Çanakkale Onsekiz Mart Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2018:4,1, 14-27

Çanakkale Onsekiz Mart University, Journal of Graduate School of Natural and Applied Sciences, 2018:4,1, 14-27

Araştırma / Research

Karmaşık Hastalıkların Teşhisinde Veri Madenciliği Yöntemlerinin Başarım Karşılaştırması

Sait Can Yücebaş

Çanakkale Onsekiz Mart Üniversitesi. Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü

15.02.2018/Geliş/Received, 18.04.2018 Kabul/Accepted

Özet

Bütünsel genom ilişkilendirme çalışmalarında (BGİÇ) ortaya çıkan verilerin yüksek miktarda ve çok boyutlu olması, profillerin hastalıklarla ilişkilendirilmesi ve buradan teşhise gidilmesi sırasında farklı veri madenciliği yöntemlerinin kullanılması ile mümkün olmaktadır. Yapılan çalışmada 1025 vaka ve 531 kontrolden oluşan melonom veri kümesi ile farklı etnik kökenli 2325 vaka ve 2350 kontrolden oluşan ve prostat kanseri veri kümesi kullanılmıştır. Bu hastalıklarla ilgili profiller Karar Ağacı, Naive Bayes, Destek Vektör Makinası gibi farklı veri madenciliği yöntemleri ile incelenmiştir. Her iki hastalık için de destek vektör makinası kullanılan yöntemler arasında en iyi başarımı sağlamıştır. İlgili yöntem prostat kanseri veri kümesinde %75.68'lık bir kesinlik değeri sunarken, melonom veri kümesi için %78,6'lik bir kesinlik değeri yakalamıştır.

Anahtar Kelimeler: veri madenciliği, karar ağacı, destek vektör makinası, naive bayes, kanser, bütünsel genom ilişkilendirme

Performance Comparison of Data Mining Methods in Diagnosis of Complex Diseases

Abstract

The data used in Genome Wide Association studies is vast in amount and high dimensional. Therefore, different data mining methods are used in order to find the relations between profiles and diseases. These methods are then used for diagnostic models. In this study two different data sets were used. The melonoma data set consists of 1025 cases and 531 controls. The multi ethnic prostate cancer data set consists of 2325 cases and 2350 controls. The underlying SNPs were searched by different data mining methods such as Decision Trees, Naive Bayes and Support Vector Machines. For both diseases support vector machine presented the best performance results. This method presented 75.68% of accuracy for prostate cancer data and 78.6% of accuracy for melonoma.

Keywords: data mining, decision tree, support vector machine, naive bayes, cancer, genome wide association

*Sorumlu Yazar (Corresponding Author): Sait Can Yücebaş
(e-posta: can@comu.edu.tr)
Bu çalışma ÇOMÜ BAP Koordinasyon Birimince desteklenmiştir. Proje numarası: FBA-2014-286

1. Giriş

Günümüzde genetik alanındaki çalışmalar İnsan Genom Projesi'nin tamamlanmasıyla birlikte büyük bir hız kazanmıştır. Bu çalışmaların bir kolu da genetik varyasyonları inceleyerek bunların hastalıklara yol açıp açmadığını inceleyen bütünsel genom ilişkilendirme (BGİÇ) çalışmalarıdır. Bu çalışmalarda karmaşık hastalıklara yol açan önemli faktörlerden biri olarak tekli nükleotit polimorfizmleri (TNP) gösterilmiş, TNP'lerin göz hastalıkları (Klein ve ark., 2005), kalp hastalıkları (Lettre ve ark., 2011), diyabet (Reddy ve ark., 2011), eklem iltihabı (Stahl ve ark., 2010), Crohn hastalığı (Lee, 2011), mental hastalıklar (Scott ve ark., 2009), MS hastalığı (Jakkula ve ark., 2010) ve birçok kanser türü (Yeager ve ark., 2007; Easton, 2008; Gerstenblith ve ark., 2010) ile ilişkisi literatürdeki farklı çalışmalarda incelenmiştir.

GWAS çalışmalarında kullanılan veri kümeleri, hem çok boyutlu hem de yüksek miktarda veri içermektedir. Bu nedenle ilgili verilerin incelenmesinde veri madenciliği yöntemlerinden yararlanılmaktadır. İlgili alanda en çok kullanılan yöntemler karar ağaçları (KA) (Huang ve ark., 2004; Fiaschi ve ark., 2009), Bayes temelli yöntemler (BA) (Jiang ve ark., 2010) ve destek vektör makinalarıdır (DVM) (Wei ve ark., 2009; Abeel ve ark., 2010).

Diğer yöntemlere göre daha basit, görsel olarak zengin ve daha az maliyetli olmalarıyla tercih edilen KA birçok BGİÇ çalışmasında kullanılmıştır. Yürütülen bir çalışmada hamile hipertansiyon hastalarının genotip özelliklerine göre teşhisi çalışılmış, geliştirilen model %65 duyarlılık ve %54 seçicilik performansı göstermiştir (Roberts, 1993).

2004 yılında yapılan bir çalışmada hipertansiyonlu 4529 hastanın genetik varyasyonları ID3, ADTree ve C4.5 gibi KA yöntemleri ile incelenmiş ve bu yöntemlerin performansları karşılaştırılmıştır (Huang ve ark., 2004).

Farklı KA yöntemlerinin incelendiği diğer bir çalışma hepatit hastalığı üzerinde yapılmış ve bu çalışmada 194 hasta ve 28 TNP incelenmiştir (Uhm ve ark., 2009).

KA yöntemleri rahim ağzı kanserine neden olan genetik faktörlerin bulunmasında karşılaştırılmış, hastalığa yol açan beş farklı yeni TNP bulunmuştur (Horng ve ark., 2011).

Farklı KA yöntemlerinin teşhis performanslarının karşılaştırıldığı bir diğer çalışma göğüs kanseri alanında yürütülmüştür. Bu çalışmada 258 hasta ve 32 TNP kullanılmış, karşılaştırılan yöntemler içerisinde en yüksek kesinlik ölçütünü C4.5 ağacı vermiştir (Anunciacao ve ark., 2010).

Otizm spektrum bozukluğu üzerinde yapılan bir çalışmada 36 hasta ve bunlara ait 25 TNP incelenmiş, genetik varyasyona bağlı teşhiste KA ile DVM yöntemlerinin karar verme performansları karşılaştırılmıştır. Bu çalışmada yöntemlerin birbirine yakın sonuçlar verdiği gözlemlenmiştir (Jiao ve ark., 2011).

TNP'lerin karmaşık hastalıklarla ilişkilendirilmesinde yaygın olarak kullanılan diğer bir veri madenciliği yöntemi de DVM'dir. DVM'nin TNP değişimlere dayanan teşhissel modellemesine ait birçok çalışma yapılmış ve ilgili yöntemin yüksek başarım performansı dikkat çekmiştir. Örneğin Tip-1 Diyabet üzerinde yapılan bir çalışmada DVM %71'lik kesinlik, %65'lik duyarlılık ve %77'lik seçicilik performansı göstermiştir. Bu çalışmada ayrıca doğrusal DVM ile doğrusal olmayan DVM yapıları karşılaştırılmış sonrasında daha iyi performans gösteren doğrusal olmayan DVM farklı bir yöntem olan lojistik regresyon ile kıyaslanmıştır. Kıyasla-

ma ROC eğrisi altında kalan alana göre yapılmış ve 0,86 – 0,89 arasında değişen duyarlılık, 0,85 – 0,88 seçicilik değerleri ile doğrusal olmayan DVM öne çıkmıştır (Wei ve ark., 2009).

Ağız kanseri üzerinde yapılan modellemede ve DVM %55,4 kesinlik sonucu ile %65.2 duyarlılık göstermiştir (Chuang, 2011)

DVM, performansının bu denli yüksek olması nedeniyle literatürde sıkça diğer veri madenciliği yöntemleri ile karşılaştırılmıştır. Bu çalışmaların birinde DVM, BA ve KA göğüs kanseri üzerindeki teşhis performansı karşılaştırılmış, ancak yöntemler arası anlamlı sonuçlar bulunamamıştır (Listgarten ve ark., 2011).

Olasılıksal modellemeye dayanan Bayes yöntemleri GWAS çalışmalarında hem yöntem karşılaştırma hem de hastalık ilişkilendirme için kullanılmıştır. Bir çalışmada 300.000 üzerinde TNP'nin alzaymır hastalığı ile ilişkisi incelenmiş, bu incelemede Naive Bayes (NB) yönteminin yanı sıra, özellik seçimli ve seçimsiz NB performansları karşılaştırılmıştır (Wei ve ark., 2011). Elde edilen sonuçlara göre model ortalaması kullanan NB, klasik NB'ye göre 0,72'lik ROC eğrisi altında kalan alan ile üstünlük sağlamıştır.

Yapılan diğer bir çalışmada hiyerarşik NB yöntemi geliştirilmiş, bu yöntemin performansı klasik NB ile farklı sayıda SNP içeren birçok veri setinde karşılaştırılmış ve son olarak gerçek veri setleri olan tip-1 ve tip-2 diyabet verileri üzerinde test edilmiştir (Malovini ve ark., 2014). NB yönteminin geliştirilerek kullanıldığı ve NB Torbası yönteminin geliştirildiği çalışmada ilgili yöntem biyo-işaretleyicilerin seçiminde kullanılmıştır (Sambo ve ark., 2012).

Literatürde bu yöntemlerin yanı sıra seçilen bir ana yöntemin genetik algoritma ile optimize edildiği genetik evrimli modeller (Turner ve ark., 2010; Jesus ve ark., 2007) ve birden fazla ana yöntemin birleştirildiği hibrit modeller (Yücebaş ve ark. 2014) de kullanılmaktadır.

Yapılan literatür taramasında edinilen bilgiler ışığında BGİÇ çalışmaları için en çok tercih edilen veri madenciliği yöntemlerinin KA, BA, DVM olduğu görülmüştür. İlgili modellerin karmaşık hastalıklardaki başarım kriterlerinin ortaya konması amacıyla yürütülen bu çalışmada ilgili yöntemler prostat kanseri ve melanom veri kümelerinde karşılaştırılmıştır.

2. Materyal ve Yöntem

2.1. Materyal

İlgili çalışma için iki farklı veri kümesi kullanılmıştır. Her iki veri kümesi de NCBI'ya ait dbGaP veri tabanından çekilmiştir. Bu veri kümelerinden ilki çok etnikli prostat kanseri verisi (Multi Ethnic Genome Wide Scan of Prostate Cancer) kümesidir. İlgili veri kümesi, Afro Amerikan, Japon ve Latin etnik kökenli, 4650 vaka ve 4795 kontrolden oluşmaktadır. Her kişi için yaklaşık 600.000 adet TNP bulunmaktadır. İkinci veri kümesi ise melanom hastalığını tanımlamaktadır. Bu veri kümesinde 2053 vaka ve 1062 kontrol bulunmaktadır. Etnik köken olarak Avrupa beyaz ırkını temsil etmektedir. Bu küme içerisindeki her birey için yaklaşık 600.000 adet TNP bulunmaktadır.

2.2. Yöntem

2.2.1 Ön İşleme

TNP profillerinin karmaşık hastalıklarla ilişkilendirildiği çalışmalarda veri kümesi içerisindeki vaka ve kontrol sayıları ile TNP sayısı fazla ise, eldeki veri kümesini en iyi şekilde temsil edecek bir alt veri kümesinin oluşturulması literatürdeki farklı çalışmalarca önerilmektedir (Wei ve ark., 2009; Zhou, 2007).

Bu doğrultuda bir veri ön işleme çalışması yapılmıştır. Çalışmanın ilk adımında veri kümeleri içerisindeki kişi sayısı azaltılmıştır. Bunun için veri kümesindeki vaka ve kontrol sayılarını yarı yarıya azaltacak bir örnekleme seçilmiştir. Örneklem seçiminde veri kümeleri içerisindeki vaka ve kontrol oranı ile etnik kökenlerin birbirine oranları korunmuştur. Belirtilen oranlar korunacak şekilde veri kümesindeki veriler rasgele olarak seçilerek, veri alt kümelerine atanmışlardır. Bu bağlamda prostat kanserini temsil eden alt veri kümesi 2325 vaka ve 2350 kontrolden oluşmuştur. Melonom için ise temsili veri alt kümesi 1025 vaka 532 kontrolden oluşturulmuştur.

Veri ön işlemenin ikinci adımında ise her iki kümede de birey başına bulunan yaklaşık 600.000 TNP'nin azaltılması amaçlanmıştır. Bu TNP'lerin tümünün belirlenen hastalıkla yüksek oranda ilişkili olmadığı gözlemlenmiştir. Bu gözlem içim METU – SNP (Ustunkar ve ark., 2011) test aracı kullanılmıştır. Bu araç, girdi olarak verilen TNP'lerin eldeki hastalıkla aralarındaki biyolojik ve istatistik ilişki analizi için kullanılmıştır. Yüzde 95 güven aralığında yürütülen bu analizin sonucunda prostat kanseri için kişi başına 2710 TNP ve melonom için kişi başına 2783 TNP anlamlı bulunmuştur.

2.2.2 Karar Ağacı Yöntemi

Karar ağaçları literatürde ikili sınıflama için oldukça tercih edilen bir yöntemdir (Rokach ve ark., 2002). Bu tercih altındaki en büyük etmenler gürültü toleransı, düşük hesaplama ihtiyaçları, uygulama kolaylığı ve sağladığı görsellik ile kolay yorumlanabilirliğidir (Coelho ve ark., 2012). Bu yöntemde eldeki her öznitelik sınıflama problemini ayırma gücü açısından özyineli olarak test edilir ve bilgi kazancı en yüksek olan öznitelik dallanma için seçilir (Quinlan, 1986). Daha formal bir notasyon ile açıklanacak olursa[1]:

Bu durumda örneklerin sınıflandırılması için gerekli bilgi şu şekilde hesaplanır:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (1)$$

Veri kümesini $\{ S_1, S_2, ..., S_v \}$şeklinde v adet alt kümeye bölecek A özniteliği için Entropi şu şekilde hesaplanır:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} I(s_{1j}i,,,s_{mj}) \quad (2)$$

s_{ij}, S_j alt kümesi içerisindeki bir C_i sınıfının örnek sayısını belirttiğinde, bu S_j alt kümesinin bilgisi şu şekilde hesaplanır:

[1] Bu alt başlıkta verilen matematiksel denklem ve notasyonlar "Data Mining: Concepts and Techniques, 3rd Ed. J.Han, M. Kamber. J. Pei." Kitabından uyarlanmıştır.

$$I(s_{1j}, s_{2j}, ..., s_{mj}\} = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3)$$

Bu durumda bilgi kazancı bilgiden entropi değerinin çıkarılması ile bulunur:

$$G(A) = I(s_1, s_2, ..., s_m\} - E(A) \quad (4)$$

S: s adet veri örneğinin bulunduğu küme
C_i: i=1 den m'e kadar sınıf etiketleri
S_i: C_i sınıfndaki örneklern sayısı
P_i : i verisinin sınıf C_i' ye ait olma olasılığı
A: Öznitelik A = {a_1, a2, ..., a_v}

Bu yöntemde eldeki her öznitelik sınıflama problemini ayırma gücü açısından test edilir ve bilgi kazancı en yüksek olan öznitelik dallanma için seçilir. Daha sonra kalan öznitelikler için yöntem öz yinelemeli olarak sürdürülür.

İlgili yöntemin uygulaması pyhton tabanlı bir açık kaynak kod aracı olan Orange Veri Madenciliği Aracı (Demsar ve ark., 2013) ile gerçekleştirilmiştir. Kullanılan veri kümesinde değerler ayrık değer aralığında olduğu için, karar ağacı için Id3 algoritması kullanılmıştır. Ağacın dallanması için bilgi kazanç oranı seçilmiş, yapraktaki en az örneklem sayısı 2 ve bir dallanmadaki oluşacak alt kümeler için en az örneklem sayısı 4, ağacının ulaşabileceği en büyük derinlik ise 100 olarak belirlenmiştir. İlgili modelin eğitimi ve sonuçların testi için 10 katmanlı çapraz geçerlilik testi uygulanmıştır.

2.2.3 Naive Bayes Yöntemi

Bu yöntem istatistiksel bir yöntem olan Bayes Teoremi'ne dayanmaktadır. Bu teorem belirli koşullar altında bir olayın gerçekleşme olasılığını gösterir. Örneğin eldeki sınıflandırma problemi prostat kanseri teşhisini PSA testinin sonuçlarına göre teşhis etmek olduğunda:

P(H): hastanın prostat kanseri olma olasılığı
P(V): PSA testi sonucunun yüksek çıkma olasılığı
P(H|V) : PSA testi yüksek çıktığı bilindiğinde kişinin prostat kanseri olma olasılığı

Sınıflama problemini özetleyen P(H|V) olasılığı Bayes Teoremi ile şu şekilde hesaplanır:

$$P(H|V) = \frac{P(V|H)P(H)}{P(V)} \quad (5)$$

Bu eşitlikte:

P(V|H): Hastanın prostat kanseri olduğu bilindiğinde PSA testinin yüksek çıkma olasılığı
P(H) : Prostat kanseri olma olasılığı
P(V) : PSA testi yüksek çıkma olasılığı

Naive Bayes yöntemi, Eşitlik 5'deki formülasyonu kullanarak eldeki sınıflama problemini çözmeye çalışır. Veri kümesinde *n* adet bağımsız sınıf C_1, C_2, ... , C_n ve sınıf bilgisi bilinmeyen yeni veri *D* olduğunda, tüm olasılıklar Eşitlik 6'daki gibi hesaplanır ve yeni veri en yüksek olasılığı veren sınıfa atanır.

$$P(C_i|D) = \frac{P(D|C_i)P(C_i)}{P(D)} \quad (6)$$

Formülasyondan doğan hesaplama karmaşıklığı ise tüm özniteliklerin bağımsız öznitelikler olduğu kabulü ile indirgenir, bu durumda da ilgili hesaplama şu şekilde yapılır:

$$P(D|C_i) = \prod_{k=1}^{n} P(d_k|C_i) \quad (7)$$

Bu kabul çoğu zaman gerçek örnekler için geçerli olmasa da ilgili yöntem daha karmaşık diğer yöntemlere yakın bir başarım performansı sergileyebilmektedir (Domingos, 1997).

İlgili modelin eğitimi ve sonuçların testi için 10 katmanlı çapraz geçerlilik testi uygulanmıştır.

2.2.4 Destek Vektör Makinası Yöntemi

Vapnik (1995) tarafından geliştirilen DVM, farklı sınıflara ait örnekleri birbirinden ayırabilecek ve bu sınıflara en uzak mesafede olacak hiperdüzlemi bulmaya çalışır. Bu hiperdüzlem:

D: Veri Kümesi
Yi : Sınıf belirteci
*Xi : veri noktasını gösteren p boyutlu reel vektör i*ken:

$$D = \{(x_i,y_i)|x_i \varepsilon R^p , y_i \varepsilon \{-1,1\}\}^n_{i=1} \text{ (Eşitlik 8)}$$

Bu durumda bir özellik uzatımdaki hiperdüzlem şu şekilde yazılabilir:

$$w.x-b=1 \quad (9)$$
$$w.x-b=-1 \quad (10)$$

Bu hiperdüzlemler arasındaki uzaklık ise:

$$\frac{2}{||w||} \quad (11)$$

Bu durumda eldeki optimizasyon problemi ||w|| minimize edilmesi halini alır. Eldeki sınıflama probleminin doğrusal olarak sınıflara ayrılamadığı durumda eldeki öznitelik uzayı daha yüksek boyutlara çıkarılmalıdır. Bunun için bir çekirdek yöntemi kullanılır (Baudat ve ark., 2001). Bu durumda X'in n boyutlu bir uzayda bir vektör ϕ(.)'nin ise girdi uzayından daha yüksek boyutlu bir uzaya eşleme yapan doğrusal olmayan bir fonksiyon olduğu kabul edildiğinde, sınıflama kararını çizecek hiperdüzlem şu şekilde ifade edilir:

$$w.\phi(x)-b = 0 \quad (12)$$

Bu eşitlikte w eğitim verisini yüksek boyutlu uzaya eşleyecek bir ağırlık katsayısı vektörünü temsil etmektedir. Bu durumda nokta çarpımı yerine çekirdek fonksiyonu kullanılırsa, girdi vektörünün yüksek boyutlu uzaya eşlenmesi gereği ortadan kaldırılabilir.

$$K(u,v) = \phi(u). .\phi(v) \quad (13)$$

Eldeki probleme göre kullanılabilecek birçok farklı çekirdek fonksiyonu bulunmaktadır (Müller ve ark., 2005). Bunların başlıcaları polinom, sigmoid, radyal temelli fonksiyon ve ANOVA'dır.

Polinom fonksiyonu verilerin normalize edildiği problemlerde yönsel çekirdek fonksiyonu olarak kullanılır. Homojen olan ve olmayan olmak üzere iki türü aşağıdaki eşitliklerde verilmiştir.

$$k(x_i,x_j) = (x_i.x_j)^d \quad (14)$$

$$k(x_i,x_j) = (x_i.x_{j+1})^d \quad (15)$$

Sigmoid: Yapay sinir ağları alanından adapte edilmiştir. Bu nedenle "*sigmoid fonksiyonu kullanan bir DVM aslında iki katmanlı perseptron sinir ağı ile eşdeğerdir* " (Lin ve ark., 2003). Sigmoid fonksiyonu:

$$k(c,x_j) = \tanh(kx_i.x_{j+c}) \quad (16)$$

Radyal Temelli Fonksiyon (RBF): Bu fonksiyon da yapay sinir ağları ile yakından ilişkili olup (Benoudjit ve ark., 2003) daha hızlı öğrenme özelliği ile bilinir (Park ve ark., 1991). Formülasyonu:

$$k(x,y) = \exp(-\gamma||x-y||^2 \quad (17)$$

ANOVA: Varyans analizi fonksiyonu RBF'in bir türevi olup özellikle yüksek boyutlu regresyon problemlerinin çözümünde tercih edilir (Hofmann ve ark., 2008).

$$k(x,y) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d \quad (18)$$

DVM uygulamalarında modelin başarılı bir performans sergilemesi için kullanılan çekirdek fonksiyonunun türü, sınıflar arasındaki sınırın uzaklığını belirleyen C katsayısı ve karar sınırının şeklini belirleyen Gamma katsayısının iyi ayarlanması gerekmektedir (Benhur ve Weston, 2010). Yapılan uygulamada çekirdek fonksiyonu olarak benzer veri kümeleri üzerindeki diğer çalışmalarda sıklıkla tercih edilen (Wei ve ark., 2009; Huang ve ark., 2009) radyal temelli fonksiyon kullanılmıştır. C ve gamma katsayıları sırası ile 10 ve 10^{-3} olarak belirlenmiştir (Yücebaş ve ark., 2014). İlgili modelin eğitimi ve sonuçların testi için 10 katmanlı çapraz geçerlilik testi uygulanmıştır.

3. Bulgular ve Tartışma

Önceki bölümde detayları verilmiş olan veri madenciliği yöntemlerinin başarımları çok etnikli prostat kanseri ve melonom veri kümesi üzerinde karşılaştırılmıştır. Yöntemlerin karşılaştırılmasında başarım kriteri olarak, kesinlik, duyarlılık ve ROC eğrisi altında kalan alan seçilmiştir.

Buna göre çok etnikli prostat kanseri veri kümesinde KA, BA ve DVM yöntemlerinin başarım kriterleri Çizelge 3.1'de sunulmuştur.

Çizelge 3.1. Prostat kanseri veri kümesi üzerinde KA, NB, DVM yöntemlerinin başarım kriterleri performansları

Yöntem / Ölçüt	Karar ağacı	Naive bayes	Destek vektör makinası
Kesinlik	%71,12	%65,12	%75,68
Duyarlılık	%60,82	%78,42	%68,91
ROC	0,811	0,605	0,835

Melonom veri kümesi üzerinde elde edilen bulgular Çizelge 3.2'de sunulmuştur.

Çizelge 3.2. Melonom veri kümesi üzerinde KA, NB, DVM yöntemlerinin başarım kriterleri performansları

Yöntem / Ölçüt	Karar ağacı	Naive bayes	Destek vektör makinası
Kesinlik	%72	%68,75	%78,6
Duyarlılık	%73,42	%67,72	%76,45
ROC	0,8	0,752	0,846

Kesinlik, duyarlılık ve ROC eğrisi altında kalan alan değerlerine göre yapılan bu karşılaştırmada DVM diğer yöntemlere göre daha iyi bir performans göstermiştir. Elde edilen sonuçlar değerlendirildiğinde DVM'nin yüksek performans vermesi şaşırtıcı değildir. Bu yöntemin doğrusal olarak ayrılmayan sınıflama problemlerinde iyi performans verdiği, benzer yöntemler arasında da global optimum değere en çok yakınsayan yöntem olduğu bilinmektedir (Xiao ve ark., 2010).

4. Sonuç

Bütünsel genom ilişkilendirme (GWAS), genetik değişimleri arayarak bu değişimlerin kanser vb karmaşık hastalıklarla ilişkili olup olmadığını arayan bir çalışma türüdür. Bu değişimler genellikle tek bir nükleotidin değişiminden kaynaklanır ve tekli nükleotit polimorfizm (TNP) olarak adlandırılır.

TNP profillerinin karmaşık hastalıklarla ilişkilerinin incelendiği çalışmalarda kullanılan veri kümeleri, veri sayısı bakımından kalabalık ve öznitelikler bakımından da oldukça yüksek boyutludur. Bu durum ilgili veri kümelerinin el ile veya basit istatistik kullanarak analizini çok zor hale getirmektedir. Bu nedenle ilgili analiz için çeşitli veri madenciliği yöntemlerinden yararlanılır.

Yapılan bu çalışmada literatürde sıkça tercih edilen ve her biri farklı algoritmik yöntemi temsil eden veri madenciliği yöntemlerinin (Özyineli mantıkla çalışan karar ağacı, olasılıksal metod olarak Bayes, olasılıksal olmayan model ve doğrusal olmayan verilerin sınıflandırması için de destek vektör makinası) başarım karşılaştırmaları melanom ve prostat kanseri veri kümeleri üzerinde yapılmıştır.

Başarım karşılaştırmasından önce veri kümelerine analize uygun hale getirmek adına veri kümelerini temsil edecek alt veri kümeleri oluşturulmuştur. Bu adımda alt veri kümesini oluşturacak veriler asıl veri kümelerinin vaka – kontrol ve etnik köken dağılımları korunacak şekilde rastgele olarak seçilmiştir. Veri kümeleri içerisinde her bireye ait yaklaşık 600.000 adet TNP bulunmaktadır. Eldeki hastalıkla yüksek ilişkisi olan TNP'ler istatistik ve biyolojik anlamlılıklarına göre seçilmiştir. Veri ön işleme adımından sonra prostat kanseri temsili alt kümesi 2325 vaka ve 2350 kontrolden oluşmuş, bu bireylere ait 2710 adet TNP kümeye dahil

edilmiştir. Melonom temsili kümesi ise 1025 vaka ve 531 kontrolden oluşurken 2783 TNP incelenmiştir.

Seçilen veri madenciliği yöntemlerinin eldeki hastalıkların teşhisindeki başarım performansları kesinlik, duyarlılık ve ROC eğrisi altında kalan alan ölçütleri ile karşılaştırılmıştır. Bu karşılaştırma sonucunda DVM her iki veri kümesi için de en iyi başarım performansı gösteren yöntem olarak bulunmuştur. Prostat kanseri için DVM %75,68'lik bir kesinlik değeri sağlarken, melonom için bu değer %78,6 şeklinde olmuştur. Doğrusal olmayan sınıflama problemlerinin çözümünde global optimum değere yakınsamadaki başarısı ile bilinen DVM'nin bu sonucu, literatürdeki benzer çalışmaların sonuçlarını destekler niteliktedir.

Teşekkür

Bu çalışma, Çanakkale Onsekiz Mart Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimince FBA-2014-286 No'lu proje olarak desteklenmiştir.

Kaynakça

Abeel T., Helleputte T., Van de Peer Y., Dupont P., Saeys Y., 2010. Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods. Advanced Access Publication. Bioinformatics. 26(3):392–398

Anunciacao O., Gomes B.C., Vinga S., Gaspar J., Oliveira A.L., Rueff J., 2010. A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups. In: Rocha M.P., Riverola F.F., Shatkay H., Corchado J.M. Eds. Advances in Bioinformatics. Advances in Intelligent and Soft Computing, Springer, Berlin, Heidelberg. 74: 43-51

Baudat G., Anouar F.M., 2001. Kernel-Based Methods and Function Approximation. International Joint Conference on Neural Networks. July 15-19. Washington D.C., USA

Ben-Hur A., Weston J., 2010. A User's Guide to Support Vector Machines. In: Carugo O., Eisenhaber F. Eds. Data Mining Techniques for the Life Sciences. Methods in Molecular Biology (Methods and Protocols), Humana Press. 609:223-239

Benoudjit N., Verleysen M., 2003. On The Kernel Widths in Radial-Basis Function Networks. Neural ProcessingLetters 18: 139–154

Chuang L.Y., 2011. Support Vector Machine-Based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes. Proceedings of International Multiconference of Engineers and Computer Scientists. March 16-18. Hong Kong, China

Coelho R., Basgalupp M.P., Carvalho A., Freitas A.A., 2012. Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactıons on Systems, Man, and Cybernetıcs—Part C: Applıcatıons and Revıews. 42(3): 291-312

Demsar J., Curk T., Erjavec A., Gorup C., Hocevar T., Milutinovic M., Možina M., Polajnar M., Toplak M., Starič A., Štajdohar M., Umek L., Žagar L., Žbontar J., Žitnik M., Zupan B., 2013. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research: 234 – 2353.

Domingos P., Pazzani M., 1997. On The Optimality of the Simple Bayesian Classifier Under Zero-One Loss. Machine Learning. 29(2):103–130

Easton D.F., Eeles R.A., 2008. Genome-Wide Association Studies in Cancer. Oxford Journals Life Sciences and Medicine Human Molecular Genetics. 17(R2):R109-R115

Fiaschi L., Garibaldi J. M., Krasnogor N., 2009. A Framework for the Application of Decision Trees to the Analysis of SNPs Data. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 30 March – 2 April. Nashville, TN, USA

Gerstenblith M.R., Shi J., LAndi M.T., 2010. Genome-Wide Association Studies of Pigmentation and Skin Cancer: A Review and Meta-Analysis. Pigment Cell & Melanoma Research. 23(5): 587–606

Guillaume L., Palmer C.D., Young T., Ejebe K.G., Allayee H., Benjamin E.J., 2011. Genome Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. Plos Genetics 7(2): e1001300

Hofmann T., Scholkopf B., Smola A .J., 2008. Kernel Methods in Machine Learning. The Annals of Statistics. 36(3):1171-1220

Horng J.T., Hu K.C., Wu L.C., Huang H P., Lin F.M., Huang S.L., Lai H.C., Chu T.Y., 2004. Identifying The Combination of Genetic Factors That Determine Susceptibility to Cervical Cancer. IEEE Transactions on Information Technology in Biomedicine. 8(1): 59-66

Huang J., Lin A., Narasimhan B., Quertermous T., Hsiung C.A., Ho L.T., Grove J.S., Oliver M., Ranade K., Risch N.J., Olshen R.A., 2004. Tree-structured supervised learning and the genetics of hypertension. Proceedings of the National Academy of Sciences of the United States of America. July 12. 101(29):10529–10534

Huang L. C., Hsu S. Y., Lin E., 2009. A Comparison of Classification Methods for Predicting Chronic Fatigue Syndrome Based on Genetic Data. Journal of Translational Medicine. 7:81

Jakkula E., Leppa V., Sulonen A.K., Varil T., 2010. Genome-wide Association Study in a -Risk Isolate for Multiple Sclerosis Reveals Associated Variants in STAT3 Gene. The American Journal of Human Genetics. 86: 285–291

Jesus K., Juan C. F.L., Enrique H.L., 2007. GPDTI: A Genetic Programming Decision Tree InductionMethod to Find Epistatic Effects in Common Complex Diseases. Bioinformatics.123(13):167-174

Jiang X., Barmada M. M., Visweswaran S., 2010. Identifying Genetic Interactions in

Genome-Wide Data Using Bayesian Networks. Genet Epidemiol, 34(6): 575–581

JiaoY., Chen R., Ke X.,Cheng L., ChuK., Lun Z., Herskovits E.H., 2011. Predictive Models for Subtypes of Autism Spectrum Disorder Based on Single-Nucleotide Polymorphisms and Magnetic Resonance Imaging. Advances in Medical Sciences. 56: 334-342

Klein R.J., Zeiss C., Chew E.Y., Tsai J.Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., Bracken M.B., Ferris F.L., Ott J., Barnstable C., Hoh J., 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 308 (5720): 385–9

Lee J. C., Parkes M., 2011. Genome-Wide Association Studies and Crohn's Disease. Oxford Journals Life Sciences Briefings in Functional Genomics. 10(2):71-76

Lin H., Lin C., 2003. A Study on Sigmoid Kernels for SVM and the Training of non- PSD Kernels by SMO-type Methods. Technical report.

Listgarten J., Damaraju S., Poulin B., Cook L., 2011. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clinical Cancer Reseach. 10:2725–2737

Malovini A., Barbarini N., Bellazzi R., Michelis F., 2014. Hierarchical Naive Bayes for Genetic Association Studies. BMC Bioinformatics. 13(Suppl 14): S6

Muller K. R., Mika S., Ratsch G., Tsuda K., Scholkopf B., 2005. An Introduction to Kernel-Based Learning Algorthims. IEEE Transactions on Neural Networks. 12(2): 181–201

Park J., Sandberg I.W., 1991. Universal Approximation Using Radial-Basis-Function Networks. Neural Comput. 3:246–257

Quinlan J.R., 1986. Induction of Decision Trees. Machine Learning. 1(1):81-106

Reddy MV, Wang H., Liu S., Bode B., Reed J.C., Steed R.D., Anderson S.W., Steed L., Hopkins D., She J.X., 2011. Association between Type 1 Diabetes and GWAS SNPs in the Southeast US Caucasian Population. Genes and Immunity. 12(3):208-212

Roberts J.M., Redman C.W. G., 1993. Pre-Eclampsia: More Than Pregnancy-Induced Hypertension. The Lancet. 341(8858):1447 – 1451

Rokach, L., Maimon, O., 2002. Top-Down Induction of Decision Trees Classifiers. IEEE

Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 35(4):476- 487

Sambo F., Trifoglio E., Di Camillo B., Toffolo G.M., Cobelli C., 2012. Bag of Naïve Bayes: Biomarker Selection and Classification from Genome-Wide SNP Data. BMC Bioinformatics. 13(Suppl 14):S2

Scott L. J., Muglia P., Kong X.Q., 2009. Genome-Wide Association and Meta-Analysis of Bipolar Disorder in Individuals of European Ancestry. PNAS. 106 (18): 7501–7506

Stahl E. A., Raychaudhuri S., Remmers E.F., 2010. Genome-Wide Association Study Meta-Analysis Identifies Seven New Rheumatoid Arthritis Risk Loci. Nature Genetics 42(10):508–514

Turner S. D., Dudek S. M., Ritchie M. D., 2010. ATHENA: A Knowledge-Based Hybrid Backpropagation-Grammatical Evolution Neural Network Algorithm for Discovering Epistasis among Quantitative Trait Loci. BioData Mining 3:5

Uhmn S., Kim D.H., Ko Y.W., Cho S., Cheong J., Kim J., 2009. A Study on Application of Single Nucleotide Polymorphism and Machine Learning Techniques to Diagnosis of Chronic Hepatitis. Expert Systems. 26(1)

Ustünkar G, Aydın Son Y., 2011. METU-SNP: An Integrated Software System for SNP-Complex Disease Association Analysis. J Integr Bioinform, 8(1):187

Vapnik V., Cortes C., 1995. Support-Vector Networks. Machine Learning. 20(3):273-297

Wei W., Visweswaran S., Cooper G. F., 2011. The Application of Naive Bayes Model Averaging to Predict Alzheimer's disease from Genome-Wide Data. JAm Med Inform Assoc. 18(4): 370–375

Wei Z., Wang K., Qu H.Q., Zhang H., 2009. From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. Plosone. 5(10): e1000678

Xiao R., Wang J., Zhang F., 2010. An Approach to Incremental SVM Learning Algorithm. 12th IEEE Proceedings on Tools with Artificial Intelligence. 268-273

Yeager M., Orr N., Hayes R.B., 2007. Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24. Nature Genetics 39: 645 – 649

Yücebaş S. C., Aydın Son Y., 2014. A Prostate Cancer Model Build by a Novel SVM ID3 Hybrid Feature Selection Method Using Both Genotyping and Phenotype Data from dbGaP. PLoS ONE 9(3): e91404

Zhou N., Wang L., 2007. Effective Selection of Informative SNPs and Classification on the Hapmap Genotype Data. BMC Bioinformatics.8:484