Gazi Üniv. Müh. Mim. Fak. Der.
Cilt 27, No 3, 659-668, 2012

Journal of the Faculty of Engineering and Architecture of Gazi University
Vol 27, No 3, 659-668, 2012

MEME KANSERİ SINIFLANDIRMASI İÇİN VERİ FÜZYONU VE GENETİK ALGORİTMA TABANLI GEN SEÇİMİ

Oktay Yıldız*, Mesut Tez, H.Şakir Bilge*, M.Ali Akcayol*, İnan Güler*****
* Gazi Üniversitesi, Mühendislik Fak., Bilgisayar Mühendisliği Böl., Ankara.
** Ankara Numune Hastanesi.
*** Gazi Üniversitesi, Teknik Eğitim Fak., Elektronik-Bilgisayar Eğitimi Böl., Ankara.
oyildiz@gazi.edu.tr, **mesuttez@yahoo.com**, **bilge@gazi.edu.tr,** **akcayol@gazi.edu.tr**, **iguler@gazi.edu.tr**

(Geliş/Received: 15.11.2011; Kabul/Accepted: 09.02.2012)

ÖZET

Ciddi rahatsızlıklardan biri olan meme kanserinin erken teşhis edilmesi hastalığın tedavisinde önemli rol oynar. Bu çalışmada meme kanseri hastalığında etkin rol oynayan genlerin belirlenmesi için veri füzyonu ve genetik algoritma tabanlı yeni bir nitelik seçme metodu önerilmektedir. Yapılan çalışma iki aşamadan oluşmaktadır: İlk aşamada filtreleme yöntemi ile gen ifade verisi indirgenmiş, ikinci aşamada genetik algoritma ile meme kanserinde etkin rol alan genlerin tespiti gerçekleştirilmiştir. Destek vektör makinesi, genetik algoritma için uygunluk fonksiyonu olarak kullanılmıştır. Yapılan çalışmada belirlenen 10 gen ile sınıflandırma doğruluk oranı %94,65 elde edilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Nitelik Seçme, Veri Füzyonu, Genetik Algoritma, Meme Kanseri, Destek Vektör Makinesi.

GENE SELECTION FOR BREAST CANCER CLASSIFICATION BASED ON DATA FUSION AND GENETIC ALGORITHM

ABSTRACT

Early diagnosis of breast cancer has been playing very important role on treatment of the disease. In this work, a new feature selection method for breast cancer classification based on data fusion and genetic algorithm is presented. The study consists of two steps: In the first step, the dimensionality of the gene expression dataset was reduced with filter method and the second step, significant genes have been identified with genetic algorithm. SVM was used for fitness function in genetic programming. In this study the classification accuracy rate was obtained 94.65 % when using selected 10 genes.

Key Words: Data Mining, Feature Selection, Data Fusion, Genetic Algorithm, Breast Cancer, Support Vector Machine.

1. GİRİŞ (INTRODUCTION)

Son yıllarda DNA analizlerinde çok büyük gelişmeler yaşanmıştır. Bunlardan biri aynı anda on binlerce genin birbirleriyle olan etkileşiminin ölçülebilmesine imkan sağlayan gen çipleri diğer bir adıyla mikrodizi (microarray) teknolojisidir. Mikrodizi teknolojisi sayesinde hastalıkların temelinde yatan genetik faktörler belirlenebilir, hastalığın erken teşhisi gerçekleştirilebilir, yeni ilaçlar geliştirilebilir veya bireye özel tedavi süreci tanımlanabilir. Günümüzde pek çok hastalığın teşhis ve tedavisinde DNA analizleri sıklıkla kullanılmaktadır [1-3,28]. Kanser gibi ciddi rahatsızlıkların temelinde yatan genetik faktörlerin belirlenmesi de bu hastalığın tedavisine çok önemli katkılar sağlayabilir.

Gen ifade verileri az örneklem ve çoğu gürültü olarak adlandırılan ilgisiz on binlerce gen bilgisini içermektedir. Bu sebeple bu verilerin sadece istatistiksel yöntemlerle analiz edilmesi oldukça zordur. Ayrıca verilerin yüksek boyutlu olması

sınıflandırma problemini de beraberinde getirmektedir. Yüksek boyutun indirgenmesi için nitelik seçme gen analizlerinde kritik öneme sahiptir. Nitelik seçme işlemlerinde makine öğrenme yöntemleri başarılı bir şekilde kullanılmaktadır [4]. Hastalıkların temelinde yatan genetik faktörlerin belirlenmesi amacıyla, nitelik seçimi gerçekleştiren pek çok çalışma yapılmıştır [5-10].

Horng ve arkadaşları [7], biyomarker genlerin belirlenmesi için üç aşamadan oluşan yeni bir yöntem önermişlerdir. Bu yaklaşım ilk aşamada veri girişi olarak adlandırılan mikrodizi verilerinin bir matrise aktarılması işlemidir. İkinci aşamada örnek sayısı Antonov yaklaşımı [12] ile çoğaltılmış ve sistem C4.5 karar ağacı ile eğitilmiştir. Son adımda WEKA' da Naive Bayes, Karar ağaçları ve Destek Vektör Makinesi sınıflandırma algoritmaları ile başarı test edilmiştir. Li ve arkadaşları [10], Genetik Algoritma ve Destek Vektör Makinesi tabanlı gen seçimi gerçekleştirmişlerdir. Genetik algoritma bir arama motoru, destek vektör makinesi ise bir sınıflandırıcı olarak kullanılmıştır. Belirlenen genler ile %99 sınıflandırma başarısı elde edildiği ifade edilmiştir. Wang ve arkadaşları [8], çeşitli filtreleme yöntemleri ile belirlenen mikrodizi verilerinde korelasyon tabanlı nitelik seçme işlemi gerçekleştirmişlerdir. Elde edilen sonuçlar farklı makine öğrenme yöntemleri ile test edilmiştir. Uriarte ve Andres [9], Rastgele Orman (Random Forest) ile gen seçimi gerçekleştirmişler, elde ettikleri sonuçları k-En Yakın Komşu ve Destek Vektör Makinesi ile test etmişlerdir. Alon ve arkadaşları [5], farklı hücre tiplerine sahip gen ifade verilerinden oluşan veri kümesinin analizi için iki yönlü kümeleme metodu ve uygulaması sunmuşlardır. Bu çalışmada 40 tümör ve 20 normal doku bilgisi ile 6500 den fazla genetik bilgiyi içeren Affymetrix oligonükleotid dizisi başarılı bir şekilde sınıflandırılmıştır. Dudoit ve arkadaşları [6], gen ifade verilerinde tümör sınıflandırmasında k-En Yakın Komşu, Doğrusal Ayırtaç Analizi (LDA, Linear Discriminant Analysis) ve Sınıflandırma Ağaçlarının performansını karşılaştırıp, bu sınıflandırıcıları lösemi, lenfoma ve 60 kanserli doku içeren veri kümesinde test etmişlerdir.

Yukarıda da belirtildiği gibi gen analizlerinde daha önce çalışılmış çeşitli nitelik seçme yaklaşımları vardır. Ancak uygulanan istatistiksel yöntemler en uygun nitelik altkümesini sağlamayı garanti etmez. Ayrıca genetik algoritma ve destek vektör makinesi ile gen seçimi çalışmalarının pek çoğu en uygun nitelik altkümesini elde etmek için önemli ölçüde hesaplama karmaşıklığına sahiptir.

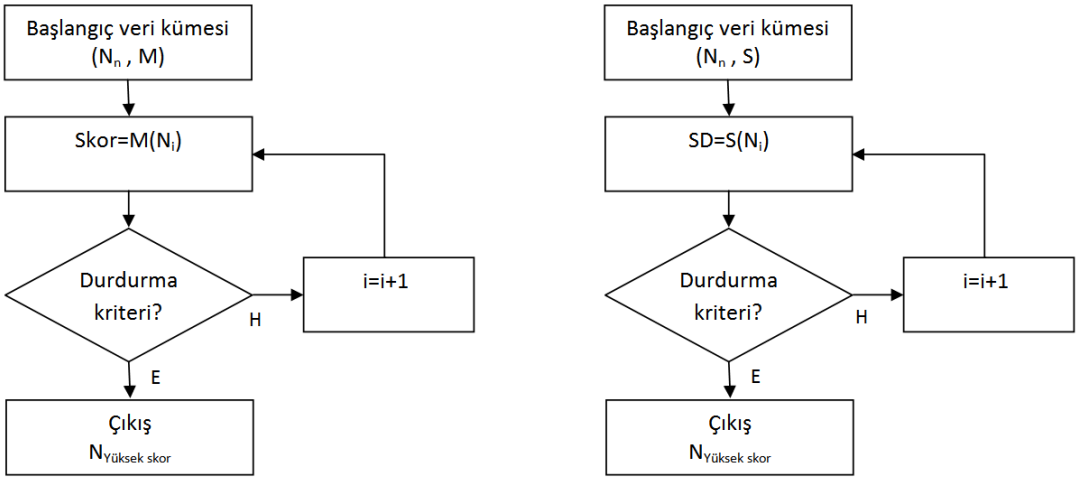
Bu çalışmada, gen ifade verilerinden genetik algoritma ve veri füzyonu tabanlı gen seçimi gerçekleştirilmiştir. Genetik algoritma, yerel en iyiye takılmaması ve nitelik seçme işlemlerinde başarılı sonuçlar vermesinden dolayı tercih edilmiştir. Çalışma iki aşamadan oluşmaktadır. İlk aşama, genetik algoritma için başlangıç veri kümesinin belirlenmesi aşamasıdır. Bunun için Fisher Korelasyon Skorlama, t-Skor ve WTS (Welch t-statistic) ile gen skorlaması yapılmıştır. Elde edilen yeni veri kümesi bu üç yöntem ile elde edilen sonuçların birleştirilmesi ile belirlenmiştir. Burada üç ayrı filtreleme yönteminde de ortak belirlenen yüksek skorlu genlerle, her bir skorlama yönteminin ayrı ayrı belirlediği genler yeni veri kümesine dahil edilmiştir. İkinci aşamada belirlenen veri kümesi içinden Genetik Algoritma ile gen seçimi gerçekleştirilmiştir. Genetik algoritma için uygunluk fonksiyonu olarak Destek Vektör Makinesi kullanılmıştır. 10 kat çapraz doğrulama ile sınıflandırıcı performansı test edilmiştir.

2. MATERYAL VE METOT (MATERIAL AND METHOD)

2.1. Nitelik Seçme (Feature Selection)

Pek çok örüntü tanıma tekniği veri kümesinde ilgisiz nitelikleri temizlemek için nitelik seçme işlemlerine ihtiyaç duyar. Nitelik seçme, izdüşüm yöntemine dayalı boyut indirgeme (Principal Component Analysis – PCA vb.) ve diğer sıkıştırma tekniklerinin aksine orijinal nitelikler içinden bir alt küme seçme işlemine dayanır [12]. Nitelik seçme, ilgisiz niteliklerin atılması ya da verinin gürültüden temizlenmesi işlemidir. Nitelik seçme hem sınıflandırma başarısını ve performansını doğrudan etkiler hem de ezberleme riskini de azaltır [4,12,13]. Nitelik seçme ile aşırı eğitim problemi aşılarak model performansı arttırılabilir, düşük maliyetli, etkin ve hızlı modeller sunulabilir, ayrıca veri elde etme süreci daha iyi detaylandırılabilir. Nitelik seçme, filtreleme ve sarmal olmak üzere iki ayrı grupta incelenebilir [14,15].

Şekil 1' de nitelik seçme yöntemleri görülmektedir. Şekil 1(a) da görüldüğü gibi filtreleme yöntemi, herhangi bir öğrenme algoritmasından bağımsız, zayıf bilgi içeren nitelikleri süzmek için istatistiksel özellikleri kullanır [13,16]. Çoğu uygulamada özellik ilişki skoru hesaplanır. Bunun sonucunda az skora sahip nitelikler atılır. Daha sonra elde edilen nitelik altkümesi sınıflandırma için kullanılır [4]. Burada N_n, n adet niteliği, M ise bağımsız testleri ifade eder. Durdurma kriteri şunlardan birine göre belirlenebilir; Herhangi bir niteliğin eklenmesi ya da çıkarılması daha iyi bir nitelik altkümesi vermiyor ise veya istenilen nitelik sayısına ulaşılmış ise durdurma gerçekleşir.



a) Filtreleme yöntemi b) Sarmal yöntem

Şekil 1. Nitelik seçme yöntemleri
(Feature selection methods)

Biyoinformatik alanında sıklıkla kullanılan Fisher Korelasyon Skorlama, t-Skor ve Welch t-İstatistik aşağıda gösterildiği gibi hesaplanmaktadır:

Fisher korelasyon skorlama:

$$FKS(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ + \sigma_i^-} \quad (1)$$

t-Skor:

$$t(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{(n^+(\sigma_i^+)^2 + n^-(\sigma_i^-)^2)/(n^+ + n^-)}} \quad (2)$$

Welch t-istatistik:

$$WTS(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \quad (3)$$

Burada $\mu_i^+ ve \mu_i^-$ sınıfların aritmetik ortalaması, $\sigma_i^+ ve \sigma_i^-$ sınıfların varyansı ve n^+ ve n^- sınıf örnek sayılarıdır. Bu çalışmada eşitlik 1, 2 ve 3 kullanılarak nitelik ön seçimi gerçekleştirilmiştir. Böylece üç yöntemle ayrı ayrı elde edilen en yüksek skorlu genler birleştirilerek genetik algoritma için başlangıç veri kümesi belirlenmiştir.

Şekil 1(b) 'de görüldüğü gibi sarmal yöntemde, bağımsız testler yerine özel makine öğrenme metotları (Destek vektör makinesi, Karar ağaçları vb) kullanır. Nitelik seçme ölçüsü, sınıflandırıcının doğruluk oranıdır. Her bir iterasyonda belirli nitelik altkümesi için sınıflandırma sonucu elde edilir. Burada, *S*, sınıflandırıcıyı, *SD*, sınıflandırma doğruluk oranını ifade etmektedir. Durdurma kriteri filtreleme yönteminde olduğu gibi gerçekleştirilir [4]. Sarmal yöntemde, nitelik alt küme uzayı üstsel büyüdükçe sezgisel arama yöntemleri tercih edilir. Sarmal yapıda, model seçimi ile nitelik alt küme araması etkileşimlidir. Ancak filtre yöntemine göre aşırı eğitim riski bulunması ve sınıflandırma maliyetinin fazla olması en büyük dezavantajıdır.

2.2. Genetik Algoritma (Genetic Algorithm)

Genetik algoritma (GA), doğal seçim teoremine dayalı bir arama algoritmasıdır [17]. Genetik algoritma, geleneksel optimizasyon yöntemlerine göre parametre kümesini değil bunların kodlanmış biçimlerini kullanırlar [18]. Olasılık kurallarına göre çalışan genetik algoritmalar, yalnızca amaç fonksiyonuna gereksinim duyar. Çözüm uzayının tamamını değil belirli bir kısmını tararlar. Böylece, etkin arama yaparak çok daha kısa bir sürede çözüme ulaşırlar. Diğer bir önemli üstünlükleri ise çözümlerden oluşan popülâsyonu eş zamanlı incelemeleri ve böylelikle yerel en iyi çözümlere takılmamalarıdır [17]. GA, en iyiyi arama aracıdır. En uygun sonucu bulmak için doğal evrim ve seçim teoremini taklit eder. GA şu adımlarla gerçekleştirilir: kodlama, başlangıç popülasyonu, uygunluk fonksiyonu, genetik operatörler (seçim, çaprazlama ve mutasyon).

Genetik algoritma en iyiyi bulduğunda son bulmak zorundadır. En iyi çözüm %100 e varmakla olur; ancak bu teorik bir beklentiden ibarettir. Bu nedenle genetik algoritmanın sonlandırılması için belirli bir kriterin olması gerekir. Bu kriter, iterasyon sınırlaması, uzman görüşü, bulanık karar verme veya elde edilen hedef değerler arasındaki farkın önceden belirlenen bir miktardan daha düşük olması ile sağlanabilir [19].

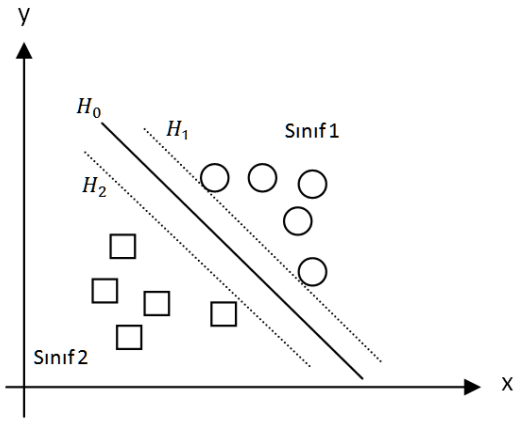
Genetik algoritma, nitelik seçme işlemlerinde sıklıkla ve başarılı bir şekilde kullanılmaktadır [20-22]. Ancak gen veri kümesinin yüksek boyutlu olması genetik algoritmanın başarısını doğrudan etkiler. Genetik

algoritmanın performansını arttırmak için gen veri kümesinin bir ön işlem ile daraltılması oldukça önemlidir. Başlangıç veri kümesinin belirlenmesinde filtreleme yöntemleri kullanılabilmektedir [23].

2.3. Destek Vektör Makinesi (Support Vector Machine)

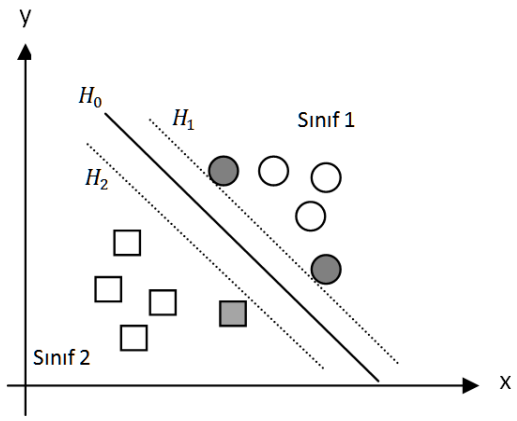
Makine öğrenme yöntemleri, büyük veri kümelerinde gizlenmiş ilişkileri ortaya çıkarmakta sıklıkla kullanılmaktadır. Destek vektör makinesi (DVM), performansıyla dikkat çeken makine öğrenme yöntemlerinde birisidir. DVM, veriyi birbirinden ayıran en uygun doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla sınıflandırma yapmaktadır [24]. Pek çok makine öğrenmesi veya istatistiksel veri analizi yönteminde olduğu gibi DVM de önceden var olan verilerden öğrenme yöntemini kullanır. DVM'ler danışmanlı veya yarı danışmanlı çalışabilen bir sınıflandırma yöntemidir. DVM'ler diğer makine öğrenme yöntemlerinin aksine yerel en iyiye takılma, ezberleme gibi problemleri aşabilmektedir.

Destek vektör makinesinin amacı her biri y={+1,-1} ile gösterilen sınıflardan birine ait olan n elemanlı bir veri kümesinde, sınıfları birbirinden ayıran altdüzlem bulmaktır. Şekil 2' de iki boyutlu düzlemde birbirinden doğrusal ayrılabilen iki sınıfa ait verilerin dağılımı ve bu verilerin birbirinden çok sayıda doğru ile ayrılabildiği görülmektedir. Çok boyutlu uzayda veriler hiper düzlemler ile ayrılabilir [24].



Şekil 2. İki boyutlu uzayda doğrusal ayrılabilen verilerin görünümü
(Two-dimensional space view of linearly separable data)

Verileri birbirinden ayıran en uygun hiper düzlem, birbirine en uzak iki hiper düzlem bulunarak elde edilir. Şekil 3' te birbirine en uzak H_1 ve H_2 düzlemleri arasından geçen H_0 hiper düzlemi, iki sınıfı birbirinden ayıran en uygun hiper düzlem olarak seçilmektedir. H_0, düzlemine *optimal ayırma hiper düzlemi* adı verilir. H_1 ve H_2 düzlemleri üzerindeki her bir veri *destek vektörü* olarak adlandırılır.



Şekil 3. İki sınıfı birbirinden ayıran en uygun hiper düzlem
(Optimal hyperplane that separates the two classes)

Şekil 3' te iki ayrı sınıf +1 ve -1 ile ifade edilir ise bu iki sınıfı birbirinden ayıran hiper düzlem $H_0 = w.x + b$ fonksiyonu ile gösterilebilir. Veri kümesinin ayrılabilir olduğu varsayılırsa n elemanlı bir kümede w ağırlık değeri ile bu fonksiyon hesaplanabilir. Şekil 3' te görüldüğü gibi bu veri kümesinde noktalardan bazıları w.x+b=1 durumunu sağlarken bazıları da w.x+b= -1 durumunu sağlamaktadır. H_0 fonksiyonu, düzlemin geçtiği noktalar cinsinden $\sum_{i=1}^{n} w_i x_i + b$ şeklinde de yazılabilir. Burada *W*, ağırlık vektörünü W={w_1, w_2, .. ., w_n} ve *n*, nitelik sayısını göstermektedir [24,25].

2.4. Alıcı İşletim Karakteristiği (Receiver Operating Characteristics)

Alıcı İşletim Karakteristiği (Receiver Operating Characteristics - ROC), sınıflandırıcı performansını test etmek için biyoinformatikte sıklıkla kullanılan bir yöntemdir [26]. İki ayrı sınıf içeren 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 4 ve 5 ile hesaplanabilir. Buradan elde edilen değerlere göre hata matrisi Şekil 4'te görüldüğü gibi olacaktır.

$$Doğru\ Pozitif\ Oranı = \frac{\mathrm{DP}}{\mathrm{DP} + \mathrm{YN}} \quad (4)$$

$$Yanlış\ Pozitif\ Oranı = \frac{\mathrm{YP}}{\mathrm{YP} + \mathrm{DN}} \quad (5)$$

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

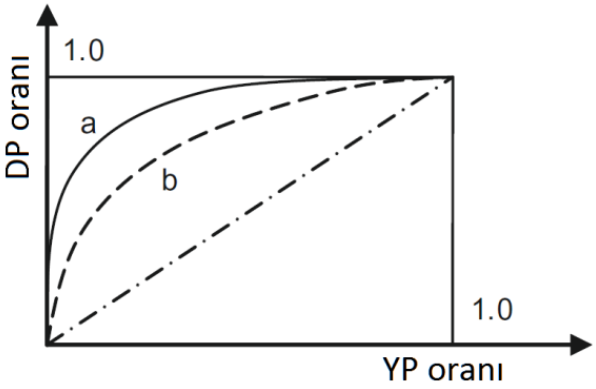
Şekil 4. Hata matrisi (Confusion matrix)

$$\text{Sınıflandırma doğruluğu} = \frac{DP + DN}{DP + YP + DN + YN} \quad (6)$$

$$\text{Duyarlılık (Sensitivity)} = DP \text{ oranı} \quad (7)$$

$$\text{Seçicilik (Specificity)} = 1 - YP \text{ oranı} \quad (8)$$

Eşitlik 6 ile sınıflandırma doğruluğu hesaplanırken, eşitlik 7 ve 8 kullanılarak, ROC eğrisi Şekil 5'te görüldüğü 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 5'te a sınıflandırıcısının b sınıflandırıcısına göre daha başarılı olduğu söylenebilir [4].



Şekil 5. ROC eğrisi (ROC curve)

2.5. Mikrodizi Veri kümesi (Microarray Dataset)

Bu çalışmada 97 meme kanseri hastasına ait mikrodizi verisi kullanılmıştır. Bu veri setinde yer alan 78 hastanın 44'ünde 5 yıl içinde uzak metastaz görülmezken (iyi prognoz), 34'ünde 5 yıl içinde uzak metastaz görülmüştür (kötü prognoz). Ayrıca 19 lenf node negatif meme kanseri hastasının 7 sinde 5 yıl içinde uzak metastaz görülmezken 12 sinde 5 yıl içinde uzak metastaz görüldüğü rapor edilmiştir. Mikrodizi veri kümesi Rosetta Inpharmatics den elde edilmiştir [27].

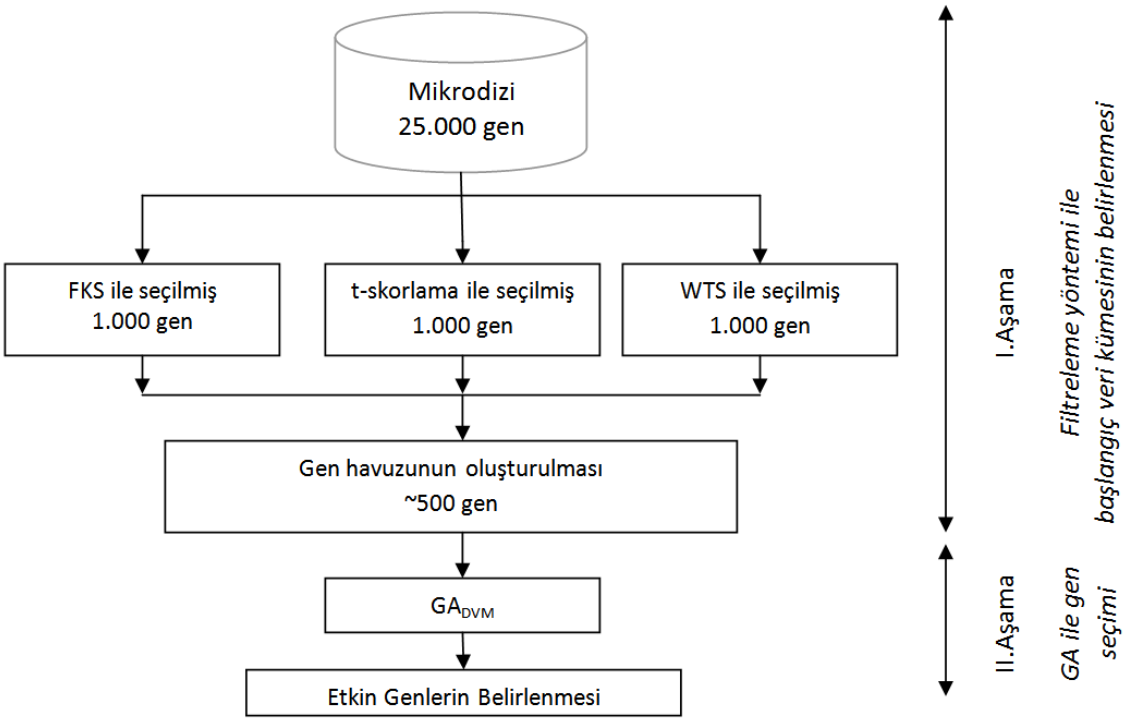
3. GA VE VERİ FÜZYONU TABANLI GEN SEÇİMİ (GENE SELECTION BASED ON GA AND DATA FUSION)

Gen ifade verilerinin büyük boyutlu olması GA' nın performansını olumsuz etkiler. Arama uzayının daraltılması, GA'nın performansını ve DVM'nın doğruluk oranını arttıracaktır. Bu amaçla önerilen algoritma iki aşamadan oluşmaktadır. İlk aşamada mevcut veri kümesi boyutu filtreleme ile azaltılmakta ve GA için başlangıç veri kümesi elde edilmektedir. İkinci aşamada elde edilen yeni veri kümesinden GA_{DVM} ile etkin genlerin tespiti gerçekleştirilmektedir. Şekil 6'da önerilen çalışmanın blok diyagramı görülmektedir.

I. Aşama: ***Filtreleme ile veri kümesinin indirgenmesi***

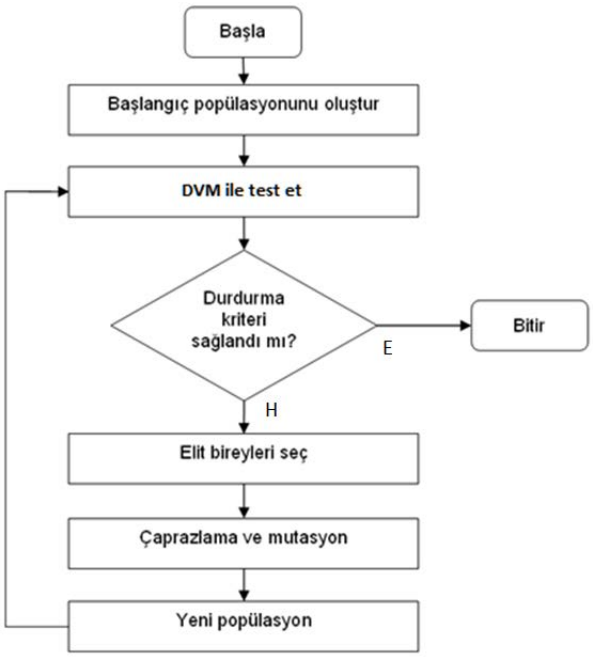
Gen ifade veri kümesinin yüksek boyutu genetik algoritmanın performansını olumsuz etkilemektedir. Arama uzayını daraltmak ve böylece genetik algoritmanın performansı artırmak için Şekil 6'da görüldüğü gibi I. Aşamada filtreleme yöntemleri kullanılarak veri kümesi boyutu azaltılmıştır. Elde edilen yeni veri kümesi genetik algoritma için başlangıç veri kümesi olarak belirlenmiştir. Bu amaçla Fisher Korelasyon Skorlama, t-Skor ve WTS kullanılmıştır. Her bir skorlama yöntemine göre en yüksek skora sahip ilk 1000 gen belirlenmiştir. Elde edilen üç ayrı veri kümesinden en yüksek skorlu genler birleştirilerek yaklaşık 500 gen bilgisi içeren başlangıç veri kümesi (Gen havuzu) oluşturulmuştur.

$$\text{Gen Havuzu} = \{FKS_{\text{Yüksek skor}}\} \cup \{\text{t-Skor}_{\text{Yüksek skor}}\} \cup \{WTS_{\text{Yüksek skor}}\}$$



Şekil 6. Önerilen çalışmanın blok diyagramı (Flowchart of the proposed method)

II. Aşama: ***Genetik algoritma*** *Destek vektör makinesi*
Filtreleme yöntemleri ile belirlenen yeni veri kümesi, meme kanseri hastalığında etkin rol alan genlerin bulunması amacıyla genetik algoritma için arama uzayı olacaktır. Şekil 7'de kullanılan GA_{DVM} algoritmasının akış diyagramı görülmektedir.



Şekil 7. GA_{DVM} akış diyagramı (Flowchart of the GA_{DVM})

Genetik algoritmanın dört önemli bileşeni vardır bunlar; genetik kodlama, başlangıç popülasyonu, uygunluk fonksiyonu, genetik operatörler (seçme, çaprazlama ve mutasyon) aşağıda gösterildiği gibi kodlanmıştır.

Genetik kodlama
Nitelik seçme işleminde belirlenen her bir nitelik alt kümesi birey olarak ifade edilmektedir. Her bir nitelik ise veri kümesinde bulunan gen bilgisidir. Bireylerin kodlanmasında ikili kodlama tercih edilmiştir. Bir (1) kodlanmış gen (nitelik) uygunluk fonksiyonuna gönderilirken, aksi halde hesaba katılmayacak anlamına gelmektedir.

Başlangıç popülasyonu
Popülasyon, bireylerden oluşan topluluktur. Her birey başlangıçta belirtilen gen sayısı kadar niteliğe sahiptir. Başlangıç popülasyonunu, rastgele gen değerlerine sahip bireylerden oluşmaktadır. Birey sayısının çok olması çalışma zamanını olumsuz etkilemektedir. Bu çalışmada başlangıç popülasyonu 200 olarak belirlenmiştir.

Uygunluk fonksiyonu
GA en yüksek başarıyı gösteren nitelik alt kümesini seçecektir. Bir veri setinde etkin nitelikler sınıflandırma başarısı da yüksek olacaktır. Uygunluk fonksiyonu DVM den elde edilen doğruluk oranı olarak belirlenmiştir. En iyi nitelik alt kümesi DVM den elde edilen doğruluk oranına bağlıdır. DVM için çekirdek fonksiyon, radyal tabanlı fonksiyon olarak belirlenmiştir

Genetik operatörler

1.Seçme ve elitizim

Rastgele belirlenen başlangıç popülasyonunda her bireyin DVM ile sınıflandırma başarı oranı belirlenir. En yüksek başarı oranına sahip bireyler, elit bireyler olarak yeni popülasyona eklenir. Elit bireyler başlangıç popülasyonunu oluşturan bireylerin %10'u olarak kabul edilmiştir. Geri kalan bireyler rulet tekerleği ile belirlenmiştir.

2.Çaprazlama ve mutasyon

Tek noktalı çaprazlama metodu kullanılmıştır. Rastgele belirlenen bir noktadan yine rastgele belirlenen iki bireyin genleri çaprazlanır. Geliştirilen uygulamada bireylerin mutasyon oranı sabit değildir. Başarı oranına göre her birey farklı oranda mutasyona tabii tutulmaktadır. Nispeten yüksek başarı gösteren birey küçük bir mutasyona uğratılırken, daha düşük başarı oranı gösteren bireyler yüksek mutasyona uğratılmıştır. Böylece GA'nın daha etkin çalışması sağlanmıştır.

4. DENEYSEL BULGULAR (EXPERIMENTAL RESULTS)

I. Aşama

Etkin genlerin belirlenmesi için elimizdeki gen ifade verilerinden en az beş örnekte iki kat daha fazla ifade olmuş ve p-değeri 0,001 in altında kalan anlamlı yaklaşık 5.000 gen bilgisi seçilmiştir. Daha sonra mikrodizi veri kümesinin boyutu filtreleme yöntemi ile azaltılmıştır. Bu amaçla FKS, t-skorlama ve WTS skorlama kullanılmıştır. Her üç skorlama yöntemi sonucunda ayrı ayrı en yüksek skora sahip ilk 1.000 gen belirlenmiştir.

Skor hesaplaması Eşitlik 1, 2 ve 3 kullanılarak gerçekleştirilmiştir. Buna göre FKS ile 0,956 – 3,669 skor değerleri aralığındaki genler, t-skor ile 1,875 - 7,399 skor değerleri aralığındaki genler ve WTS ile 8,172 - 31,4 skor değerleri aralığındaki genler seçilmiştir. Tablo 1'de ilk 5.000 ve ilk 1.000 gene ait skor ortalama ve standart sapma değerleri görülmektedir.

Tablo 1. FKS, t-Skor ve WTS ile elde edilen niteliklerin skor ortalama ve standart sapma değerleri (Standard deviations and means that obtained from FCS, t-Score and WTS)

		Tüm genler	İlk 1000 gen
FKS	μ	0,5906	1,4096
	σ	0,5088	0,4081
t-Skor	μ	1,1584	2,7641
	σ	0,9974	0,7943
WTS	μ	5,0168	12,0720
	σ	4,3780	3,5374

FKS, t-Skor ve WTS ile ayrı ayrı her bir niteliğin skoru hesaplanmıştır. Başlangıç veri kümesi belirlenirken, her bir filtreleme yönteminde de yüksek skora sahip nitelikler seçilmiştir. Böylece FKS, t-Skor ve WTS 'nin tek başına ayırt edemediği nitelikler de başlangıç veri kümesine dahil edilmiştir. Yeniden elde edilen başlangıç veri kümesi yaklaşık 500 nitelik (gen) bilgisi içermektedir. Tablo 2'de genetik algoritma için başlangıç veri kümesini oluşturan niteliklerin skor ortalama ve standart sapma değerleri görülmektedir. Elde edilen sonuçlar Tablo 1 ile karşılaştırıldığında skor ortalama ve standart sapma değerlerinin yükseldiği görülmüştür. Bu durum yüksek skorlu niteliklerin yanı sıra bazı düşük skorlu niteliklerin de başlangıç veri kümesine dahil edilmesinden kaynaklanmaktadır.

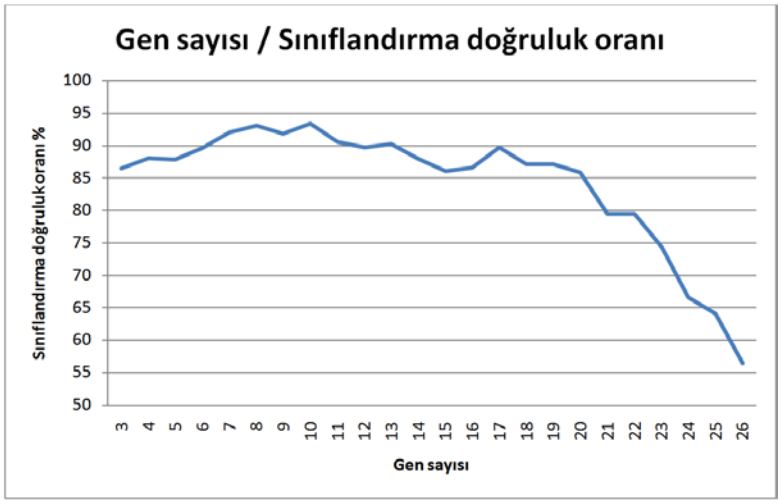
Tablo 2. Başlangıç veri kümesinde yer alan niteliklerin skor ortalama ve standart sapma değerleri (Means and standard deviations for initial dataset)

	FKS	t-Skor	WTS
μ	1,4404	2,8431	12,4370
σ	0,6263	1,2180	5,4087

II. Aşama

I.Aşamada genetik algoritma için başlangıç veri kümesi belirlenmişti. Bu veri kümesinde FKS, t-Skor ve WTS yöntemleri ile belirlenen ortak nitelikler bulunduğu gibi, her bir yöntemin ayrı ayrı tespit ettiği nitelikler de dahil edilmişti. Elde edilen yeni veri kümesi genetik algoritma için başlangıç veri kümesi olarak belirlenmiştir.

II. Aşamada, I. Aşamada belirlenen nitelik alt kümesi GA için başlangıç popülasyonu olarak kullanılmış, en iyi sınıflandırma doğruluğu gösteren nitelikler tespit edilmiştir. Her iterasyonda belirlenen nitelik alt kümesi için destek vektör makinesi ile sınıflandırma doğruluk oranını elde etmiştir. Başlangıç popülasyonu 200 belirlenerek, genetik algoritma 200, 300 ve 500 iterasyon için ayrı ayrı çalıştırılmış, 300 ve üstü iterasyonda çalışma zamanın yüksek olduğu ve ayrıca iterasyon sayısının daha fazla artmasının seçilen nitelik alt kümelerini değiştirmediği gözlemlenmiştir. Şekil 8'de GA_{DVM} ile belirlenen niteliklere ait sınıflandırma doğruluk oranları görülmektedir. Tek nitelik bilgisi ile sınıflandırma doğruluğunun çok düşük olduğu görülürken 3 ve üstü nitelik bilgisine sahip alt kümelerin bazılarında sınıflandırma doğruluk oranının %80 'nin üzerinde olduğu gözlemlenmiştir. Nitelik sayısı 20 ve üzerinde olduğunda sınıflandırma doğruluk oranının düştüğü gözlemlenmiştir. En yüksek sınıflandırma doğruluk oranı 10 nitelik (gen) ile tespit edilmiştir. %94,65 doğruluk oranı ile belirlenen 10 genin en yüksek doğrulukta sınıflandırma yaptığı belirlenmiştir.



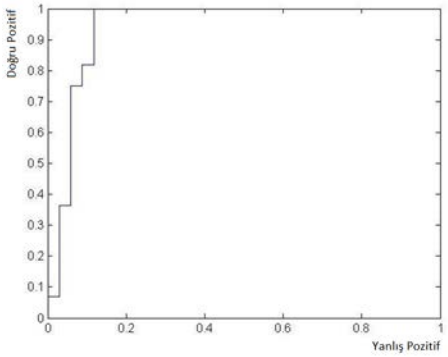
Şekil 8. Seçilen genler ve sınıflandırma doğruluk oranları (Classification accuracy for selected genes)

Tablo 3'te GA_{DVM} ile belirlenen 10 gene ait skor ortalama ve standart sapma değerleri görülmektedir. Sonuçlar Tablo 1 ve Tablo 2 ile karşılaştırıldığında, ortalama ve standart sapma değerlerinin düştüğü gözlemlenmiştir. Tablo 3'te filtreleme yöntemi ile belirlenen en yüksek skorlu niteliklerin, GA ile belirlenen niteliklerden farklı olduğu açıkça görülmektedir. Diğer bir ifadeyle GA ile belirlenen 10 nitelik, filtreleme yöntemlerinde en yüksek skora sahip nitelikler değildir. Ancak bu nitelikler ile en yüksek sınıflandırma doğruluğu elde edilmiştir.

Tablo 3. Belirlenen 10 gene ait skor ortalama ve standart sapma değerleri (Means and standard deviations for selected 10 genes)

		FKS	t-Skor	WTS
Belirlenen 10 gen	μ	1,2325	2,1710	9,9367
	σ	0,4982	0,8660	4,5598

Şekil 8'de görüldüğü gibi 10 gen ile en yüksek başarı oranı %94,65 elde edilmiştir. 26 gen ve üstünde sınıflandırma doğruluk oranının düştüğü gözlemlenmiş ve sonraki değerlerde de değişme görülmemiştir. Seçilen 10 gen için sınıflandırma başarısı 10-kat çapraz geçerleme ile test edilmiştir. Böylece daha güvenilir hata kestirimi gerçekleştirilmiştir. Şekil 9'da seçilen 10 gene göre sınıflandırıcının ROC eğrisi görülmektedir.



Şekil 9. ROC eğrisi (ROC curve)

Sınıflandırıcı başarısı ROC eğrileri ile de gösterilebilmektedir. Başarılı bir sınıflandırmada doğru pozitif (DP) oranının daha yüksek, yanlış pozitif (YP) oranının da daha düşük olması beklenir. Ayrıca eğri altında kalan alanın (Area Under the Curve-AUC) 1'e yakın olması sınıflandırma başarısının yüksek olduğu anlamına gelmektedir. Şekil 9'da da açıkça görüldüğü gibi sınıflandırıcının DP oranının yüksek çıkması ve ayrıca AUC değerinin 0,9512 elde edilmesi seçilen 10 genin sınıflandırma başarısının oldukça yüksek olduğunu göstermektedir.

5. SONUÇ VE DEĞERLENDİRME (CONCLUSION AND EVALUATION)

Günümüzde genetik çalışmalar her alanda artarak devam etmektedir. Hastalıkların temelinde yatan genetik faktörlerin belirlenmesi, teşhis ve tedaviye çok önemli katkılar sağlayabilir. Bu sebeple gen analizleri çok önemlidir. Ancak gen ifade verileri çok az örnekleme sahipken çok büyük miktarda gen bilgisi içerirler. Bu niteliklerin pek çoğu ilgisiz ya da gürültü olarak adlandırılan gen bilgileridir. İlgisiz genlerin atılması ya da etkin genlerin bulunması ciddi bir problemdir. Gen ifade verilerinin indirgenmesinde istatistiksel yöntemler çoğu kez başarısız olmaktadır.

Bu çalışmada, meme kanseri sınıflandırmasında etkin genlerin belirlenmesi amacıyla gen ifade verilerinden veri füzyonu ve GA_{DVM} tabanlı gen seçme işlemi gerçekleştirilmiştir. Belirlenen genler ile gerçekleştirilen sınıflandırma başarısı 10 kat çapraz doğrulama ile test edilmiştir.

Genetik algoritma nitelik seçme işlemlerinde sıklıkla kullanılmasına rağmen arama uzayının çok büyük olması genetik algoritmanın başarısını ve çalışma süresini olumsuz etkilemektedir. Bu sebeple arama uzayının daraltılması gerekmektedir. Bu amaçla, filtreleme yöntemleri ile gen ifade veri kümesi indirgenerek GA için başlangıç veri kümesi elde

edilmiştir. Daha önce yapılan nitelik seçme çalışmalarında görülmüştür ki genetik algoritma arama uzayında veri kümesi ya doğrudan kullanılmış ya da filtreleme yöntemlerinden biri tercih edilerek yeni arama uzayı elde edilmiştir. Yapılan çalışmada sadece bir filtreleme yöntemine bağlı kalınmamış, üç ayrı filtreleme yöntemi sonucunda yüksek skor elde edilen tüm genler yeni arama uzayına dahil edilmiştir. Böylece genetik algoritma performansı ve sınıflandırma başarısı da arttırılmıştır.

GA ile belirlenen genlerin, filtreleme yöntemi ile belirlenen genler arasında en yüksek skora sahip olan genler olmadığı görülmüştür. Bu durum, filtreleme yönteminin tek başına gen ifade verilerinde nitelik seçmede yeterli olamayacağını göstermektedir. Filtreleme yöntemleri her ne kadar en uygun alt veri kümesini vermeyi garanti etmese de, GA için başlangıç veri kümesini belirlemede bizim için önemli bir adım olmuştur.

Yapılan çalışma ile belirlenen 10 gen, meme kanseri hastalarında %94,65 doğruluk oranında sınıflandırma başarısı sağlamıştır. Sınıflandırma, 10 kat çapraz doğrulama ve ROC eğrisi ile sınıflandırma performansı test edilmiştir. AUC değeri 0,9512 ile sınıflandırma performansı oldukça başarılı çıkmıştır.

Tespit edilen 10 gen ile yüksek doğrulukta meme kanseri teşhisi yapılabilir. Ayrıca yeni gen çipleri tasarlanabilir. Daha az gen daha az maliyetli gen çipi tasarımı anlamına gelmektedir. Günümüzde kanser gibi ciddi hastalıkların tedavisinde ve yeni ilaç geliştirilmesinde hastalığa etken genlerin belirlenmesi üzerine yaygın çalışmalar yapılmaktadır. Önerilen yöntem farklı gen ifade veri kümelerinde ve farklı hastalıklarda gen seçimi için uygulanabilir.

KAYNAKLAR (REFERENCES)

1. Segal, E.,Wang, H. ve Koller, D., "Discovering Molecular Pathways From Protein Interaction And Gene Expression Data", **Bioinformatics**, Cilt 19, 264-272, 2003.
2. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., Trent, J., "Gene-Expression Profiles in Hereditary Breast Cancer", **The New England Journal of Medicine**, Cilt 344, 539-548, 2001.
3. van de Vijver, M.J, He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., Bernards, R., "A Gene-Expression Signature As A Predictor Of Survival In Breast Cancer", **The New England Journal of Medicine**, Cilt 347, 1999-2009, 2002.
4. Peng, Y., Wu, Z., Jiang, J., "A Novel Feature Selection Approach For Biomedical Data Classification", **Journal of Biomedical Informatics**, Cilt 43, 15-23, 2010.
5. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. ve Levine, A. J., "Broad Patterns Of Gene Expression Revealed By Clustering Analysis Of Tumor And Normal Colon Tissues Probed By Oligonucleotide Arrays", **Cell Biology**, Cilt 96, 6745-6750, 1999.
6. Dudoit, S., Fridlyand, J ve Speed, T., P., "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", **American Statistical Association**, Cilt 97, 77-87, 2002.
7. Horng, J.T., Wub, L.C., Liu, B.J., Kuo, J.L., Kuo, W.H., Zhang, J.J., "An Expert System To Classify Microarray Gene Expression Data Using Gene Selection By Decision Tree", **Expert Systems with Applications**, Cilt 36, 9072-9081, 2009.
8. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K., Mewes, H., "Gene Selection From Microarray Data For Cancer Classification-A Machine Learning Approach", **Computational Biology and Chemistry**, Cilt 29, 37-46, 2005.
9. Uriarte, R., D. ve Andrés, S.,A., "Gene Selection And Classification Of Microarray Data Using Random Forest**", BMC Bioinformatics**, Cilt 7, 1-13, 2006.
10. Li, L., Jiang, W., Li, X., Moser, K.L., Guo, Z., Du, L., Wang, Q., Topol, E., Wang, Q., Rao, S., "A Robust Hybrid Between Genetic Algorithm And Support Vector Machine For Extracting An Optimal Feature Gene Subset", **Genomics**, Cilt 85, 16-23, 2005.
11. Antonov, A.V., Tetko, I.V., Kosykh, D., Surmeli, D., Mewes, H.W., "Exploiting Scale-Free Information From Expression Data For Cancer Classification", **Computational Biology and Chemistry**, Cilt 29, 288–293, 2005.
12. Saeys, Y., Inza, I. ve Larranaga, P., "A Review Of Feature Selection Techniques In Bioinformatics", **Bioınformatics**, Cilt 23, 2507–2517, 2007.
13. Maldonado, S., Weber, R., "A Wrapper Method For Feature Selection Using Support Vector Machines", **Information Sciences**, Cilt 179, 2208–2217, 2009.
14. Huang, C., L. Ve Wang, C., J., "A GA-Based Feature Selection And Parameters Optimization For Support Vector Machines", **Expert Systems with Applications**, Cilt 31, 231–240, 2006.
15. Liu, H. ve Motoda, H., "Computational Methods Of Feature Selection", **Chapman & Hall/CRC**, Taylor & Francis Group 6000 Broken Sound Parkway NW, 26-27, 2008.

16. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.J., "Filter Versus Wrapper Gene Selection Approaches In DNA Microarray Domains", **Artificial Intelligence in Medicine**, Cilt 31, 91-103, 2004.

17. Goldberg, E.D., "Genetic Algorithms in Search, Optimization, and Machine Learning", **Addison-Wesley Longman, Inc**., New York, 1989.

18. Michalewicz, Z., "Genetic Algorithms + Data Structures = Evolution Programs", **Springer**, Berlin, 1992.

19. Şen, Z., "Genetik Algoritma Ve En İyileme Yöntemleri", **Su Vakfı Yayınları**, 2004

20. Lee, C., P., Lin, W., S., Chen, Y., M. ve Kuo B., J., "Gene Selection And Sample Classification On Microarray Data Based On Adaptive Genetic Algorithm/K-Nearest Neighbor Method", **Expert Systems with Applications**, Cilt 38, 4661–4667, 2011.

21. Li, L., Weinberg, C., R., Darden, T.A. ve Pedersen, L.G., "Gene Selection For Sample Classification Based On Gene Expression Data: Study Of Sensitivity To Choice Of Parameters Of The GA/KNN Method", **Bioinformatics**, Cilt. 17, 1131–1142, 2001.

22. Hernandez, J.,C., Duval, D., ve Hao, J., K., "A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data", EvoBIO'07 Proceedings of the 5th European conference on Evolutionary computation, machine learning and data mining in bioinformatics, **Springer-Verlag Berlin**, Heidelberg 2007.

23. Lee, C., P. ve Leu, Y., "A Novel Hybrid Feature Selection Method For Microarray Data Analysis", **Applied Soft Computing**, Cilt 11, 208–213, 2011.

24. Özkan, Y., "Veri Madenciliği Yöntemleri", **Papatya Yayıncılık**, İstanbul, 2008.

25. Cristianini, N. ve Taylor, J.S., "An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods", **Cambridge Univ. Press**, 2000.

26. 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.

27. http://bioinformatics.nki.nl/data.php

28. Veer,L.J,V.,Dai,H., Vijver, M.J.V., He, Y.D., Hart, A. A. M., Mao, M., Peterse, H.L., Kooy, K.V.D., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., "Gene Expression Profiling Predicts Clinical Outcome Of Breast Cancer", **Nature**, Cilt 415, 530-536, 2002.