Tıbbi Tahminde Alternatif Bir Yaklaşım: Destek Vektör Makineleri

An Alternative Approach in Medical Diagnosis: Support Vector Machines

Özge YILMAZ AKŞEHİRLİ,[a]
Handan ANKARALI,[a]
Duygu AYDIN,[a]
Özge SARAÇLI[b]

[a]Biyoistatistik ve Tıbbi Bilişim AD, Düzce Üniversitesi Tıp Fakültesi, Düzce
[b]Psikiyatri AD, Bülent Ecevit Üniversitesi Tıp Fakültesi, Zonguldak

Geliş Tarihi/*Received:* 13.02.2012
Kabul Tarihi/*Accepted:* 05.03.2012

Yazışma Adresi/*Correspondence:*
Handan ANKARALI
Düzce Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Düzce, TÜRKİYE/TURKEY
hankarali@yahoo.com

ÖZET Amaç: Bu çalışma, birçok alanda sıklıkla kullanılan destek vektör makinelerinin (DVM) tıbbi araştırmalarda kullanımına yönelik bir uygulama olarak düşünülmüş ve tıbbi bir çalışma verisi kullanılarak DVM yönteminin veriyi doğru sınıflama başarısını ortaya koymak amaçlanmıştır. **Gereç ve Yöntemler:** Çalışmamızda, verileri sınıflandırmak veya tahmin yapmak amacıyla kullanılan, eğiticili (supervised) bir makine öğrenmesi yöntemi olan DVM kullanılmıştır. Burada, doğrusal olmayan ilişkiler için iki sınıflı DVM yönteminin bir uygulaması yapılmıştır. DVM'nin temelini, verilerin bir düzlem veya hiper düzlem ile ayrılarak sınıflandırılması işlemi oluşturmakta ve DVM bu işlemi, iki sınıf arasındaki marjini maksimum yaparak gerçekleştirmektedir. Bu şekilde veri eğitildikten sonra, DVM yeni gelen veriyi doğru sınıflamayı amaçlamaktadır. Tıpta DVM özellikle, kanser morfolojisinde, tedavi başarısının ve ilgili genin belirlenmesinde, çeşitli hastalıkların teşhisinde kullanılmaktadır. Araştırmanın uygulama bölümünde, Zonguldak Karaelmas Üniversitesi Tıp Fakültesi psikiyatri polikliniğine 1-31 Ocak 2011 tarihleri arasında gece yeme sendromu şikâyetiyle başvuran 433 hastaya ilişkin bilgiler kullanılmıştır. **Bulgular:** Kullanılan 17 değişken için tanımlayıcı istatistikler elde edilmiş ve univariate analizlerden elde edilen sonuçlara göre, GYA, BSQ ve SCL puanları, medeni durum, sigara kullanımı ve psikolojik tanı değişkenlerinin gece yeme sendromu tanısı koymada tek başına etkileri olduğu sonucuna varılmıştır. Doğrusal olmayan destek vektör makineleri kullanılarak elde edilen sonuçlar incelendiğinde, eğitim ve test verileri için doğruluk ve ROC eğrisi altında kalan alanlara bakılarak, modelin tanı koyma başarısının oldukça iyi derecede olduğu görülmüştür. **Sonuç:** DVM yöntemi, istatistiksel öğrenme teorisine dayanan yeni ve doğrusal olmayan, karmaşık yapıya sahip verileri sınıflamada etkin bir yöntemdir ve bu nedenle birçok sınıflama yöntemine tercih edilebilmektedir.

Anahtar Kelimeler: Sınıflandırma; veri madenciliği; destek vektör makinaları; gece-yeme sendromu

ABSTRACT Objective: This study was intended as an application in medical research of support vector machines (SVM) that is commonly used in many domains and aimed to determine the success of correct classification of SVM method by using a medical data set. **Material and Methods:** In this study, supervised SVM that is used for classification or regression was performed. Here an application of two class SVM was performed for nonlinear relationships. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, since in general the larger the margin the lower the generalization error of the classifier. So, SVM training algorithm builds a model that assigns new examples into one category or the other. In medicine, SVM is used for cancer morphology, identifying success of treatment and related gene, diagnosing various diseases. In application of the study, informations about 433 patient who were refer to the outpatient department of Zonguldak Karaelmas University Faculty of Medicine between date of 1-31 January 2011 for complaints of night eating syndrome were used. **Results:** Descriptive statistics for 17 variables were calculated and according to the results obtained from univariate analysis, GYA, BSQ, SCL scores, marital status, smoking status, psychological diagnosis are effective variables for diagnosis of night eating syndrome. When the results obtained by using nonlinear SVM were examined, the success of diagnosis of the model was observed very well according to accuracy and area under ROC curve. **Conclusion:** SVM method based on statistical learning theory is a new and effective method for classification nonlinear data with complex structure. For this reason it can prefer to many classification method.

Key Words: Classification; data mining; support vector machine; night-eating syndrome

Turkiye Klinikleri J Biostat 2013;5(1):19-28

Sağlık alanında çok çeşitli araştırmalar yapılmakta olup bu alan önemli bir veri kaynağıdır. Bilgisayar teknolojisindeki ilerleme ve bilgisayar donanımlarının ucuzlamasıyla, yararlı bilgilerin ortaya çıkarılacağı çok büyük boyutlu verilerin kayıt altına alınması ve saklanması olanaklı hale gelmiştir. Bu verilerden elde edilen bilgiler, çeşitli hastalıkların sınıflandırılması, tanımlanması, tanı ve tedavisinde ve aynı zamanda hastalıklara karşı koruyucu önlemler almada kullanılır.[1] Çeşitli makine öğrenmesi yöntemleri ile bu verilerden faydalı bilgileri ya da gizli kalmış ilişkileri ortaya çıkarmak mümkündür. Makine öğrenmesi metotları, geçmişteki verileri kullanarak veriye en uygun modeli bulmaya çalışırlar ve yeni gelen verileri de bu modele göre analiz ederler.[1] Makine öğrenmesi uygulamalarından biri olan veri madenciliği de, istatistiksel yöntemler ile çeşitli bilgisayar algoritmalarını kullanarak, veri tabanlarındaki veriden gerekli bilgi keşfini sağlamak için geliştirilmiş yöntemlerden birisidir.

Veri madenciliği yöntemlerinden biri olan destek vektör makineleri (DVM), veriyi sınıflandırmak (DVS) veya tahmin yapmak (DVR) amacıyla kullanılan, eğiticili (supervised) bir makine öğrenmesi yöntemidir.

Günümüzde, DVM'nin birçok dünya problemine uyarlanabilir olması, DVM yöntemine olan ilgiyi arttırmakta ve bununla birlikte, bu yöntemle yapılan çalışmalar her alanda ağırlık kazanmaktadır. DVM, görüntü ve metin sınıflandırma, nesne tanıma, el yazısı tanıma, ses tanıma ve yüz tanıma gibi çeşitli örüntü tanıma uygulamalarında sıkça kullanılmaktadır.[2] DVM, aynı zamanda biyolojik uygulamalarda da yükselen bir başarı göstermektedir.[3] Tıpta ise kanser morfolojisinde, tedavi başarısının ve ilgili genin belirlenmesinde, çeşitli hastalıkların teşhisinde kullanılmaktadır.[4]

DVM pratikte daha çok sınıflama amacıyla kullanılmaktadır ve sağlık alanı araştırmalarında bir tanı yöntemi olarak tercih edilebilir. DVM yardımıyla sınıflamada, en az iki grup sahip oldukları özellikler bakımından doğrusal veya doğrusal olmayan modellerle ayırt edilebilmektedir. Sağlık araştırmalarında genellikle sağlıklı kontrol ve hasta grubu olarak iki grup ayrımı yapılmak istenir.

Bu çalışmanın amacı, DVM yönteminin teorik özelliklerini tanımlamak ve sağlık alanında kullanımını yaygınlaştırmaktır.

GEREÇ VE YÖNTEMLER

VERİLER

Çalışmada kullanılan veri seti, Zonguldak Karaelmas Üniversitesi Tıp Fakültesi psikiyatri polikliniğine 1-31 Ocak 2011 tarihleri arasında gece yeme sendromu şikayetiyle ayaktan başvuran ve çalışmaya katılmayı kabul eden 433 hastaya ait bilgileri içermektedir. Bu veriler amaca uygun olarak çeşitli klasik istatistik yöntemlerle değerlendirilmiş ve yazarları tarafından yayınlanmak amacıyla dergiye gönderilmiştir. Verilerin kullanımı için bu çalışmada ilk isim olarak bulunan yazardan izin alınmıştır. Bu hastalardan, yaş, cinsiyet, eğitim yılı, kardeş sayısı, medeni durum, çocuk sayısı gibi demografik özellikler, bel çevresi, kalça çevresi, beden kitle indeksi, fiziksel hastalık varlığı, psikolojik hastalık varlığı gibi çeşitli değişkenler sorgulanmıştır. Hastalara ayrıca, gece yeme anketi (GYA), beden şekli anketi (BSQ), semptom tarama listesi (SCL-90) ölçekleri de uygulanmıştır.

Psikiyatri kliniğine başvuran 433 hastanın 97'si klinik görüşmelerle gece yeme sendromu tanısı almış, geriye kalan 336 kişi ise gece yeme sendromu tanısı almamıştır.

Çalışmanın uygulama bölümünde, yukarıda tanımlanan sosyo-demografik özellikler, bazı fiziksel ölçümler ve psikolojik durumları gösteren ölçek puanları dikkate alınarak, gece yeme alışkanlığı tanısı konulması amaçlanmaktadır. Böylece klinik olarak konulan tanı ile bu özelliklere göre konulan tanının uyumu ölçülecek ve söz konusu özelliklerin tanı başarıları belirlenecektir. Tanı başarısının yüksek olması durumunda klinik tanıya gerek duyulmadan geliştirilen bir model yardımıyla başarılı bir tanı konulabilecektir. Bu amacı gerçekleştirmek için 17 doğal nitelik yardımıyla doğrusal olmayan DVM yöntemi kullanılmıştır.

Doğrusal olmayan DVM yönteminin kullanılabilmesi için gerekli olan çekirdek fonksiyonlardan literatürde sıkça kullanılan ve diğer çekirdek fonksiyonlara göre bazı üstünlükleri çeşitli çalış-

malarla kanıtlanmış olan radyal tabanlı fonksiyon seçilmiştir. Çekirdek fonksiyon parametrelerinin tahmininde ise ızgara arama (grid search) metodu kullanılmıştır.

Verinin analizinde, doğrusal olmayan DVM yöntemi yardımıyla gece yeme alışkanlığı varlığının tahmin edilmesi için DTREG paket programı kullanılmıştır.

DESTEK VEKTÖR MAKİNELERİ

DVM, 1960'lı yılların sonunda Vladimir Vapnik ve Alexey Chervonenkis tarafından geliştirilmiş, temel olarak istatistiksel öğrenme teorisine dayanan bir makine öğrenmesi yöntemidir. DVM metodu, son yıllarda özellikle veri madenciliğinde değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki sınıflama problemleri için sıklıkla kullanılmaktadır. Bu metot, temelde iki sınıflı problemlerin çözümünde doğrusal bir sınıflayıcı olarak düşünülmüş, daha sonra doğrusal olarak ayrılamayan veya çok sınıflı sınıflama problemlerinin çözümüne de genelleştirilerek, bu problemlerin çözümünde de yaygın olarak kullanılmaya başlanmıştır.

DVM algoritması öğrenme teorisinin ve pratiğinin kesiştiği bir uygulamadır. Gerçek dünya uygulamaları, teorik olarak çözülmesi zor ve karmaşık olan uygulamalardır. DVM algoritması bu iki zorluğu da basitçe kaldırabilir ve karmaşık modellere de çözüm getirebilir.

DVM yöntemi, istatistiksel öğrenme teorisinde iyi şekilde kurulmuş bir teoriye sahiptir ve sınıflandırma ile regresyon problemlerine çözüm için uygundur. Vapnik'in teorisi eğitim kümesindeki hata ile VC boyutuna göre ifade edilen hipotez uzayının karmaşıklığının her ikisini de küçükleyen çözümün bulunduğunu göstermektir.

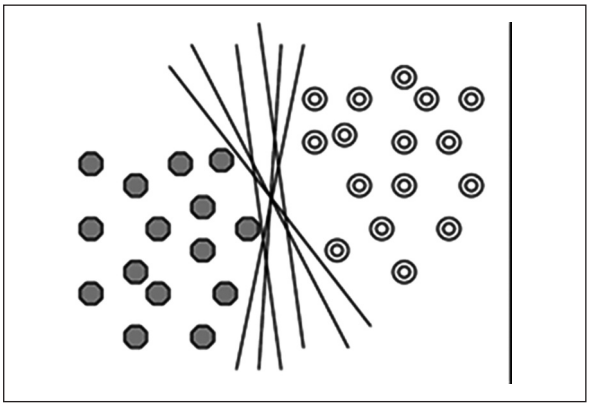
DVM eğitim esnasında gözlenmemiş yeni verileri de sorunsuz olarak sınıflandırabilmektedir ve bu durum DVM'nin genelleştirebilme yeteneğini göstermektedir. Genelleştirebilme özelliği DVM'yi diğer tekniklere göre (YSA, karar ağacı vs..) iyi bir alternatif yapmaktadır.

DVM'nin temelini, verilerin bir düzlem veya hiper düzlem ile ayrılarak sınıflandırılması işlemi oluşturmaktadır. Yani, iki sınıfa ait verileri ayırabilecek en uygun düzlemi veya hiper düzlemi belirlemektir. Doğrusal olarak ayrılabilen verileri, ait oldukları boyutta bir düzlem ile ayırabilmek mümkünken, doğrusal olarak ayrılamayan verilerin ait oldukları boyuttan daha yüksek boyutlu bir uzaya taşınarak, burada bir hiper düzlem ile ayırmak mümkün olacaktır. DVM, doğrusal olarak ayrılabilen veriler söz konusu olduğunda, verileri ayırabilecek sonsuz sayıdaki doğru içerisinden marjini en yüksek yapacak olan doğruyu seçmeyi hedeflemektedir. Doğrusal olarak ayrılamayan verilerin olduğu durumda ise DVM, bir haritalama yöntemi ile orijinal veriyi daha yüksek boyutlu bir uzaya taşır ve burada verileri sınıflandırmak için optimum olabilecek doğrusal ayırıcı hiper düzlemi bulmaya çalışır.[3,5]

DVM literatüründe, tahmin edici veya bağımsız (predictor) değişkene *doğal nitelik (attribute)*, optimum hiperdüzlemi belirlemek için kullanılan dönüştürülmüş doğal niteliğe *belirleyici nitelik (feature)* ve bir deneği (gözlemi) tanımlayan belirleyici nitelik setine ise, *vektör* denilmektedir.

VERİLERİN TAMAMININ DOĞRUSAL OLARAK AYRILABİLDİĞİ DURUM (SERT MARJİN)

DVM'lerin bu şekli sadece doğrusal olarak ayrılabilen belirleyici nitelik uzayı için geçerlidir ancak gerçek hayatta bu tip problemlerle nadiren karşılaşılır. Doğrusal olarak ayrılabilen sınıflama problemleri, DVM'nin temelini oluşturur ve tanımı çok daha gelişmiş sistemleri anlamak için önemlidir (Şekil 1).[6]



ŞEKİL 1: Doğrusal olarak ayrılabilen iki sınıflı sınıflama problemi.

Şekil 1'de gösterilen iki sınıflı veriler doğrusaldır ve bu verileri birbirinden direkt olarak ayırabilen birçok hiperdüzlem (doğru) çizilebilmektedir.[5] Ancak DVM'nin amacı, bilinmeyen veri seti ile karşılaştığında sınıflama hatasını en küçük yapacak hiperdüzlemi seçmektir. Bu hiperdüzlem, iki örnek grubuna eş uzaklıkta olacaktır. Bunun için maksimum marjinli hiperdüzlem tekniği önerilmiştir.[7]

Şekil 1'deki hiperdüzlemler;

$$< w.x > + b = 0 \qquad (1)$$

şeklinde formüle edilir.

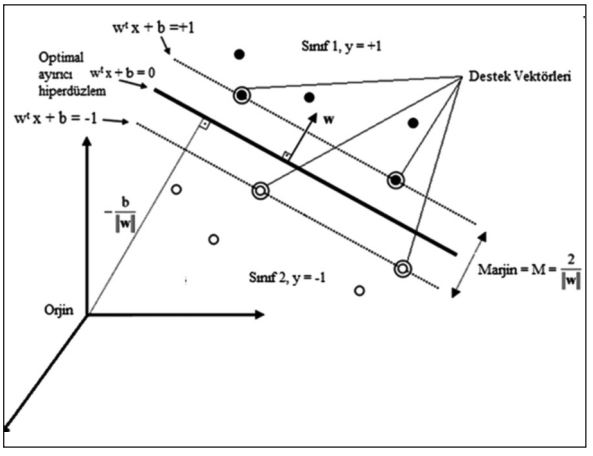
Formülde *w* hiperdüzlemin normalini (ya da ağırlık vektörü) ve *b* yanlılığı (bias) gösteren bir değerdir. *x* ise $<w.x>+b= 0$ hiperdüzlemi üzerinde herhangi bir noktadır. Burada, $<w.x>$ bir iç çarpımı göstermektedir ve $<w.x> = w^t x$ şeklinde ifade edilebilir.

Doğrusal olarak ayrılabilen iki sınıflı bir sınıflandırma probleminde DVM'nin eğitimi için k sayıda örnekten oluşan eğitim seti; $\{x_i,y_i\}$ $i=1,2,...,k$ için, sınıflar $y_i \in \{-1, +1\}$ ve girdi vektörü (doğal nitelikler) $x_i \in R^d$ olacak şekilde bir veri seti göz önüne alınsın. Özellik vektörlerinin ayrılabilir olduğu ve doğrusal bir karar sınırı tarafından ayrılabildiği varsayılmaktadır. Tanımlanan veri kümesi için, iki veri sınıfını ayırabilen, hiperdüzlemler kümesi bulunmaktadır bulunmaktadır.[8] Doğrusal ayrılabilir veriler için, veri kümesini verilen etiketlere göre bir hiperdüzlemle ayırıp, aynı sınıfa ait bütün veri noktalarını hiperdüzlemin aynı tarafında bırakmak mümkündür (Şekil 2).

$\frac{|b|}{||w||}$: hiperdüzlemden orjine olan dik uzaklık,

$||w||$: w'nun Öklid normu olarak ifade edilir.

Şekil 2'de görülen, sınırı maksimuma çıkararak en uygun ayrımı yapan hiperdüzlem "optimum ayırıcı hiperdüzlem" ve sınır genişliğini belirleyen noktalar ise "destek vektörleri (dv)" olarak adlandırılır. Optimum hiperdüzlemin belirlenebilmesi için bu düzleme paralel ve sınırlarını oluşturacak iki hiperdüzlemin belirlenmesi gerekir.[9] Şekil 2'de kesikli çizgilerle gösterilen ve ayırıcı hiperdüzleme paralel olarak çizilmiş eşit uzaklıkta iki hiperdüzlem bulunmaktadır. Bu iki hiperdüzlem arasındaki



ŞEKİL 2: Verilerin tamamının ayrılabildiği durum için doğrusal ayırıcı hiperdüzlem.

uzaklığa "marjin" adı verilmektedir. Bu hiperdüzlemlerin fonksiyon gösterimleri aşağıdaki gibidir:

$$w^t x^+ + b = +1, \qquad y_i = +1 \text{ için} \qquad (2)$$

$$w^t x^- + b = -1, \qquad y_i = -1 \text{ için} \qquad (3)$$

Bu iki eşitlik birbirinden çıkarılırsa, $w^t (x^+ - x^-) = 2$ eşitliği elde edilir. Bu eşitliğin her iki tarafı $||w||$'ye bölünürse; $\frac{w^t (x^+ - x^-)}{||w||} = \frac{2}{||w||}$ elde edilir. Burada $||w|| = \sqrt{\sum_{i=1}^{n} w_i^2}$, *w*'nin Öklid normudur ve dik uzaklıkların hesabı için kullanılmaktadır. $\frac{w^t}{||w||}$ ise, birim vektördür (uzunluğu 1 olan vektör). Böylece marjin uzunluğu:

$$M = ||x^+ - x^-|| \frac{2}{||w||} \qquad (4)$$

olarak bulunur.

Ayırıcı hiperdüzlem, veri örneklerinin ayrılmasını tanımlayan aşağıdaki koşulları yerine getirir:

$$w^t x_i + b \geq 1, \qquad y_i = +1 \text{ için} \qquad (5)$$

$$w^t x_i + b \leq -1, \qquad y_i = -1 \text{ için} \qquad (6)$$

Formül (5) ve (6) tek formül olarak ifade edilecek olursa:

$$y_i(w^t x_i + b) \geq 1 \qquad i = 1, 2, \ldots\ldots, n \qquad (7)$$

Verilen eğitim verisi için tüm ayırıcı hiperdüzlemler bu formda gösterilebilir.

(7) eşitsizliğini sağlayan hiperdüzlemin iki tarafındaki en yakın örneklere olan dik uzaklıkları toplamı marjindir ve optimum ayırıcı hiperdüzlem, marjini maksimum yapan hiperdüzlemdir. Böylece, optimum ayırıcı hiperdüzlemin bulunması problemi, Formül (4)'te verilen $M = \frac{2}{\|w\|}$ marjinini maksimum yapan *w* değerinin bulunması işlemine dönüşmüş olur. $\frac{2}{\|w\|}$ değerini maksimum yapmak için, $||w||$ değerinin, dolayısıyla $||w||^2$ değerinin minimize edilmesi gerekmektedir. Burada $||w||^2 = w^t w$'dir. Bu durumda, en iyi ayırıcı düzlemi bulmak için, aşağıdaki denklemlerin çözümü gerekir:

$$\min_{wb} \frac{1}{2}\|w\|^2 \quad (8)$$

$$y_i(w^t x_i + b) \geq 1, \quad \forall_i \quad (9)$$

Burada Formül (8) çözülecek problem (nesne fonksiyonu) ve Formül (9) problemin çözümü sırasında kullanılan koşul yani eşitsizlik kısıtıdır. (8)'deki ifade ikinci dereceden eşitsizlik kısıtlı bir doğrusal olmayan optimizasyon problemidir. Bu optimizasyon problemi Lagrange çarpanları yöntemi ile çözüldüğünde, pozitif Lagrange çarpanları olan α_i'ler kullanılarak dönüştürülen yeni optimizasyon problemi aşağıdaki gibidir:

$$L_p = \frac{1}{2} w^t w - \sum_{i=1}^{n} \alpha_i y_i (w^t x_i + b) + \sum_{i=1}^{n} \alpha_i \quad (10)$$

Bu Lagrangian formülü birincil (primal) değişkenler *w* ve *b* bakımından minimumlaştırılmalı, ikincil (dual) değişkenler bakımından maksimumlaştırılmalıdır.

Uygulamada, Formül (10)'da görülen primal problemi çözmek yerine, yaygın olarak dual (Wolfe dual) karesel optimizasyon problemi çözülmektedir ve (8) ile aynı sonucu vermektedir.

Formül (10)'da ifade edilen karmaşık formülasyonun çözülmesi için, formüldeki *w* ve *b* parametrelerinin sadece α_i parametresiyle ifade edilmesini sağlayacak olan Karush-Kuhn-Tucker (K.K.T.) koşulları olarak bilinen yöntem kullanılır ve bu durumda, Formül (10) sadece α_i Lagrange çarpanlarına göre maksimumlaştırılması istenen bir dual probleme dönüştürülür. K.K.T. koşulları ile çözüm sağlamak için öncelikle Formül (10)'un *w* ve *b*'ye göre türevleri alınır:

$$\frac{\partial}{\partial b} L_p = 0 \Rightarrow \sum \alpha_i y_i = 0 \qquad \alpha_i \geq 0 \quad (11)$$

$$\frac{\partial}{\partial w} L_p = 0 \Rightarrow w = \sum \alpha_i y_i x_i \qquad \alpha_i \geq 0 \quad (12)$$

w ve *b* parametrelerini bulmak için ulaşılan bu formüller, bilinmeyen Lagrange çarpanlarını (α_i) içerdiği için halen çözüm üretmemektedir. Çözüm için *(xi,yi)* noktasında Formül (9)'u eşitlik haline dönüştürecek 0'dan farklı α_i'leri sağlamalıdır:

$$\alpha_i [y_i < x_i, w > + b) - 1)] = 0 \qquad i = 1, \ldots \ldots \ldots, m \quad (13)$$

α_i'lerin 0 olmadığı yerlerde eşitliği sağlayan *(xi,yi)* noktaları destek vektörleridir ve ayırıcı vektöre paralel olan marjin doğrusu üzerinde yer alırlar.

(11) ve (12) ile ifade edilen koşullar Formül (10)'de yerine yazılırsa, aşağıdaki dual problem elde edilir:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i{}^t x_j, \qquad \alpha_i \geq 0 \quad \forall_i \quad (14)$$

Artık *w* ve *b* parametreleri için çözüm üretebilecek Formül (12) ve Formül (13) kullanılarak, sınıfları ayıracak karar fonksiyonunu belirlenebilir:

$$f(x) = (\sum \alpha_i y_i x_i \,.\, x) + b \quad (15)$$

Karar fonksiyonu bulunduktan sonra, yeni gelen bir örneğin hangi sınıfa ait olacağına aşağıdaki eşitsizlikler yardımıyla karar verilir:

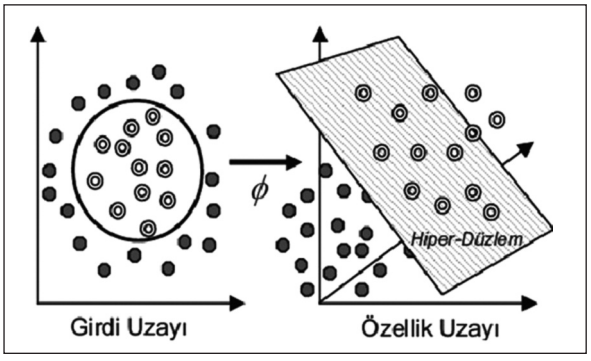
$$f(x) > 0 \Rightarrow sınıf\ 1 \quad (16)$$

$$f(x) < 0 \Rightarrow sınıf\ 2 \quad (17)$$

DOĞRUSAL OLMAYAN DESTEK VEKTÖR MAKİNELERİ İLE SINIFLAMA

Pratikte verilerin doğrusal olarak ayrılabildiği durumlarla pek karşılaşılmamaktadır ve bu durumda, pratik uygulamaların çoğu, doğrusal DVM'lerle çözülememektedir. DVM'ler böyle doğrusal olmayan problemlerle karşılaştığında, orijinal verilerden sınıflandırma özelliklerini çıkarmak için, doğrusal olmayan haritalama (mapping) yaparak, verileri n boyutlu orjinal girdi uzayından daha yüksek boyuta sahip belirleyici nitelik (feature) uzayına taşır (Şekil 3).

$$x \in R^n \rightarrow \Phi(x) \in R^f \quad (18)$$



ŞEKİL 3: Doğrusal olarak ayrılamayan verilerin farklı boyutlardaki uzaylara aktarılması.

DVM daha sonra, belirleyici nitelik uzayında maksimum marjini bulmak için doğrusal sınıflandırma kuralını öğrenir. Sınıflandırma kuralı, belirleyici nitelik uzayında doğrusal olması gerçeğine karşın, orijinal girdi uzayına izdüşüm yapıldığında doğrusal değildir.[10]

Doğrusal olmayan DVM, verilerin taşındığı bu yeni boyutta doğrusal DVM gibi çalışarak verileri ayıracak optimum hiperdüzlemi arar. Dönüştürme işlemi için kullanılacak olan fonksiyon *Φ(x)* olarak belirlensin. Bu durumda doğrusal DVM'den tek farkı x yerine *Φ(x)* kullanılması olacaktır. Verilerin doğrusal olarak ayrılamadığı bu durum için kullanılan karar fonksiyonu,

$$f(x) = \left(\sum \alpha_i y_i \Phi(x_i)^T . \Phi(x_j)\right) + b \qquad \textbf{(19)}$$

olarak ifade edilir.

Genelde *Φ(x)* fonksiyonu elde edilebilir değildir, hesaplanamaz hatta mevcut değildir. Uygulanan haritalama fonksiyonu bilinse dahi, kurulan optimizasyon probleminin yüksek boyutlu belirleyici nitelik uzayında çözümü karmaşık ve zor hesaplamalar gerektirecektir. Bu sorunu önlemek amacıyla çekirdek düzenlemesi olarak adlandırdığımız *kernel trick* yöntemi önerilmiştir. Doğrusal olmayan haritalama *Φ*; destek vektörleri $\Phi(x_i)$ ile belirleyici nitelik uzayındaki örüntü vektörü *Φ(x)* arasındaki iç çarpımı hesaplamak için Mercer koşullarına uyan çekirdek fonksiyonlarını ($K(x_i, x)$) kullanır.

ÇEKİRDEK FONKSİYONLARI (KERNEL FUNCTIONS)

DVM'de genellikle, doğrusal olmayan iki sınıf arasındaki en iyi sınırı bulabilmek için, veri, giriş uzayından daha yüksek boyutlu bir belirleyici nitelik uzayına haritalama yoluyla taşınır.[7] Ancak, bu haritalama fonksiyonunu bilmek veya elde etmek genellikle zordur. Bu nedenle, haritalama işlemini yapabilmek için çekirdek fonksiyonlarından faydalanılır. Çekirdek fonksiyonları, bilinen bir haritalama fonksiyonu olmadan yüksek boyutlu uzayda nokta çarpımları hesaplamaya olanak sağlar ve belirleyici nitelik uzayındaki iç çarpımı gerçekleştirdiğinden *Φ* dönüşümünün analitik olarak bilinmesine gerek yoktur. Yani, vektörleri daha yüksek boyutlu bir belirleyici nitelik uzayına açık olarak haritalayıp iç çarpımı orada hesaplamak yerine, değeri doğrudan iki vektörün iç çarpımını veren bir çekirdek fonksiyonu kullanmak daha uygun olacaktır. [11]

Doğrusal olarak ayrılamayan verilerde kullanılan DVM metodunun performansı, sınıflandırma yapılırken seçilmesi gereken çekirdek fonksiyonu ile direk alakalıdır. Bu fonksiyon yardımıyla DVM, sınıflandırılacak doğrusal olmayan verileri daha yüksek boyutlu belirleyici nitelik uzayına taşır.

Çekirdek fonksiyonlarının dayandığı temel fikir, girdi uzayındaki bir takım bileşenlerin belirli bir kurala göre dönüştürülmesidir. Bu dönüşüm, aslında girdi uzayında da gerçekleştirilebilecek bazı işlemlerin, çok boyutlu bir uzayda gerçekleştirilmesi üzerine kurulmuştur ve böylece çekirdek fonksiyonlarını kullanan metotlar girdi uzayında karmaşık olan birçok uygulama üzerinde de çalışabilmektedirler.

Verilerin doğrusal olarak ayrılamadığı durum için kullanılan karar fonksiyonu,

$$f(x) = \left(\sum \alpha_i y_i \Phi(x_i)^T . \Phi(x_j)\right) + b \qquad \textbf{(20)}$$

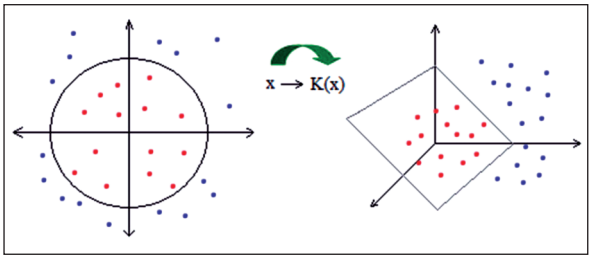
olarak tanımlanmıştı. Burada, $\Phi(x_i)^T \Phi(x_j)$şeklinde gösterilen bir iç çarpım söz konusudur. Çekirdek fonksiyonu, beklenen belirleyici nitelik uzayında iki özellik vektörünün iç çarpımı ile ilgili olan bir fonksiyon olarak ifade edilmektedir.[5] Tüm özellik vektörleri arasındaki iç çarpımlar kullanılarak oluşturulan matris Gram matris (çekirdek matrisi) ola-

rak adlandırılmaktadır. DVM'de, tüm veri girdileri çekirdek fonksiyonundan geçer ve çekirdek matrisinde sonlanır. Yani, çekirdek fonksiyonları, bütün veriyi özetleyen çekirdek matrisini oluştururlar. DVM yönteminde, matematiksel olarak $K(x_i,x_j) = \Phi(x_i)^T \Phi(x_j)$ şeklinde ifade edilen bir çekirdek fonksiyonu yardımıyla doğrusal olmayan dönüşümler yapılabilmekte ve bu şekilde verilerin yüksek boyutta doğrusal olarak ayrımına imkan sağlanmaktadır.[12] Bir çekirdek fonksiyonu olan $K(x_i,x_j)$ ifadesi aslında bir transfer fonksiyonudur. Bu durumda, doğrusal olarak ayrılamayan veriler için kullanılacak karar fonksiyonu aşağıdaki gibidir,

$$f(x) = \sum \alpha_i y_i K(x_i, x_j) \quad (21)$$

Ancak burada doğrusal hiperdüzlem söz konusu olmadığından b terimi ihmal edilir. Çünkü b terimi çekirdek fonksiyonu içinde kapalı biçimde yer almaktadır (Şekil 4).

Doğrusal olarak ayrılamama durumu için DVM'de çeşitli çekirdek fonksiyonları kullanılmaktadır. Farklı çekirdek fonksiyonlarının seçimiyle, farklı DVM'ler üretilir ve bu da farklı sınıflama performanslarına neden olabilir.



ŞEKİL 4: Verilerin çekirdek fonksiyonu ile belirleyici nitelik uzayına taşınması.

TABLO 1: Yaygın olarak kullanılan çekirdek fonksiyonlar.

Çekirdek fonksiyonların isimleri	Çekirdek fonksiyonların matematiksel ifadeleri
Doğrusal fonksiyon	$K(x_i, x_j) = x_i^T x_j$
Polinomiyal fonksiyon	$K(x_i, x_j) = (1 + x_i^T x_j)^d$
Sigmoid fonksiyon	$K(x_i, x_j) = tanh\,(k x_i^T x_j - \delta)$
Radyal tabanlı fonksiyon	$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma^2}\right)$

Literatürde DVM'de kullanılan birçok çekirdek fonksiyon bulunmaktadır (Tablo 1). Ancak bu fonksiyonlardan doğrusal fonksiyon, polinomiyal fonksiyon, sigmoid fonksiyon ve radyal tabanlı fonksiyon (RTF) en çok kullanılanlarıdır ve çoğu tahmin problemi polinomiyal ve radyal tabanlı fonksiyondan biriyle çözülebilmektedir. Polinomiyal çekirdek fonksiyon, RTF'ye göre daha fazla parametre içerir. Bu yüzden, RTF'nin hesaplanması daha kolaydır. Ancak, düşük derecelerde polinomiyal çekirdek fonksiyon ile iyi sonuçlar alınabiliyorsa polinomiyal çekirdek, sınıflandırılamayan kısımlar oluyorsa (C hata maliyeti parametresini içerdiğinden dolayı) RTF kullanılmalıdır. Doğrusal çekirdek fonksiyon ise, özellik sayısının çok fazla olduğu durumlarda daha iyi sonuçlar vermektedir.

Bu çalışmanın uygulama bölümünde, doğrusal olarak ayrılamayan verilerin sınıflandırılması için DVM'de radyal tabanlı çekirdek fonksiyon kullanılmıştır. Ayrıca çekirdek fonksiyon ve hata maliyeti parametrelerinin tahmininde 4-katlı çapraz geçerlilik yöntemi, eğitim ve test setlerinin oluşturulmasında ise 10-katlı çapraz geçerlilik yönteminden yararlanılmıştır.

BULGULAR

Çalışmaya alınan bireylerden elde edilen, sayısal yapıdaki değişkenlerin tanımlayıcı istatistikleri Tablo 2' de topluca verilmiştir. Bu değişkenler bakımından, klinik olarak gece yeme sendromu tanısı alan ve almayan gruplar karşılaştırıldığında, yaş, eğitim yılı, kardeş sayısı, bel çevresi, kalça çevresi ve beden kitle indeksi (BKİ) ortalamaları arasında istatistiksel olarak anlamlı fark bulunmazken (p değerleri sırasıyla 0.702, 0.420, 0.686, 0.734, 0.186, 0.247), GYA puanı, BSQ puanı ve SCL ortalama puanı bakımından, gece yeme sendromu tanısı alan ve almayan gruplar arasında istatistiksel olarak anlamlı fark bulunmuştur (her bir $p<0.001$) (Tablo 2).

Çalışmaya alınan bireylerden elde edilen, kategorik yapıdaki değişkenlerin tanımlayıcı istatistikleri Tablo 3'de verilmiştir. Cinsiyet, çocuk sayısı, kilo vermek, fiziksel hastalık ve antidepresan kullanımı ile gece yeme sendromu tanısı alıp almama durumu arasında istatistiksel olarak anlamlı bir

TABLO 2: Sayısal özelliklere ait tanımlayıcı istatistikler.

	Tanı Aldı (n=336)	Tanı Almadı (n=97)	
	Ort ± Std Sapma	Ort ± Std Sapma	P Değeri
Yaş	36.6 ± 10.6	38.02 ± 12.6	0.702
Eğitim Yılı	8.03 ± 3.3	8.25 ± 3.8	0.420
Kardeş Sayısı	3.9 ± 1.8	4.07 ± 2.07	0.686
Bel Çevresi	91.6 ± 12.9	90.9 ± 15.1	0.734
Kalça Çevresi	107.2 ± 11.1	105.7 ± 10.9	0.186
BKİ	27.6 ± 5.5	27.0 ± 5.1	0.247
GYA Puanı	27.6 ± 7.4	15.6 ± 5.6	<0.001*
BSQ Puanı	93.6 ± 44.0	67.2 ± 33.3	<0.001*
SCL Ortalama Puanı	1.8 ± 0.8	1.2 ± 0.7	<0.001*

BKİ: Beden kitle indeksi; GYA: Gece yeme alışkanlığı; BSQ: Beden şekil indeksi; SCL: Semptom tarama listesi; Ort: Ortalama; Std Sapma: Standart sapma; n: Örneklem genişliği

ilişki bulunmazken (p değerleri sırasıyla 0.369, 0.849, 0.303, 0.169, 0.936), medeni durum, sigara kullanımı, psikolojik tanı ile gece yeme sendromu tanısı alıp almama durumu arasında istatistiksel olarak anlamlı bir ilişki (p değerleri sırasıyla 0. 036, 0. 003, 0. 003) bulunmuştur (Tablo 3).

Tek değişkenli analizlerden elde edilen sonuçlar değerlendirildiğinde, GYA puanı, BSQ puanı ve SCL ortalama puanı, medeni durum, sigara kullanımı, psikolojik tanı değişkenlerinin gece yeme sendromu tanısı koymada etkili oldukları söylenebilir.

Doğrusal olmayan destek vektör makineleri kullanılarak elde edilen sonuçlar incelendiğinde, uygun parametrelerin tahmini için yapılan arama sırasında, değerlendirilen noktalarının sayısı 144, arama ile optimize edilen en iyi duyarlılık ve seçicilik değeri %71 ve modelde kullanılan destek vektörlerin sayısı 207 olarak bulunmuştur. Durdurma kriteri ε (stopping criterion), 0.001 olarak belirlenmiş ve tahmin edilen parametre değerlerleri, hata maliyet değeri, $C = 0.10846446$ ve RTF parametresi olan gamma değeri, $\gamma = 0.02714418$ olarak bulunmuştur (Tablo 4).

TABLO 3: Kategorik özelliklere ait tanımlayıcı istatistikler

		Tanı Aldı		Tanı Almadı		
	Kategori	Sayı	%	Sayı	%	P Değeri
Cinsiyet	Kadın	65	67.0	241	71.7	0.369
	Erkek	32	33.0	95	28.3	
Medeni Durum	Bekar	25	25.8	91	27.1	0.036*
	Evli	59	60.8	226	67.3	
	Dul	13	13.4	19	5.7	
Çocuk Sayısı	0	33	34.0	130	38.8	0.849
	1	18	18.6	60	17.9	
	2	29	29.9	84	25.1	
	3	10	10.3	32	9.6	
	4 ve daha fazla	7	7.2	29	8.7	
Kilo Vermek	Evet	52	55.9	159	49.8	0.303
	Hayır	41	44.1	160	50.2	
Sigara	Kullanmıyor	51	52.6	231	68.8	0.003*
	Kullanıyor	46	47.4	105	31.3	
Fiziksel Hastalık	Yok	55	57.3	217	65.0	0.169
	Var	41	42.7	117	35.0	
Psikolojik Tanı	Majör Depresyon	32	33.0	62	18.5	0.003*
	Anksiyete Bozukluğu	29	29.9	157	46.7	
	Bipolar Afektif Bozukluk	14	14.4	43	12.8	
	Psikotik Bozukluk	7	7.2	40	11.9	
	Diğer	15	15.5	34	10.1	
Antidepresan	Almıyor	22	25.6	81	25.2	0.936
	Alıyor	64	74.4	241	74.8	

TABLO 4: DVM yönteminde 17 belirleyici nitelik yardımıyla tahmin edilen optimum parametre değerleri.

		Tahmin Edilen Parametreler		
Değerlendirilen Nokta Sayısı	**Optimize Edilen En Büyük Duyarlılık ve Seçicilik Değeri**	**ε**	**C**	**γ**
144	0.708927	0.001	0.10846446	0.02714418

ε : Durdurma kriteri; C: Hata maliyeti; γ: RTF parametresi.

TABLO 5: DVM yöntemi ile elde edilmiş eğitim ve test verilerinde modelin sınıflama başarıları.

	Eğitim (Training)		Test (Validation)	
Klinik Tanı	**Var**	**Yok**	**Var**	**Yok**
Var	62	35	54	43
Yok	16	320	23	313
Doğruluk	% 88.22		% 84.76	
Duyarlılık	% 63.92		% 55.67	
Seçicilik	% 95.24		% 93.15	
Pozitif tahmini değer (PTD)	% 79.49		% 70.13	
Negatif tahmini değer (NTD)	% 90.14		% 87.92	
F-Measure	0.7086		0.6207	
ROC eğrisinin altında kalan alan	0.915378		0. 861392	

Belirlenen kesim noktalarına göre, gece yeme sendromu tanısı alan (Var=Tanı alan) ve almayan (Yok=tanı almayan) grupları doğru sınıflayabilmek için modelin sınıflama başarıları Tablo 5'de verilmiştir. Tablo incelendiğinde, doğruluk, seçicilik ve negatif tahmini değerin hem eğitim, hem de test verilerinde yüksek, duyarlılık ve pozitif tahmini değerin ise her iki grupta da nispeten daha düşük olduğu görülmüştür. Ayrıca ROC eğrisinin altında kalan alan eğitim seti için 0.915378, test seti için 0.861392 olarak bulunmuştur. Bulunan bu değerler ve ROC eğrisinin altında kalan alanlar birlikte incelendiğinde, modelin tanı koyma başarısının oldukça iyi derecede olduğu görülmüştür (Tablo 5).

TARTIŞMA VE SONUÇ

DVM metodu, diğer bazı yöntemlere göre daha yeni ve çeşitli problemler için doğru modeller üretebilme yeteneğine sahip bir modelleme yöntemidir ve teorik istatistik temellere dayanarak, özellikle verinin doğrusal olarak ayrılamama durumunda doğru ve güçlü sonuçlar üretir.[13,14]

DVM yöntemi, doğrusal olarak ayrılamayan verileri sınıflandırmak için çeşitli çekirdek fonksiyonlarını kullanmaktadır ve iyi sonuçlar elde etmeyi sağlayacak olan bu fonksiyonların doğru olarak seçilmesi yöntemin en önemli unsurlarındandır. Çekirdek dönüşümü için seçilmesi gereken fonksiyona karar vermede karşılaşılabilecek zorluk, DVM'nin bir dezavantajı olarak düşünülebilir. Ancak çeşitli çalışmalarda, radyal tabanlı çekirdek fonksiyonunun birçok koşulda yüksek sınıflama doğruluğu verdiği bulunmuş ve bu nedenle DVM yöntemiyle sınıflama yapıldığında bu fonksiyonun kullanılması önerilmiştir.[9]

DVM yöntemi, yapay sinir ağları gibi bazı yaklaşımlardan radikal farklılıklar gösterir. Örneğin, DVM eğitimi her zaman global bir minimum bulur ve basit geometrik yorumu daha fazla araştırma için verimli bir zemin sağlar.[5]

Bunların yanı sıra DVM yönteminde bir karar fonksiyonu elde edilebilmekte ve bu fonksiyon yardımıyla yeni bir gözlem ait olduğu sınıfa sınıflandırılabilmektedir. Verilerin dağılımı ve nitelikler arası ilişkilerin şeklinde her hangi bir varsayım gerektirmemesi ise yöntemin birçok alanda kullanımını yaygınlaştıracaktır.

Son yıllarda sınıflama, regresyon ve zaman serileri tahmininde kullanılan DVM yöntemi, bu çalışmanın uygulama bölümünde, tıbbi verilerin

sınıflandırılmasında kullanılmış ve bu amaçla iki sınıflı sınıflama yapılmıştır. RTF çekirdek fonksiyonu kullanılarak uygulanan DVM yönteminden elde edilen sonuçların sınıflama başarısının yüksek olduğu gözlenmiştir.

Yapılan yurt dışı çalışmalar incelendiğinde, veri madenciliği yöntemlerinin tıp alanında yoğun bir şekilde kullanıldığı görülür. Bu yöntemler içerisinde de random forest ve destek vektör makinaları taşıdıkları önemli avantajlar nedeniyle, hastalıkların risk faktörlerini belirlemede klasik yöntemlere tercih edilmektedir. Elde edilen sonuçlar değerlendirildiğinde ise risk faktörlerinin başarılı bir şekilde belirlendiği ve model performanslarının oldukça yüksek olduğu gözlenmiştir.[15,16]

DVM nin tıp alanında en yaygın kullanım alanlarından birisi de değişken sayısı ve gözlem sayısı açısından oldukça kapsamlı veri içeren gen çalışmalarıdır. Son yılların popüler konularından birisi olan gen-hastalık ilişkisinin DVM yöntemi gibi veri madenciliği yöntemleri ile incelenmesi, hem güvenilir hem de kullanılabilir sonuçlar vermektedir.[17,18]

Gece yeme sendromunu etkileyen risk faktörlerinin belirlenmesi amacıyla yapılan çalışmalarda, potansiyel risk faktörleri ya tek değişkenli analiz yöntemleriyle incelenmiş veya bilinen çoklu modeller kullanılmıştır.[19,20] Bu çalışmada ise ilk defa gece yeme sendromunu etkileyen risk faktörleri DVM yöntemiyle incelenmiştir.

Konu ile ilgili dikkate alınmayan veya ölçülmeyen veya ölçülemeyen risk faktörleri söz konusu olabilir. Ayrıca dikkate alınan risk faktörlerinin, gece yeme sendromu ile doğrudan ilişkileri incelenmiş, risk faktörleri arasındaki interaksiyonlar değerlendirilmemiştir. Bu iki durum çalışmanın bir sınırlılığı olarak düşünülebilir.

KAYNAKLAR

1. Koyuncugil AS, Özgülbaş N. [Data mining: using and applications in medicine and healthcare]. Bilişim Teknolojileri Dergisi 2009;2(2): 21-31.
2. Moguerza JM, Muñoz A. Support vector machines with applications. Stat Sci 2006; 21(3):322-36.
3. Noble WS. What is a Support vector machine? Nat Biotechnol 2006;24(12):1564-67.
4. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machine. Mach Learn 2002; 46(1-3):389-422.
5. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 1998;2(2):121-67.
6. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-7.
7. Cortes C, Vapnik V. Support-vektor networks. Mach Learn 1995;20(3):273-97.
8. Herbrich R. Learning Kernel Classifiers: Theory and Algorithms. 1st ed. Massachusetts: The MIT Press; 2002. p.1-384.
9. Kavzoğlu T, Çölkesen İ. [Investigation of the effects of Kernel functions in satellite image classification using support vector machines]. Harita Dergisi 2010;144(7):73-82.
10. Joachims T. Learning to Classify Text Using Support Vektor Machines: Methods, Theory and Algorithms. 1st ed. London: Kluwer Academic Publishers / Springer; 2002. p.1-228.
11. Huang CL, Wang CJ. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications 2006;31(2):231-40.
12. PingWu K, De Wang S. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognition 2009;42(5):710-7.
13. Cortes C, Vapnik V. Support vector networks. Machine Learning 1995;20(3):273-97.
14. Pochet NL, Suykens JA. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. Ultrasound Obstet Gynecol 2006;27(6):607-8.
15. Zhou XH, Li SL, Tian F, Cai BJ, Xie YM, Pei Y, et al. Building a disease risk model of osteoporosis based on traditional Chinese medicine symptoms and western medicine risk factors. Stat Med 2012;31(7):643-52.
16. Fitzgerald AJ, Pinder S, Purushotham AD, O'Kelly P, Ashworth PC, Wallace VP. Classification of terahertz-pulsed imaging data from excised breast tissue. J Biomed Opt 2012; 17(1):016005. doi: 10.1117/1.JBO.17.1. 016005.
17. Duttagupta R, DiRienzo S, Jiang R, Bowers J, Gollub J, Kao J, et al. Genome-wide maps of circulating miRNA biomarkers for ulcerative colitis. PLoS One 2012;7(2):e31241.
18. Yi Z, Li Z, Yu S, Yuan C, Hong W, Wang Z, et al. Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder. PLoS One 2012;7(2):e31283.
19. Gallant AR, Lundgren J, Drapeau V. The night-eating syndrome and obesity. Obes Rev 2012;13(6):528-36.
20. Jacobi C, Fittig E, Bryson SW, Wilfley D, Kraemer HC, Taylor CB. Who is really at risk? Identifying risk factors for subthreshold and full syndrome eating disorders in a high-risk sample. Psychol Med 2011;41(9):1939-49.