. J., Makond, B., & Wang, K. M. (2014). Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: a case study of Taiwan. Computers in biology and medicine, 47, 147-160.

[2] Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with applications, 31(2), 231-240.

[3] Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of Rotasyon Ormanıs and support vector machines for microarray-based cancer classification. BMC bioinformatics, 9(1), 319.

[4] Korkem, E. (2013). Mikroarray Gen Ekspresyon Veri Setlerinde Rotasyon Ormanı ve Naive Bayes Sınıflama Yöntemleri Yaklaşım.

[5] Cai, Z., Xu, D., Zhang, Q., Zhang, J., Ngai, S. M., & Shao, J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molecular BioSystems, 11(3), 791-800.

[6] Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., ... & Loda, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences, 98(24), 13790-13795.

[7] Gök, M. (2015). An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. International Journal of Systems Science, 46(6), 1108-1112.

[8] Whitley, D. (1994). A genetic algorithm tutorial. Statistics and computing, 4(2), 65-85.

[9] Koza, J. R., & Koza, J. R. (1992). Genetic programming: on the programming of computers by means of natural selection (Vol. 1). MIT press.

[10] Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering.

[11] Valdes, A. D. J., Fong, M. W., & Porras, P. A. (2008). U.S. Patent No. 7,379,993. Washington, DC: U.S. Patent and Trademark Office.

[12] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

[13] Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a Rotasyon Ormanı?. In International Workshop on Machine Learning and Data Mining in Pattern Recognition(pp. 154-168). Springer, Berlin, Heidelberg.

[14] 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.

[15] Schaffer, C. (1993). Selecting a classification method by cross-validation. Machine Learning, 13(1), 135-143.

[16] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.