Available online at www.alphanumericjournal.com

alphanumeric journal
The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 5, Issue 2, 2017

Received: October 19, 2017
Accepted: November 13, 2017
Published Online: November 29, 2017

AJ ID: 2017.05.02.STAT.03
DOI: 10.17093/alphanumeric.345115

Classification of Gene Samples Using Pair-Wise Support Vector Machines

Engin Taş, Ph.D. *

Assist. Prof, Department of Statistics, Faculty of Science and Literature, Afyon Kocatepe University, Afyonkarahisar, Turkey, engintas@aku.edu.tr

* Afyon Kocatepe Üniversitesi Fen Edebiyat Fakültesi Ahmet Necdet Sezer Kampusü, 2. Eğitim Binası, 03200 Afyonkarahisar / Türkiye

ABSTRACT

The main problem in the classification problems encountered with gene samples is that the dimension of the data is high although the sample size is small. In such problems, the classifier to be used must be a classifier that allows the processing of high dimensional data and extracts maximum information from a small number of samples at hand. In this context, a classification methodology has been developed, which first transforms the problem of binary or multiple classification into separate pair-wise classification problems. To this end, an online classifier has been adapted to solve pair-wise binary classification problems. The resulting classifier performed better on most of the real problems compared to other popular classifiers.

Keywords: Tumor Classification, Pair-Wise Classification, Support Vector Machine, Kernel Methods

Gen Örneklerinin Eşli Destek Vektör Makinesi ile Sınıflandırılması

ÖZ

Gen örnekleriyle ilgili karşılaşılan sınıflandırma problemlerinde en büyük sorun az sayıda örnek elde edilmesine karşın verinin büyük boyutlu olmasıdır. Bu tür problemlerde kullanılacak sınıflandırıcının büyük boyutlu verinin işlenmesine olanak sağlayan ve eldeki az sayıda örnekten maksimum bilgiyi çıkaran bir sınıflandırıcı olması gerekir. Bu kapsamda, öncelikle ikili/çoklu sınıflandırma problemlerini ayrı ayrı eşli ikili sınıflandırma problemlerine çeviren bir sınıflandırma metodolojisi geliştirilmiştir. Bunun için, çevrimiçi bir sınıflandırıcı eşli ikili sınıflandırma problemlerini çözecek şekilde tekrar düzenlenmiştir. Oluşan sınıflandırıcı gerçek problemlerin çoğu üzerinde diğer popüler sınıflandırıcılara göre oldukça iyi bir performans göstermiştir.

Anahtar Kelimeler: Tümör Sınıflandırması, Eşli Sınıflandırma, Destek Vektör Makinesi, Çekirdek Yöntemler

1. Giriş

Kanser vakalarında kanserli tümörlerin hassas bir şekilde tespit edilmesi, zor olmasına karşın başarılı bir tedavi süreci için çok değerlidir. Bilgisayar teknolojisindeki son gelişmeler doğrultusunda elde edilen büyük boyutlu, yüksek verimli gen örneği çıkarma yöntemleri ve buna uygun tutarlı istatistiksel metotlar kullanılarak, biyomoleküler bilgi kanser tedavilerinde çok önemli bir konuma ulaşmıştır. Bu çalışmalardaki temel problem, mikrodizi deneyleri sonucu elde edilen veri kümelerinin binlerce gene ait örnekten oluştuğu için büyük boyutlu olması ve bunun yanında birkaç düzine mikrodizi örneğinden oluşmasıdır. Başka bir ifadeyle, gözlem sayısından çok daha fazla sayıda tahmin edici bağımsız değişken bulunmasıdır. Bu durum yeni istatistiksel problemlerin doğmasına ve bu problemlerin çözümüne yönelik yeni tekniklerin geliştirilmesine yol açmıştır.

Mikrodizi verisine dayanarak kanserli hastaların tespit edilmesi problemi bir istatistiksel sınıflandırma problemidir. Bu problemin çözümü için literatürde diskriminant analizi, cezalandırılmış regresyon teknikleri ve en yakın komşu kuralı gibi klasik parametrik olmayan yöntemlerden, yapay sinir ağları ve destek vektör makineleri gibi modern yapay öğrenme tekniklerine kadar birçok yöntem bu problemlerin çözümünde kullanılmıştır. Bu çalışmaları gözden geçirmek için (Dudoit, Fridlyand, & Speed, 2003)'e bakılabilir. Bu çalışmada ise var olan yöntemlere alternatif olarak istatistiksel sınıflandırma problemine farklı bir açıdan yaklaşan bir sınıflandırma metodolojisi geliştirilmiştir.

Şimdi ikili bir sınıflandırma problemini ele elalım. Elimizde belirli bir hastalığa ilişkin sağlıklı ve hasta insanlara ait gözlemlerden oluşan bir veri kümesi olsun. Aynı sınıftan gelen iki örnek arasında pozitif bir bağ, farklı sınıftan gelen iki örnek arasında negatif bir bağ olduğu düşünülebilir. Bu şekilde ele aldığımız bu problem, insanların varlıkları ve aynı sınıfa (sağlıklı/hasta) ait olup olmama bilgisinin de ilişkileri oluşturduğu bir ağ şeklinde temsil edilebilir. Bu sayede ikili sınıflandırma problemi ile bir ağdaki bağ tahmini arasında bir benzerlik kurulmuş olur. Diğer bir ifadeyle, ikili sınıflandırma probleminde yeni bir örneğin ilgili sınıfa atanması problemi, bir ağa yeni eklenen bir düğümün hangi düğümlerle bağlantı oluşturacağını tahmin etme problemi gibi düşünülebilir. Bu yaklaşımı pratikte gerçekleştirebilmek için mevcut verinin farklı bir şekilde temsil edilmesi gerekir. Örneğin, klasik sınıflandırma probleminde bir gözlem sadece bir kişiye ait özelliklerden oluşan bağımsız değişkenler ve bu kişinin ait olduğu sınıfı belirleyen ikili bir etiketten oluşurken, yeni yaklaşımda bir gözlem iki farklı kişiye ait özelliklerden oluşan bağımsız değişkenlerden ve bu iki kişinin aynı sınıfa ait olup olmadığını belirleyen bir etiketten oluşur.

Destek Vektör Makinesi'nin (DVM) temel fikri, doğrusal olarak ayrılabilen veriler üzerinden optimal bir ayırma hiperdüzlemi oluşturmaktır (Boser, Guyon, & Vapnik, 1992). Aynı zamanda çekirdekleri ve yumuşak marj formülasyonlarını kullanarak doğrusal olarak ayrılamayan verilerde geniş marjlı bir hiperdüzlemi de öğrenebilir. Bununla birlikte, DVM başlangıçta ikili sınıflandırma için tasarlanmıştır ve DVM'yi çok sınıflı senaryoya genişletmek için iki ana yaklaşım vardır. Bir yaklaşım ikili algoritmayı çok sınıfa genelleştirmektir (Weston & Watkins, 1999, Mayoraz & Alpaydin, 1999), başka bir yaklaşım ise çok sınıflı sınıflandırma problemini bir dizi ikili problem haline dönüştürmektir. En eski ve en yaygın kullanılan uygulamalardan biri, her biri her sınıfı

diğerlerinden ayıran m ikili DVM sınıflandırıcılarını oluşturan, tümüne karşı tek yaklaşımıdır (Dietterich & Bakiri, 1995). İ. DVM, i. sınıfın tüm örneklerini pozitif etiketlerle ve diğer tüm örnekleri negatif etiketlerle ele alarak eğitilir. Eşli sınıflandırma ise, iki sınıflı problemlerin her birinden elde edilen eşli karşılaştırmalar göz önüne alınarak, çok sınıflı problemleri çözmek için alternatif bir tekniktir (Friedman, 1996). Test kümesinden bir örnek sınıflandırılırken, en çok eşli karşılaştırmayı kazanan sınıfa atanır. Bu çalışmada, çiftleri sınıflandırmada kullanılmak üzere her bir sınıf için eşli bir DVM modeli oluşturduk. Burada aynı sınıftan gelen iki yumurta pozitif bir çifti, farklı sınıftan gelen iki yumurta negatif bir çifti oluşturur. Bu eşli DVM modeli, herhangi bir çiftin pozitif bir çift olup olmadığını belirleyebilir. Diğer taraftan, eşli bir düzenlemede n örnek n^2 eşli örneğe karşılık gelir ve büyük ölçekli bir veri kümesi ile birlikte bir destek vektör makinesinin eğitilmesi çoğu durumda ciddi hesaplama maliyetleri getirir. Veriler toplu olarak işlendiğinde, DVM'ler her adımda amaç fonksiyonun hesaplanmasını gerektirir, ve bu temel olarak önceden tanımlanmış bir kayıp fonksiyonunun eğitilecek bir veri seti üzerinden hesaplanmasını gerektirir. Gradyana dayalı yöntemler, amaç fonksiyonunun her bir değerlendirmesinde sırasıyla gradyanı hesaplarken, Newton yöntemi ve eşlenik gradyan algoritması gibi standart sayısal optimizasyon teknikleri, amaç fonksiyonunun ikinci dereceden bilgisine ihtiyaç duyar. Mevcut veri setleri gittikçe büyüdükçe, bu tür klasik ikinci dereceden yöntemler neredeyse tüm durumlarda uygulaması pratik değildir.

Buna karşın, algılayıcı (perceptron) (Rosenblatt, 1958, Minsky & Papert, 1969) ve varyantları (Freund & Schapire, 1999, Li & Long, 2002, Gentile, 2002, Anlauf & Biehl, 2007) gibi çevrimiçi gradyan tabanlı yöntemler , büyük ve tekrarlı veri kümelerinde büyük bir avantaja sahiptir. Aslında, basit çevrimiçi gradyan düşümü yöntemleri (Bottou & LeCun, 2004, Shalev-Shwartz, Singer, & Srebro, 2007, Xu, 2011) genelde sofistike ikinci derece toplu algoritmalardan daha iyi performans sergiler, çünkü çevrimiçi yöntemlerin hesaplama gereksinimleri, eğitim verilerini tek tek örnekler veya küçük alt örnekler şeklinde işledikleri için oldukça düşüktür. Dolayısıyla bu çalışma, problemi farklı tanımlayarak ve buna uygun bir yöntemin seçimi ile ilgili iki temel fikre dayanır. İlk olarak, herhangi bir ikili veya çok sınıflı sınıflandırma probleminin eşli sınıflandırma problemine dönüştürülebileceğini biliyoruz. Bu, daha zengin bir veri kümesinden öğrenmenin avantajını getirir ve eşli model, örneklerden öğrendiğimizden daha fazlasını öğrenebilir. İkincil olarak, geleneksel öğrenme algoritmaları bu kapsamda daha verimsiz hale geldiğinden ve bazı durumlarda uygulanamadığı için (ör. toplu yöntemler), çevrimiçi DVM algoritmasını örnek çiftlerle çalışacak şekilde değiştirilmesini öneriyoruz. Bu yaklaşım, çiftleri tek tek işleme avantajını getirir ve daha büyük boyutlara sahip büyük ölçekli verilerden kaynaklanan zorlukların üstesinden gelir.

2. Gereç ve Yöntem

$\mathcal{X} = (x_1, x_2, \ldots, x_m), \forall x_i \in \mathbb{R}^n$ şeklinde bir örnek kümemiz olsun. İki örneğin herhangi bir kombinasyonu $p = (x_i, x_j) \in \mathcal{P} \subseteq \mathcal{X}^2$ çifti olarak düşünülebilir. Bu çiftlerden oluşan $T = \{(p, y_p): p \in \mathcal{P}\}$ dizisini $\mathbb{Z} = \mathcal{P} \times \{+1, -1\}$ olasılık dağılımına sahip bir kitleden çekilmiş bir eğitim örneklemi olduğunu düşünelim. T'den gelen bir örnek, n-boyutlu bir sütün vektörü çifti ve bu iki örneğin aynı sınıftan gelip gelmediğini belirleyen bir y_p $(+1, -1)$ etiketinden oluşan bir üçlüdür. Amaç eğitim örnekleminden uygun bir $f: \mathcal{P} \rightarrow$

$\{+1,-1\}$ fonksiyonunu öğrenmektir. Karar fonksiyonunun, $f(p)=\langle w,\Phi(p)\rangle$ şeklinde temsil edildiği doğrusal durumu ele alalım, burada $w\in\mathbb{R}^n$ eğitim örneklemi T'ye dayanarak tahmin edilmesi gereken parametre vektörüdür ve Φ ise çiftleri daha büyük bir uzaya gönderen bir özellik fonksiyonudur.

Optimal bir karar fonksiyonu (en büyük marjine sahip optimal hiperdüzlem) özellik uzayında aşağıdaki amaç fonksiyonunu minimize ederek bulunur (Schölkopf & Smola, 2001):

$$\min_{w}\|w\|^2+C\sum_{i=1}^{n}\xi_i \; with \begin{cases}\forall_i & y_i\hat{y}(p_i)\geq 1-\xi_i\\ \forall_i & \xi_i\geq 0\end{cases} \tag{1}$$

$\xi=\xi_1,\xi_2,\ldots,\xi_n$ gevşek değişkenleri bazı çiftlerin marjinin yanlış tarafında yer almasına izin verir. Düzeltme parametresi C'nin büyük değerleri için, ayrılamayan çiftlere büyük bir ceza verilir ve daha fazla sayıda destek çifti oluşur. C'nin küçük değerleri bu cezanın etkisini yumuşatır ve gürültülü problemlerde daha iy sonuçlar elde etmemizi sağlar ama yetersiz uyuma sahip bir destek çifti modeli oluşturabilir.

Bu konveks optimizasyon probleminin eşini (dual) maksimize etmek esas problemden daha basit bir konveks kuadratik programlama problemidir. DVM çekirdek genişlemesinin α_i katsayıları aşağıdaki eş amaç fonksiyonunu

$$W(\alpha)=\sum_i\alpha_i y_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j K(p_i,p_j) \tag{2}$$

tanımlayarak ve

$$\max_{\alpha}(\alpha)\; with \begin{cases}\sum_i\alpha_i=0\\ A_i\leq\alpha_i\leq B_i\\ A_i=\min(0,Cy_i)\\ B_i=\max(0,Cy_i)\end{cases} \tag{3}$$

DVM kuadratik programlama problemini çözerek bulunabilir. Orta büyüklükteki veri kümelerinde bile eşli öğrenme gerçekleştirdiğimizde çok büyük sayıda örnek çiftlerini işlememiz gerekir, bu nedenle etkin ve hızlı bir SVM sınıflandırıcısına ihtiyaç duyarız. Bu zorlukların üstesinden gelebilmek için çevrimiçi ve aktif öğrenme özellikleriyle çekirdeklerle çalışan hızlı bir sınıflandırıcı olan LASVM (http://leon.bottou.org/projects/lasvm) algoritması (Bordes, Ertekin, Weston, & Bottou, 2005) kullanılmıştır. Çevrimiçi yapısından dolayı eşli öğrenme kapsamında bu algoritma çeşitli avantajlara sahiptir. LASVM örnekleri tek tek işleyerek ve en çok bilgilendirici olan destek vektörlerini açılımında tutarak hesaplama karmaşıklığı ve maliyetiyle başa çıkabilir. Bu gerekli olan hesaplama miktarını ciddi anlamda düşürür. LASVM algoritmasının başka bir avantajı da seyrek veri kümelerinde kullanılmasına uygun olmasıdır zira veri boyutunun büyük olduğu çoğu veri kümesinde örnekler önemli derecede seyrek bir yapıya sahiptir yani örnek vektörünün bazı elemanları değer almaz. LASVM bu problemi bu örneklere uygun hızlı seyrek vektör çarpımları kullanarak bir avantaja dönüştürür. Bu sayede örneklerin ikili kombinasyonlarından oluşan çiftler tek tek işlenerek daha zengin bir veri kümesinden öğrenmenin avantajları değerlendirilmiş olur. LASVM ayrıca herhangi bir zamanda çekirdek

açılımında toplanan vektörlerin çevrimiçi süreçte açılımdan çıkarıldığı bir destek vektörü çıkarma adımına da sahiptir. Bu da o anda çekirdek açılımındaki etkinliğini yitirmiş olan destek vektörlerinin temizlenmesi anlamına gelir. Bu sayede çekirdek açılımı en etkin ve kompakt şekilde süreç boyunca korunmuş olur. LASVM algoritması aynı zamanda sıralı minimal optimizasyon (SMO) (Platt, 1999) algoritmasıyla ilişkilidir ve SVM kuadratik programlama probleminin çözümüne yakınsar.

Diğer yandan, LASVM eşli öğrenme için uygun değildir. Çünkü, bu haliyle, çekirdek belleği eşler için hesaplanan çekirdek değerlerini tutar ve bu da hem bellek kapasitesi hem de hesaplama yükü olarak oldukça ciddi maliyetler oluşturur. Bu durum, algoritmayı büyük veri kümeleri için elverişsiz hale getirir. Ayrıca eşler için çekirdek değerlerini saklamak anlamsızdır, zira örnekler için çekirdek değerleri bir kez hesaplandığında herhangi bir çift için çekirdek değerleri hesaplamak daha sonra göreceğimiz üzere 3 basit aritmetik işlemden oluşur. Bu nedenle, bu çalışmada LASVM algoritmasını eşli öğrenme durumunda çalıştırabilmek için algoritmanın yapısında ciddi değişiklikler yapılmıştır. Oluşan algoritma eşli-LASVM olarak adlandırılmıştır. Bu çalışmanın birinci ana katkısı, aşağıda maddeler halinde özetlenen, eşli-LASVM algoritmasını oluşturabilmek için LASVM algoritmasında gerçekleştirilen temel değişikliklerdir.

- Destek çiftlerinin indislerini ve karşılık gelen örneklerin indislerini tutmak için sırasıyla P ve S kümeleri tanımlanmıştır. Eşli-LASVM çekirdek açılımına bir çift eklediğinde (işleme), çiftin indisi P kümesine eklenir ve aynı zamanda bu çifti oluşturan iki örneğin indisleri de S kümesine eklenir. P kümesi sadece çiftlerin indislerini tutmak için kullanılır, herhangi bir çekirdek önbelleği kullanılmaz. Çekirdek değerleri sadece ilgili çifti oluşturan örnekler için hesaplanır ve çekirdek önbelleğinde tutulur. Dolayısıyla, S kümesiyle beraber örneklerin çekirdek değerlerini tutan bir çekirdek önbelleği kullanılır.

- LASVM algoritmasında bir örneği işlemek için gerekli olan tüm yordamlar, bir çifti işleyecek şekilde yeniden düzenlenmiştir. Bu bir çift için ilgili gradyanın hesaplanması, maksimum gradyana sahip τ-bozan dörtlünün (iki çift tarafında oluşturulan) belirlenmesi ve uygun adım yönlerinin belirlenmesini içerir.

- Yeniden işleme bazı çiftleri P'den çıkartır. Buna karşılık S kümesinden bu çiftle ilgili olarak iki örnek çıkartılır. Sonuç olarak sapma terimi b ve P kümesindeki en çok τ-bozan dörtlünün gradyanı δ değerlerinin tümü hesaplanır.

LASVM öncelikle çekirdek açılımına en az bir çift ekleyerek sürece başlar ve daha sonra o anki çekirdek açılımındaki var olan gereksiz destek çiftlerini arar. Çevrimiçi durumda bu, t anında yeni bir çifti işlemek için kullanılabilir. Çekirdek açılımında herhangi bir t anında $\alpha_i \neq 0$ katsayısına sahip çiftler destek çiftleri olarak tanımlanır. Destek çiftlerine karşılık gelen çiftlerin indisleri P kümesinde tutulurken, bu çiftlere karşılık gelen örneklerin indisleri S kümesinde tutulur. Eğer $i \notin S$ ise karşılık gelen α_i'lerin değer almadığı varsayılır.

Eşlerden öğrenme durumunda, eşli-LASVM'ın yapısı daha fazla önem arz eder, çünkü (işleme) ve (yeniden işleme) adımları daha fazla bilgiye sahip eşleri elinde tutarken, gerek duyulmayan önemini yitirmiş çiftleri de çekirdek açılımında çıkartarak bunlardan kurtulur. Bu çiftlerin sayısının kuadratik olarak büyüdüğü eşli öğrenme durumunda en kullanışlı durumdur. Diğer taraftan, çiftlerden bir model öğrenmek için, ortak bir özellik alanında bir çifti temsil edecek ek bir yapı türüne ihtiyaç duyarız. Basilico &

Hofmann (2004) kullanıcı derecelendirmelerini ve öğe özelliklerini ortak bir öğrenme mimarisinde birleştirmiş, farklı örnek çiftleri için uygun çekirdeklerin tasarımında iyi örnekler vermiştir ve çekirdeklerin tek bir çekirdeğe birleştirilmesinin birkaç yolunu göstermiştir. Birbirinden farklı özellik haritalarını basitçe birleştirmek için tensör çarpımlarını kullanmışlardır. Oyama & Manning (2004) eşli sınıflandırıcıları öğrenmede örnek çiftleri arasındaki özelliklerin kombinasyonlarını kullanmak için bir çekirdek önermiştir. Bu kapsamda, Ben-Hur & Noble (2005) proteinler arasındaki bir çekirdeği protein çiftleri arasındaki bir çekirdeğe çeviren tensör çarpım eşli çekirdeğini (TÇEÇ) önermiştir. Vert, Qiu, & Noble (2007) biyolojik ağların yeniden inşası için metrik öğrenme eşli çekirdeğini (MÖEÇ) geliştirmiştir. Kashima, Oyama, Yamanishi, & Tsuda, 2009) çevrimiçi öğrenme sürecini hızlandırmak için Kartezyen çekirdeğini kullanarak bu genel çerçevenin özel bir durumunu önermiştir. Kartezyen çekirdek TÇEÇ ve MÖEÇ' den daha seyrek bir yapıya sahiptir. Ayrıca çekirdek matrislerinin özdeğer analizine dayanarak iki farklı eşli çekirdek için genelleştirme sınırları verilmiştir.

Bu çalışmada, eşleri temsil etmek için TÇEÇ kullanılmıştır. $p_1 = (x_1, x_2)$ ve $p_2 = (x_3, x_4)$ şeklinde iki çift için TÇEÇ

$$K_{TPPK}\left((x_1,x_2),(x_3,x_4)\right) = K(x_1,x_3)K(x_2,x_4) + K(x_1,x_4)K(x_2,x_3) \qquad (4)$$

şeklinde verilir. Bunun yanında örnekler arası RBF çekirdeği gibi bir çekirdek kullanarak, TÇEÇ çekirdeği çiftler arasındaki herhangi bir ilişkiyi öğrenen, evrensel olarak yaklaşan bir fonksiyonlar sınıfı H üretir. Bunun yanında, tüm denemelerimizde MÖEÇ'nin de performansını değerlendirmemize rağmen TÇEÇ'den anlamlı bir farklılık görülmemiştir. Bu nedenle, esas amaçtan uzaklaşmamak için uygulamada sadece TÇEÇ kullanılmıştır.

3. Sonuçlar

Çalışmanın uygulaması Tablo 1'de verilen altı farklı problem üzerinde gerçekleştirilmiştir. Bu problemlerde temel özellik daha önce belirttiğimiz gibi değişken sayısının çok fazla olması ama bunun yanında gözlem sayısının çok az olmasıdır. Verilere uygulanan ön işlemlerden sonra, tüm gen örnekleri logaritma 10 tabanına dönüştürülmüş ve sıfır ortalamaya ve birim varyansa sahip olacak şekilde standartlaştırılmıştır.

Veri kümesi	Yayın	*n*	*p*	*M*	Bulgu
Lösemi	Golub vd. (1999)	72	3571	2	Lösemi'nin alt türleri
Kolon	Alon vd. (1999)	62	2000	2	Tümor/normal doku
Prostat	Singh vd.(2002)	102	6033	2	Tümor/normal doku
Lenfoma	Alizadeh vd. (2000)	62	4026	3	Lenfoma'nın alt türleri
SRBCT	Khan vd. (2001)	63	2308	4	Farklı tümör türleri
Beyin A	Pomeroy vd. (2002)	42	5597	5	Farklı tümör türleri

Tablo 1. Çalışmada kullanılan veri kümeleri ve özellikleri

Gözlemlerin üçte ikisinden dengeli bir eğitim kümesi oluşturulmuştur. Eğitim kümesi kullanılarak sınıflandırıcının ceza parametrelerinin ve çekirdek türünün belirlenmiştir. Tablo 2'de sınıflandırıcıların eğitimi aşamasında belirlenen en iyi ceza parametresi değerleri ve çekirdek türleri verilmiştir. Daha sonra nihai sınıflandırıcının eğitimi gerçekleştirilmiştir. Geriye kalan gözlemler test kümesi olarak kullanılmış ve bu gözlemlerin hangi sınıfa ait olduğu tahmin edilmiştir. Gerçek sınıf ile tahmini sınıf

üzerinde önerilen sınıflandırıcının yanlış sınıflandırma hatası hesaplanmıştır. Her bir problem için toplamda 50 deneme gerçekleştirilmiş ve hata tahminlerinin ortalaması Tablo 3'de verilmiştir.

Veri kümesi	Sınıf	C	Kernel
Lösemi	0	100	Doğrusal
	1	100	
Kolon	0	100	RBF
	1	100	
Prostat	0	10	Doğrusal
	1	10	
Lenfoma	0	10	Doğrusal
	1	10	
	2	10	
SRBCT	0	10	Doğrusal
	1	10	
	2	10	
	3	10	
Beyin	Tek model	1000	Doğrusal

Tablo 2. En iyi ceza parametresi değerleri ve çekirdek türleri

Yanlış sınıflandırma oranları karşılaştırıldığında, önerilen eşli-LASVM sınıflandırıcısının 6 veri kümesinin 4'ünde diğer sınıflandırıcılara göre daha iyi performans göstermiştir. Lenfoma ve SRBCT veri kümelerinde ise daha kötü bir hata oranına sahiptir. Bu iki veri kümesinin ortak özelliği sınıflandırılacak sınıf sayısının diğer veri kümelerine göre daha fazla olmasıdır. Bu nedenle pw-LASVM tarafından oluşturulan ikili sınıflandırma problemi sayısı artar, m sınıflı bir sınıflandırma problemi için pw-LASVM m(m*1)/2 adet ikili SVM modeli oluşturur. Ayrıca, sınıflardaki örneklerin dengesiz dağılımı yine önerilen algoritmanın performansını düşürmüş olabilir. Bu gibi problemlerde, bu çalışmada önerilen yaklaşım çok da avantajlı olmayabilir. Diğer dört veri kümesinde ise önerilen yaklaşımın diğer sınıflandırıcılara göre daha üstün performans sergilediği açıktır.

	Lösemi	Kolon	Prostat	Lenfoma	SRBCT	Beyin
	(%)	(%)	(%)	(%)	(%)	(%)
Eşli-LASVM	**1.25**	**7.24**	**4.98**	3.81	3.81	**23.57**
BagBoost	4.08	16.10	7.53	1.62	1.24	23.86
Boosting	5.67	19.14	8.71	6.29	6.19	27.57
RanFor	1.92	14.86	9.00	1.24	3.71	33.71
SVM	1.83	15.05	7.88	1.62	2.00	28.29
PAM	3.75	11.90	16.53	5.33	2.10	25.29
DLDA	2.92	12.86	14.18	2.19	2.19	28.57
***k*NN**	3.83	16.38	10.59	1.52	1.43	29.71

Tablo 3. Eşli-LASVM sınıflandırıcısı ile 7 farklı sınıflandırıcının 6 farklı mikrodizi veri kümesi üzerindeki hata oranlarının ortalamaları

4. Tartışma

Mikrodizi verisine dayanarak sınıflandırma problemlerinde temel amaç kanserli tümörlerin erken bir aşamada büyük bir doğrulukla tespit edilmesi ve bunun sonucunda ilgili tespite yönelik daha başarılı tedavilerin uygulanabilmesidir. Bu kapsamda, bu tür sınıflandırma problemlerinde karşılaşılan temel problem örnek sayısının az ama verini boyutunu büyük olmasıdır. Dolayısıyla büyük boyutlu gen örnekleriyle çalışabilen ve elde edilen az sayıdaki örneğin içerdiği bilgiden olabildiğince faydalanan sınıflandırma algoritmalarının geliştirilmesi gerekir. Bunun için, bu tür problemlerde karşılaşılan ikili/çoklu sınıflandırma problemlerini etkin bir şekilde yeniden düzenleyerek ikili eşli problemlere dönüştürüp daha başarılı bir şekilde sınıflandıran bir sınıflandırma metodolojisi geliştirilmiştir. Bu yaklaşım eldeki veriyi eşli hale dönüştürerek verinin genişletilmesine olanak sağlamıştır. Bu kapsamda, LASVM algoritması eşli veri kümleriyle çalışacak şekilde yeniden geliştirilmiş ve oluşan algoritma eşli-LASVM olarak adlandırılmıştır. Eşli-LASVM örnekleri çevrimiçi işleyerek oldukça hızlı bir şekilde eşli-SVM modelini kurabilir. Önerilen yaklaşımın gen örnekleriyle ilgili ikili/çoklu sınıflandırma problemlerinde oldukça başarılı bir performans gösterdiği gerçek veri kümeleri kullanılarak gösterilmiştir.

Kaynakça

Anlauf, J.K., & Biehl, M. (1989). The adatron: an adaptive perceptron algorithm. *EPL (Europhysics Letters)*. *10*(7): p. 687.

Basilico, J., & Hofmann, T. (2004). Unifying collaborative and content-based filtering. In *Proceedings of the twenty-first international conference on Machine learning*. p. 9. ACM.

Ben-Hur, A. & Noble, W.S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1): pp.i38-i46.

Bordes, A., Ertekin, S., Weston, J., & Bottou, L. (2005). Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*. *6*: pp.1579-1619.

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *In the Proceedings of the fifth annual workshop on Computational learning theory*. pp: 144-152. ACM.

Bottou, L., & LeCun, Y. (2003). Large scale online learning. In *NIPS*. 30: p. 77.

Dietterich, T., & Bakiri G. (1995). Solving multiclass learning problems via error correcting output codes, *Journal of Artificial Intelligence Research*, 2: 263-286.

Dudoit, S., Fridlyand, J., & Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using geneexpression data. *J. Am. Stat. Assoc.* 97: 77–87.

Freund, Y., & Schapire, R.E. (1999). Large margin classification using the perceptron algorithm. *Machine learning*. *37*(3): pp.277-296.

Friedman J. (1996). Another approach to polychotomous classifcation, Technical Report, Technical report, Stanford University, Department of Statistics.

Gentile, C. (2001). A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*. *2*: pp.213-242.

Kashima, H., Oyama, S., Yamanishi, Y., & Tsuda, K. (2009). On pairwise kernels: An efficient alternative and generalization analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. pp. 1030-1037. Springer Berlin Heidelberg.

Li, Y., & Long, P.M. (2002). The relaxed online maximum margin algorithm. *Machine Learning*. *46*(1-3): pp.361-387.

Mayoraz, E., & Alpaydin E. (1999). Support vector machines for multi-class classification. *In the International Work-Conference on Artificial Neural Networks*. Springer Berlin Heidelberg.

Minsky, M., & Papert, S. (1969). Perceptrons.

Oyama, S., & Manning, C.D. (2004). Using feature conjunctions across examples for learning pairwise classifiers. In *European Conference on Machine Learning.* pp. 322-333. Springer Berlin Heidelberg.

Platt, J.C. (1999). 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*. pp.185-208.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6): p. 386.

Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*. pp. 807-814. ACM.

Schölkopf, B., & Smola, A.J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Vert, J.P., Qiu, J., & Noble, W.S. (2007). A new pairwise kernel for biological network inference with support vector machines. *BMC bioinformatics*, *8*(10): p.S8.

Weston, J., & Watkins, C. (1999). Support vector machines for multi-class pattern recognition. *In ESANN*. 99: pp. 219-224.

Xu, W. (2011). Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*.