DVM Tabanlı Kalın Bağırsak Kanseri Tanısı İçin Performans Geliştirme
Performance Improvement for SVM Based Colon Cancer Recognition

Sebahattin Babur[1], Uğur Turhal[2], Ahmet Akbaş[3]

[1,2,3] Bilgisayar Mühendisliği Bölümü
Yalova Üniversitesi
sebahattin_babur@hotmail.com, ugurturhal@hotmail.com,
ahmetakbas@yalova.edu.tr

Özet

Tanıma problemlerinde doğru sınıflandırma oranını artırmak ve gereksiz işlem gücü harcamasından kaçınmak için, öznitelik vektörlerinin olabildiğince az sayıda veriyle ve isabetli bir şekilde belirlenmesi önem arz eder. Bu çalışmada böyle bir yaklaşımla kalın bağırsak kanserinin tanınması amacıyla kullanılan 2000 öznitelik verisi içinden tanımada yüksek etkinlik sağlayan 13 tanesi seçilerek sınıflandırıcının performansı artırılmıştır. Çalışma 3 aşamadan oluşmaktadır. İlk aşamada, her bir öznitelik verisinin sınıflandırmadaki ağırlığı, destek vektör makinelerinin (DVM) kuadratik fonksiyonu kullanılarak tespit edilmiştir. İkinci aşamada ağırlıkları tespit edilen öznitelik verileri içinden yüksek ağırlığa sahip olanlar seçilmiştir. Son aşamada rastgele alt örnekleme (random subsampling) yöntemi kullanılarak 62 denekten elde edilen verilerin tamamı içerisinden seçilen 20 farklı eğitim ve test grubu ile denemeler yapılmıştır. Elde edilen sonuçlar, bu çalışmada kullanılan yaklaşımın hastalıklı ve sağlıklı kişilerin sınıflandırılma doğruluklarının %88.55±4.74 aralığında değiştiğini göstermiştir. Bu sonuç, kullanılan yaklaşımın DVM tabanlı kalın bağırsak kanseri tanısında oldukça tatmin edici bir performansa sahip olduğunu göstermiştir.

Abstract

To solve the recognition problems it is an important issue reducing the size of feature vectors in order to improve the correcness of classification, and in order to reduce the need of computation power. In this study, the feature vectors having 13 data have been produced from the feature vectors having 2000 data, in order to improve the correctness of classification of colon cancer. The study have 3 stages. In the first stage the weight of each data in the 2000 elements of original vectors have been calculated by using quadratic function of support vector machines (SVM). In the second stage 13 data having highest weights have been selected. In the last stage 20 different training/test data set have been produced from 62 data set by using random subsampling. For 20 sets training and testing have been completed. The results have shown that the correctness of classification can be increased up to %88.55±4.74. This is an sufficient performance for generalization of training methode selected.

1. Giriş

İstatistiklere göre batı ülkelerinde en yüksek ölüm oranına sahip olan hastalıklar arasında ikinci sırayı kalın bağırsak kanseri almaktadır. Dünya Sağlık Örgütünün 2006 raporuna göre, dünyanın değişik bölgelerinde, yılda 655,000 kişi kolorektal kanser nedeniyle hayatını kaybetmektedir [1]. Dolayısıyla bu hastalığın teşhisi ve tedavi sürecinde elde edilecek başarı, öncelikle teşhis için kullanılacak verilerin hızlı ve doğru bir şekilde belirlenmesine bağlıdır.

Gereksiz yere kullanılan öznitelik vektörleri, tanıma ve sınıflandırma algoritmalarının performansını azaltır. Dolayısıyla, bir yandan doğru tanıma performansını artırmak, diğer yandan gereksiz işlem gücü harcamasından kaçınmak için, kullanılacak olan öznitelik vektörlerinin isabetli bir şekilde belirlenmesi önemlidir. Bu amaçla yapılacak çalışmalar için destek vektör makineleri (DVM) etkin bir çözüm sunmaktadır [2,3].

DVM'nin veri madenciliği, biyoinformatik, bilgisayarlı görme gibi birçok çalışma alanında uygulamaları vardır. Bu çalışmada DVM'nin sağladığı avantajlardan yararlanılarak, kalın bağırsak kanserinin tanınması amacıyla kullanılan çok sayıda öznitelik verisi içerisinden, tanımada yüksek etkinlik sağlayan öznitelik verileri seçilerek sınıflandırma performansının artırılması amaçlanmıştır.

2. Materyal ve Metod

Çalışmada 62 deneğe ait veriler kullanılmıştır [4]. Veri setinde bunlardan 40 deneğe bağırsak kanseri teşhisi konulmuş, 22 denek ise sağlıklı olarak belirtilmiştir. Hasta kişiler "1", sağlıklı kişiler "0" olarak etiketlenmiştir. Her denek için bağırsak kanseri teşhisinde anlamlı olduğu düşünülen 2000 adet öznitelik verisi verilmiştir. Veri setinin bu özellikleri Tablo 1'de verilmiştir.

*Tablo 1 :*Veri seti özellikleri

Veri Seti	Veri Sayısı	Öznitelik Sayısı	Kaynak
Bağırsak Kanseri	62	2000	[4]

Çalışmada kullanılan tanıma yaklaşımının ön işlem aşamasında veriler normalize edilmiştir. Bunun için aşağıdaki eşitlik kullanılmıştır.

$$Z = \frac{X-\mu}{\sigma} \quad (1)$$

Burada X veri setini (x_k; k =1; 2; _ _; K), μ aritmetik ortalamayı ve σ standart sapmayı, Z normalize edilmiş veriyi temsil etmektedir. Veriler normalize işleminden geçirildikten sonra, DVM ile öznitelik seçimi yapılmıştır. Bunlardan anlamlı olanlarını seçmek için literatürde en çok 't-testi've 'F-Score' yöntemleri kullanılmaktadır [5,6]. Bu çalışmada da bu testlerden yararlanılmıştır. Çalışmada DVM uygulamaları MATLAB'ın 'svm' fonksiyonu kullanılarak gerçekleştirilmiştir.

Verilerin anlamlılık derecelerini belirlemek için t-testi ile öznitelik verilerinin istatistiksel p değerleri belirlenmektedir. Özniteliğin p değeri ne kadar küçükse, iki sınıfı ayırmadaki anlamlılık derecesi de o kadar yüksek olmaktadır [5]. F-Score yöntemi ile elde edilen verilerin ağırlıklarını hesaplamak için aşağıdaki eşitlik kullanılmıştır [6].

$$F_i = \frac{(\bar{X}_i^{(+)} - \bar{X}_i)^2 + (\bar{X}_i^{(-)} - \bar{X}_i)^2}{\frac{1}{(n_+ - 1)}\sum_{k=1}^{n_+}(\bar{X}_{k,i}^{(+)} - \bar{X}_i^{(+)})^2 + \frac{1}{(n_- - 1)}\sum_{k=1}^{n_-}(\bar{X}_{k,i}^{(-)} - \bar{X}_i^{(-)})^2} \quad (2)$$

Burada F_i, özniteliğin F-skorunu; i öznitelik indisini; x_k; k =1; 2; _ _; K veri setini; n_+ ve n_- hasta ve normal sınıfları temsil etmektedir.

2.1. DVM ile Sınıflandırma

DVM, istatistiksel öğrenme teorisine dayalı sınıflandırma yapan bir algoritmadır. Başlarda sadece iki sınıflı doğrusal yapıları sınıflandırmak için kullanılan DVM, sonraları doğrusal olmayan verileri sınıflandırmak için de kullanılmıştır [7].

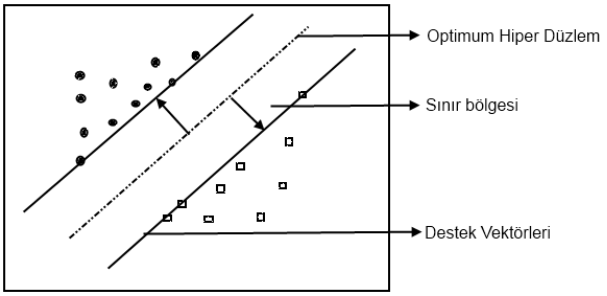
DVM algoritmasına göre sınıflar genellikle {-1,+1} şeklinde gösterilir. Bu iki sınıfa ait eğitim verilerine göre ayrım fonksiyonu oluşturulur. Bu ayrım fonksiyonu, test verilerini her zaman doğru sınıflandıramayabilir. Çünkü eğitim sırasında veriler her zaman lineer bir dağılım göstermezler. Böyle durumlarda sistem test edildiğinde hatalı bir sınıflandırma yapılmış olur.

Şekil 1'de gösterildiği gibi iki yapıyı birbirinden ayırabilen birçok düzlem çizilebilir. Ancak DVM'nin amacı kendisine en yakın noktalar arasındaki mesafeyi maksimuma çıkaran düzlemi bulmaktır. En uygun ayrımı yapan düzleme optimum hiper-düzlem ve sınır genişliğini sınırlandıran noktalara destek vektörleri denir [7].

$$\omega \cdot x_i + b \geq +1 \quad \text{her} \quad y = +1 \quad \text{için} \quad (3)$$

$$\omega \cdot x_i + b \leq -1 \quad \text{her} \quad y = -1 \quad \text{için} \quad (4)$$

Burada $x \in R^N$olup N-boyutlu bir uzayı, $y \in \{-1, +1\}$ ise sınıf etiketlerini temsil etmektedir. ω ağırlık vektörünü ve b eğilim değerini göstermektedir [8].



*Şekil 1:*İki sınıflı bir problem için Hiper-Düzlem ve destek vektörleri

İkiden fazla boyuta sahip sınıfların doğrusal olarak ayrılması zordur. Girdi uzayında doğrusal olarak ayrılamayan sınıflar yüksek boyutlu bir uzaya yerleştirilir. Böylece sınıflar arasında bir ayrım yapılabilmekte ve hiper-düzlem çizilebilmektedir. Buradan da anlaşılacağı gibi paremetreler için uygun değerlerin seçimi, sınıflandırıcıların ayrım fonksiyonu oluşturmasındaki başarısını oldukça etkilemektedir [7].

2.2. Öznitelik Seçimi/Boyut İndirgeme

Boyut indirgeme yöntemleri; temel ya da bağımsız bileşenlere en az katkı yapan özniteliklerin düşük ayırt edici özelliğe sahip olduğu fikrini esas alarak, daha anlamlı olan özniteliklerin yeni bir sınıflandırma uzayı altında toplanmaları esasına dayanır. Bunun için kullanılan bir çok yöntem bulunmaktadır.

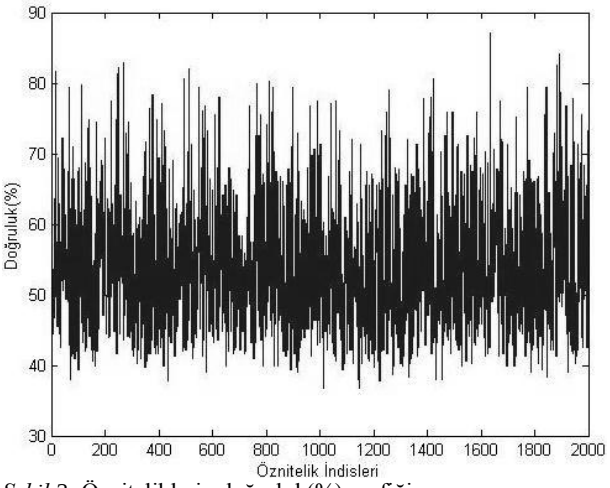
Özellikle kolorektal kanser hastalığı teşhisinde, öznitelik vektörlerinin mümkün olan en az sayıda veriyle tespit edilmesi, hesaplama karmaşıklığı açısından da önemlidir.

Bizim çalışmamızda, öznitelik uzayındaki her bir bileşen için bir DVM yapısı oluşturulmuş ve sonucunda performans verisi elde edilmiştir. Bileşenlerin performansını hesaplamak için aşağıdaki eşitlik kullanılmıştır.

$$Doğruluk = \frac{GPS+GNS}{TS} \quad (5)$$

Burada *GPS* doğru pozitiflerin sayısı, *GNS* doğru negatiflerin sayısı *TS* test verilerinin sayısını temsil etmektedir.

2000 adet özniteliğin her biri için bu yolla hesaplanan performans değerleri Şekil 2'deki grafikte toplu olarak gösterilmiştir. Grafikte özniteliklerin doğruluğu 0 ile 100 arasındaki yüzdelik oranları ile verilmiştir.



Şekil 2: Özniteliklerin doğruluk(%) grafiği

Doğruluk oranı en yüksek olan öznitelikler, sınıflandırma sonucunu en fazla etkileyen bileşenlerdir. 2000 adet öznitelik için yapılacak indirgeme işleminde bir eşik değeri belirlenir ve doğruluk oranı bu eşik değerine eşit veya bunun üzerinde olan tüm öznitelikler yeni öznitelik uzayına aktarılır. Bu sistemde, Şekil 2 deki doğruluk grafiği göz önüne alınarak, eşik değeri P=80 olarak belirlenmiştir. Buna göre belirlenen 13 indis Tablo 2'de gösterilmiştir.

Tablo 2 : Eşik değerine göre belirlenen öznitelikler

Seçilen Özniteliklerin İndis Numaraları	Doğruluk Yüzdeleri(%)
F14	80.65
F245	82.26
F249	83.87
F267	81.61
F377	83.23
F493	83.55
F765	80.97
F822	80.97
F1423	82.58
F1582	80.97
F1635	87.74
F1771	80.97
F1892	82.26

2.3. Algoritma

Doğruluk tabanlı öznitelik seçme algoritması aşağıda özetlenmiştir:

Giriş: Eğitim örnekleri
$Edata = \{(x_i , \quad y_i) | x_i \in R^n, y_i \in \{1,0\}, i = 1,2,\dots,l\}$
Çıkış: Öznitelik ağırlıkları r=[ ]
Veriler $G = \{F_i \mid i = 1,\dots,n\}$
Eşik parametresi P

For i = 1 : n
DVM Yapısı (x_i , y_i)
Formüle Göre (5)
Eğer Doğruluk ≥ P
$F_j = Edata\,(x_i , y_i)$
End

2.4. Öznitelikler kullanılarak DVM ile Sınıflandırma

DVM, istatistiksel öğrenme teorisine dayanan, sınıflandırma ve uyumlandırma problemlerinin çözümü için önerilen bir yöntemdir [9].

Kanser hastası deneklerini normal deneklerden ayırmak için kullanılan sınıflandırma yapısının son aşamasında; lineer, rbf (radial basis function), mlp (multilayer perceptron) ve kuadratik fonksiyonları kullanılmıştır. Sınıflandırıcının testi için 20 katlı Random SubSampling metodu uygulanmıştır. Bu yöntemle toplam verinin %50 'si eğitim grubu, kalanı test grubu olarak belirlenmiş ve sınıflandırma işlemi gerçekleştirilmiştir. Aynı işlem, veri grubunun sırasıyla %60'ı %70'i ve %80'nin eğitim grubu olarak belirlenmesiyle tekrarlanmıştır [10,11]. Buna göre elde edilen sınıflandırma performanslarının ortalama ve standart sapma değerleri Tablo 3 ve Tablo 4'te verilmiştir. Tablo 3'te verilen değerler, öznitelik vektörlerinin bu çalışmada indirgenmiş 13 veri ile oluşturulmuş hali için; Tablo 4'de verilen değerler, öznitelik vektörlerinin orijinal veri setinde verildiği şekliyle 2000 veri ile oluşturulmuş hali için elde edilen sonuçları yansıtmaktadır.

Tablo 3 : 13 Öznitelikli sınıflandırıcının performans sonuçları

	(%50)	(%60)	(%70)	(%80)
Lineer	82.90 ±4.69	85.60 ±7.27	86.84 ±5.79	86.67 ±8.72
Rbf	80.32 ±6	80.20 ±7.62	82.37 ±8.41	77.50 ±12.99
Mlp	**88.55 ±4.74**	85.80 ±4.94	87.37 ±4.95	85.83 ±8.59
Kuadratik	79.52 ±6.80	81 ±6.73	81.05 ±5.51	83.75 ±7.39

Tablo 4 : 2000 Öznitelikli sınıflandırıcının performans sonuçları

	(%50)	(%60)	(%70)	(%80)
Lineer	82.58± 4.49	**83± 6.07**	82.11± 7.12	80± 8.29
Rbf	64.52± $1x10^{-14}$	64± $1x10^{-14}$	63.16± $1x10^{-14}$	66.67± $2x10^{-14}$
Mlp	64.84± 10.56	64.4± 7.99	67.89± 10.79	64.17± 14.83
Kuadratik	56.77± 6.05	58± 6.16	58.95± 7.56	59.17± 10.78

3. Sonuçlar

Orijinal veri setindeki 2000 veri içerisinden seçilen ve Tablo 2'de indis numaraları verilen 13 öznitelik verisinin kullanıldığı vektörlerle elde edilmiş DVM tabanlı sınıflandırma sonuçları Tablo 3'de verildiği gibidir. Buna göre, hasta ve normal deneklerin doğru sınıflandırılma oranı için elde edilen en büyük performans değeri %88.55±4.74 olarak bulunmuştur. Bunun için kullanılan sınıflandırma fonksiyonu mlp, eğitim verisinin tüm veri seti içindeki oranı da %50 olarak belirlenmiştir.

Buna karşılık orijinal 2000 öznitelik vektörünün kullanıldığı DVM tabanlı sınıflandırma sonuçları Tablo 4'de verildiği gibidir. Buna göre, hasta ve normal deneklerin doğru sınıflandırılma oranı için elde edilen en büyük performans değeri %83±6.07 olarak bulunmuştur. Bunun için kullanılan sınıflandırma fonksiyonu lineer, eğitim verisinin tüm veri seti içindeki oranı da %60 olarak belirlenmiştir.

Bu sonuçlara göre, öznitelik verilerinin azaltılması suretiyle elde edilen tanıma performansları, elde edilen ortalamalar itibariyle, seçilen bütün kombinasyonlar için %80'in üzerinde gerçekleşmiştir. Bu oran verilerin azaltılmamış hali için %57 oranına kadar düşmektedir. Benzer şekilde elde edilen standart sapmalar da verinin azaltılmış olması hali için daha az bulunmuştur. Bu sonuçlar, veri indirgemenin sınıflandırmada önemli bir performans artışı sağladığını göstermektedir. En iyi sonuç esas alındığında elde edilen 4.74 standart sapma değeri, 62 veri seti ile yapılan eğitimin genelleştirilebileceğini göstermektedir.

Bu sonuç aynı zamanda, 13 elemanlı verilerin oluşturduğu vektörlerle çalıştırılan tanıma algoritmalarının, orijinal olarak 2000 elemanlı verilerin oluşturduğu vektörlerle çalıştırılan tanıma algoritmalarına nazaran daha az işlem gücü gerektirmesi açısından önemli bir avantaj sağladığını göstermektedir.

4. Kaynaklar

[1] Alladi, S. M. ve Santosh P., Shinde "Colon cancer preiction with genetic profiles using intelligent techniques", *Bioinformation by Biomedical Informatics Publishing Group*, 3(2)., 130-133, 2008

[2] Cortes, CorinnaveVapnik, V., "Support-Vector Networks", *Machine Learning*, 20,1995

[3] Vapnik, V., *Statistical Learning Theory*, Wiley, Newyork, 1998

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[5] Yıldız, A., Akın, M., Poyraz, M., "EKG kayıtlarından Tıkayıcı Uyku Apne Sendremonun Otomatik Teşhisi", *Sinyal işleme ve iletişim uygulamaları kurultayı*,2010, 97-100

[6] Aruna, S., Nandakishore, L.V., "Application of Gist SVM in Cancer Detection", *Anale. Seria Informatica*,2011, 39-48

[7] Kavzoğlu, T., Çölkesen, İ., "Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi", *Harita Dergisi*, Sayı.144., 73-82, 2010

[8] Osuna, E.E., Freund, R., Girosi, F., "Support Vector Machines: Training and Applications", *Massachusetts Institute of Technoloy and Artifical Intelligence Laboratory*, 1602, 144, 1997

[9] Burges, C.J.C., "A tutorial on support vector machines for pattern recognotion", *Knowledge Discovery Data mining*, Vol.2, no.2, 1998

[10] Chawla, N.V., Bowyer, K.W., "Random subspaces and sub sampling for 2-D face recognition", *Computer Vision and Pattern Recognition*, Vol.2, 582-589, 2005

[11] Kohavi, R., "A study of cross validation and bootstrap for accuracy estimation and model selection*", in Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137-1143, 2002