DEÜ MÜHENDİSLİK FAKÜLTESİ
MÜHENDİSLİK BİLİMLERİ DERGİSİ
Cilt: 16 Sayı: 48 sh. 30-41 Eylül 2014

BİYOMEDİKAL VERİ KÜMELERİ İLE MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ İSTATİSTİKSEL OLARAK KARŞILAŞTIRILMASI

(STATISTICAL COMPARISON OF MACHINE LEARNING CLASSIFICATION ALGORITHMS USING BIOMEDICAL DATA SETS)

Murat KARAKOYUN[1], Mehmet HACIBEYOĞLU[2]

ÖZET/*ABSTRACT*

Günümüzde bilişim teknolojileri hemen hemen her alanda kullanılmaktadır. En çok kullanılan alanlardan bir tanesi de sağlık sektörüdür. Dijital hastane sistemlerinin kullanılmasıyla birlikte hasta verileri artık bilgisayar ortamında saklanmakta ve böylelikle biyomedikal veri kümeleri oluşmaktadır. Boyut olarak çok büyük olan bu veri kümelerinin bir insan tarafından analiz edilmesi ve yorumlanması çok zordur. Bunun için bilgisayar mühendisliği çalışma alanlarından biri olan makine öğrenmesi algoritmaları kullanılır. Bu çalışmada 6 tane makine öğrenmesi algoritmalarının başarımları 9 farklı biyomedikal veri kümesi üzerinde test edilmiştir ve elde edilen sonuçlar istatistiksel olarak karşılaştırılmıştır. Deneysel ve istatistiksel sonuçlar birlikte incelediğinde küçük ve orta büyüklükteki biyomedikal veri kümeleri için Yapay Sinir Ağları algoritması sınıflandırma başarımı açısından ve K-en Yakın Komşu algoritması ise çalışma zamanı açısından daha başarılı olmuştur. Bu çalışmanın bir bölümü ASYU 2014/İzmir sempozyumunda bildiri olarak sunulmuştur.

Nowadays, information technology is used in nearly every field. One of the most used is the health sector. Patient datas are stored in computers with the using of digital hospital systems and in this way biomedical data sets are consisted. The size of these data sets is too large and it is very difficult to be analyzed and interpreted by a human. The machine learning algorithms which are workspace of computer engineering are used for analyzing and interpreting these data sets. In this study 6 machine learning algorithms' performance has been tested with using 9 different biomedical data sets and the obtained results were compared statistically. According to the experimental and statistical results of this study, for the small and medium sized datasets Artificial Neural Network algorithm and K-Nearest Neighbor algorithm are succeeded in terms of classification accuracy performance and cpu time performance, respectively. A part of this work was presented at the ASYU 2014/Izmir symposium.

ANAHTAR KELİMELER/*KEYWORDS*

Makine Öğrenmesi, Sınıflandırma, Biyomedikal
Machine Learning, Classification, Biomedical

[1] Necmettin Erbakan Ü. Müh. Mim. Fak. Bilgisayar Mühendisliği, mkarakoyun@konya.edu.tr
[2] Necmettin Erbakan Ü. Müh. Mim. Fak. Bilgisayar Mühendisliği, hacibeyoglu@konya.edu.tr

1. GİRİŞ

Günümüzde sağlık alanındaki problemlerin çözümü için, özellikle hastalıkların teşhisi konusunda, doktorların kararına destek olabilecek makine öğrenmesi temelli önemli çalışmalar yapılmaktadır. Makine öğrenmesinin bir alt dalı olan sınıflandırma algoritmaları da bu alanda kullanılmaktadır. Sınıflandırma algoritmaları öncelikle hastalıklar ile ilgili geçmişe yönelik hasta verilerini kullanarak kendisini eğitir. Eğitimi tamamlanmış sınıflandırma algoritması daha sonra olası hastalar veya hastalıklar için tahminde bulunur. Doktorlara hastalık teşhisinde yardımcı olan sınıflandırma algoritmaları temelli bu sistemlere karar destek sistemleri adı verilir. Karar destek sistemlerini oluşturan sınıflandırma algoritmaları konusunda biyomedikal veri kümeleri konusunda günümüze kadar birçok çalışma gerçekleştirilmiştir.

Jin Huang ve arkadaşları Naive Bayes (NB), Karar Ağaçları ve Destek Vektör Makinelerini (DVM) kullanarak çeşitli veri kümeleri üzerinde veri sınıflandırma yapmışlardır. Bu çalışmada DVM algoritmasının, uygulanan veri kümeleri için NB ve C4.5 algoritmalarından daha yüksek başarımlı olduğu ancak yapılan istatistiksel testler sonucunda aradaki farkın kayda değer oranlarda olmadığı bulunmuştur (Huang vd., 2003).

E. Kaya ve arkadaşları, NB, k-En Yakın Komşu (k-EYK) , C4.5 Karar Ağaçları Algoritması ve DVM algoritmalarını kullanarak Parkinson veri kümesi üzerinde sınıflandırma yapmışlardır. Bu çalışmada sınıflandırma öncesi veriler üzerinde ayrıştırma yapmanın sınıflandırmaların başarısı üzerine etkileri gözlemlenmiştir. Çalışmaya göre Parkinson veri kümesi üzerinde yapılan ayrıştırma işlemi kullanılan tüm sınıflandırıcı algoritmalar için iyi sonuçlar vermesine karşın, en iyi sonucu DVM algoritması elde etmiştir (Kaya vd., 2011).

Kai Fu ve arkadaşları, Hilbert -Huang dönüşüm metodu ile beraber DVM algoritmasını kullanarak EEG veri kümesi üzerinde sınıflandırma işlemi gerçekleştirmiştir. Bu çalışmadan elde edilen sonuçlara göre kullanılan yöntemin, % 99.125 gibi iyi başarım oranına ulaştığı görülmüştür ve daha önce üzerinde çalışılan birçok yaklaşımdan daha iyi olduğu gözlemlenmiştir (Fu vd., 2014).

I. Nitze ve arkadaşları, arazi örtüsü görüntülerini kullanarak Rastgele Orman (RO) algoritması ve MODIS zaman serisi ile sınıflandırma yapmışlardır. Çalışmada RO algoritması kullanılarak NDVI ve EVI zaman serileri kıyaslanmıştır (Nitze vd., 2015).

U. Rajendra Acharya ve arkadaşları, Yapay Sinir Ağları (YSA) ve bulanık denklik bağıntısını kullanarak kalpten elde edilen veriler üzerinde hastalık teşhisine yönelik sınıflandırma işlemi gerçekleştirmişlerdir. Çalışmadan elde edilen sonuçlara göre YSA algoritmasının ortalama % 85 civarında başarım oranı gösterirken bulanık denklik sisteminin başarımının ortalama %90 oranlarında olduğu gözlemlenmiştir (Rajendra vd., 2003).

Bu çalışmada 9 adet biyomedikal veri seti üzerinde deneyler yapılmıştır. Kullanılan sınıflandırma algoritmaları ise literatürde sıklıkla kullanılan k-EYK, NBCN2, DVM, RO ve YSA algoritmalarıdır (Altman, 2007; Farid vd., 2014; Clark ve Niblett, 1989; He vd., 2014; Breiman, 2001; McCulloch ve Pitts, 1990).

İkinci bölümde kullanılan biyomedikal ve sınıflandırma algoritmaları açıklanacaktır. Daha sonra üçüncü bölümde, yapılan deneysel çalışmalar gösterilecek ve algoritmaların istatistiksel karşılaştırmaları dördüncü bölümde yapılacaktır. En son bölüm olan beşinci bölümde ise yapılan çalışma yorumlanacak ve bu çalışmayla ilgili gelecekte yapılabilecekler açıklanacaktır.

2. YÖNTEM

2.1.Materyal

Bu çalışmada kullandığımız, UCI veri ambarından alınan veri kümelerinin sınıf sayısı, özellik sayısı ve örnek sayısı özellikleri Çizelge 1'de gösterilmektedir.

Çizelge 1. Kullanılan veri kümelerinin özellikleri

No	Veri Seti	Sınıf Sayısı	Özellik Sayısı	Örnek Sayısı
1	Dermatology	6	34	366
2	Echocardio	2	12	132
3	Breast cancer	2	10	699
4	Heart Disease	5	13	303
5	Parkinson	2	22	197
6	Diabetes	2	19	155
7	Thyroid	3	5	215
8	Pima	2	8	768
9	Bupa	2	6	345

Dermatology: Ciltte oluşan kızarıklık, sertleşme, kaşıntı, yumuşama, kepeklenme, çatlama; kafa derisi, diz ve el dirseklerinde hastalığın görülme düzeyi gibi medikal özellikler kullanılarak hastanın hastalık sınıfı tahmin edilmektedir. Hastalık sınıfları: Sedef Hastalığı, Seboreik Egzama, Liken Planus, Pitriyazis Rozea, Kronik Deri İltihabı, Pitriyazis Rubra Pilar.

Echocardio: Kalp krizi yaşandıktan sonra hastanın kalp hareketlerinden oluşan nitelikler kullanılmaktadır. Örnek olarak kalp kesesinde sıvı toplanıp toplanmadığı, kalpte kısmı yağlanma oranı, kalbin kasılma durumu, kalbin sol karıncıklarındaki hareket durumu gibi değerler göz önünde bulundurulur. Bu özelliklerin yanında hastanın krizden sonra hayatta olup olmadığı, değilse hayatta kaldığı süre dikkate alınmaktadır. Tüm bu özelliklerin kullanılması ile hastanın 1 yıla yakın bir süre içinde hayatta kalıp kalmadığı sonucu elde edilmektedir.

Breast Cancer: Hastalık bölgesindeki tümörlü alanın özellikleri sınıflandırmada nitelik olarak kullanılmaktadır. Örneğin: tümörlü bölgenin kalınlığı, tümörlü hücrelerin şekil ve büyüklük olarak birbirine benzemesi oranı, tümörün vücuda yapışma oranı gibi özellikler. Bu özelliklerin kullanılması ile tümörün iyi huylu veya kötü huylu olması sonucu elde edilir.

Heart Disease: Hastanın yaş ve cinsiyet gibi özelliklerinin yanında göğsündeki ağrı tipi, kandaki en düşük ve en yüksek şeker seviyesi ile tansiyon değerleri gibi verilerin nitelik olarak kullanılmasıyla hastanın kalp rahatsızlık seviyesi tahmin edilmektedir. 0 (sıfır) kalpte rahatsızlığın hiç olmaması olmak üzere, kalpteki rahatsızlık oranı 1, 2, 3 ve 4 değerleri ile nitelendirilmektedir.

Parkinson: Hastanın konuşma durumunda sesinde oluşan dalgalanmalar, duyarlı olduğu maksimum, minimum, ortalama ses frekansları gibi değerleri nitelik olarak kullanarak hastanın parkinson hastası olma veya sağlıklı bir insan olma tahminini gerçekleştirmek üzere kullanılmaktadır.

Diabetes: Hastanın yaş ve cinsiyet özelliklerinin yanı sıra vücutta rahatsızlık, kırgınlık, iştahsızlık, karaciğerde büyüklük olup olmadığı; safra düzeyindeki düşüş miktarı, vücuttaki fosfat ve albümin gibi bazı değerlerin miktarlarının özellik olarak kullanıldığı bir veri kümesidir.

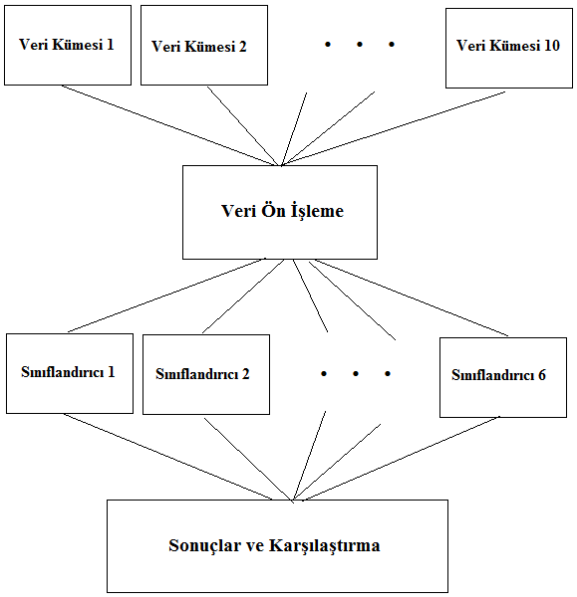
Thyroid: RT3U (tiroid hormonlarını birbirine bağlayan uçların ölçümü) testinden elde edilen değerler ile öz sıvıdaki bazı değerlerin özellik olarak kullanılması ile hastanın normal olduğu, yüksek veya düşük derecede tiroit hastalıklı olduğu tahmini için kullanılan veri kümesidir.

Pima: Hastanın hamilelik sayısı, yaşı, öz sıvıdaki insülin miktarı, glükoz yoğunluğu, derideki kıvrım kalınlığı ve vücut kitle indeksi gibi değerleri özellik olarak kullanan bir veri kümesidir.

Bupa: Hastanın kanı ile ilgili yapılan testlerden elde edilen sonuçlar ile günlük alınan alkol miktarının özellik olarak kullanılması sonucu oluşan bir veri kümesidir.

2.2. Metot

Çalışmamızın akış diyagramı Şekil 1'de gösterilmektedir.



Şekil 1. Yapılan çalışmanın akış diyagramı

2.2.1. K-En Yakın Komşu Algoritması

K-En Yakın Komşu algoritması öznitelik uzayındaki en yakın eğitim örneklerine dayanarak nesneleri sınıflandıran, en basit örüntü tanıma yöntemlerinden birisidir. Bu algoritma verilen *k* değeri kadar en yakın komşunun sınıfına göre sınıflandırma işlemi yapmaktadır. k-EYK algoritmasında bir vektörün sınıflandırılması, sınıfı bilinen vektörler kullanılarak yapılmaktadır. Test edilecek örnek, eğitim kümesindeki her bir örnek ile tek tek işleme alınır. Test edilecek örneğin sınıfını belirlemek için eğitim kümesindeki o örneğe en yakın *k* adet örnek seçilir. Seçilen örneklerden oluşan küme içerisinde hangi sınıfa ait en çok örnek varsa test edilecek olan örnek bu sınıfa aittir denilir. Örnekler arası uzaklıklar *Öklid (Euclidean)* uzaklığı ile bulunur (Küçük vd., 2013). Eşitlik 1, n boyutlu 2 nokta arasındaki uzaklığı veren Öklid uzaklık formülüdür.

$$d(x, y) = \sqrt{\{\sum_i^n (X_i - Y_i)^2\}} \quad (1)$$

2.2.2. Naive Bayes Algoritması

Naive Bayes Sınıflandırıcısı Bayes teoremine dayanan basit bir olasılıksal sınıflandırma yöntemidir. Mevcut sınıflanmış durumdaki örnek verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine ait olma olasılığını hesaplayan bir yaklaşımdır. Bu sınıflandırıcıda nitelikler birbirinden bağımsız olarak kabul edilir. Örneklerin hepsi aynı derecede öneme sahiptir. Bir özelliğin değeri başka bir özellik değeri hakkında bilgi içermez.

Her biri n adet nitelikten oluşan ve m adet sınıftan herhangi birine dâhil olan bir veri seti üzerinde çalıştığımızı düşünelim. Bu durumda hangi sınıfa ait olduğu bilinmeyen yeni bir X örneği sınıflandırılmak istendiğinde, Eşitlik 2 kullanılarak örneğin her sınıf için, o sınıfa ait olma olasılığı hesaplanır. Bu değerler içerisinde en yüksek olasılığa sahip olan sınıf örneğin ait olduğu sınıf olarak kabul edilir.

$$P(S_i|X) = \frac{P(X|S_i)*P(S_i)}{P(X)} \quad (2)$$

$P(S_i|X)$: X olayı gerçekleştiğinde S_i olayının gerçekleşme olasılığı,
$P(X|S_i)$: S_i olayı gerçekleştiğinde X olayının gerçekleşme olasılığı,
$P(S_i)$,$P(X)$: S_i ve X olaylarının önsel olasılığıdır.

Her bir X örneğinin aynı derecede öneme sahip olması sebebiyle P(X) değeri her örnek veri için aynıdır. Bu durumda Eşitlik 2, Eşitlik 3 biçimine sadeleştirilebilir.

$$P(S_i|X) = P(X|S_i) * P(S_i) \quad (3)$$

Her bir sınıf için Eşitlik 3 uygulanıp olasılıklar hesaplandıktan sonra örneğin ait olduğu sınıf bulunur (Bermejo vd., 2011).

2.2.3. CN2 Algoritması

CN2 bir öğrenme algoritmasıdır ve üzerinde çalışılacak veri kümesinde kurallar oluşturmaya yönelik geliştirilmiştir. Veri kümesindeki gürültülü örneklerden kaynaklanan problemlerle ilgilenir. Bunun için bir dizi kural oluşturur ve karar ağacında budama yapabilecek istatistiksel teknikler sunar (Hacıbeyoğlu vd., 2011).

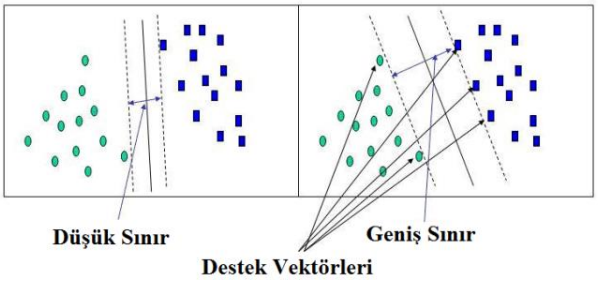
2.2.4. Destek Vektör Makineleri (DVM)

DVM, Vapnik tarafından geliştirilmiş ve istatistiksel öğrenme teorisi alanında ortaya çıkmış bir öğrenme metodudur (Cortes ve Vapnik, 1995). Sınıflandırma için, bir düzlemde bulunan örnekler arasına bir sınır çizerek iki gruba ayırır. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. Sınırın çizilmesi için iki gruba da yakın ve birbirine paralel iki çizgi çizilir ve bu çizgiler birbirine yaklaştırılarak sınır çizgisi üretilir. DVM yöntemi, veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır.

DVM, basit bir yapısı olması ve pratik uygulamalarda yüksek performans göstermesi bakımından oldukça kullanışlıdır. DVM'lerde kullanılacak örnek sayısı önemli değildir. DVM eğitim esnasında görülmemiş verileri de sorunsuz olarak sınıflandırır. Bu DVM'nin genelleştirebilme yeteneğini gösterir. Genelleştirebilme özelliği DVM' yi diğer tekniklere göre iyi bir alternatif yapmaktadır (Kecman, 2001).

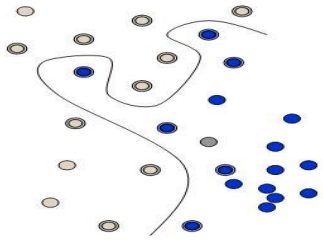
Sınıflandırılacak örnekler Şekil 2'deki gibi doğrusal bir düzlemle ayrıştırılabilecek seviyede olabilir. DVM analizi, olayları ayıran 1-boyutlu düzlemi, hedef kategorilerini temel alarak

bulmaya çalışır. Mümkün çizgilerin sınırsız sayısı vardır. Hangi çizginin, daha iyi olduğunu bulmak ve optimal çizgiyi nasıl bulacağımız önemlidir. Noktalı gösterilen çizgiler en yakın vektörler arasında mesafeyi ayıran çizgiye paralel olarak çekilir. Noktalı çizgilerin arasındaki mesafe kenarı çağırır. Kenarın genişliğini zorlayan vektörler, destek vektörleridir (DTREG, 2014).



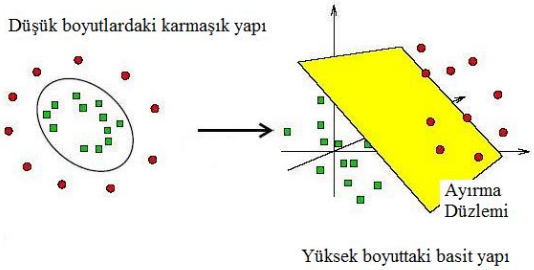
Şekil 2. Destek vektörlerinin gösterimi

Sınıflandırılacak örnekler Şekil 3'deki gibi doğrusal bir çizgi ile ayrılamayacak durumda olabilir.



Şekil 3. DVM doğrusal olmayan sınıflandırma örneği

Bu durumda doğrusal olmayan bir çizgiye ihtiyaç duyulur. Veriye doğrusal olmayan eğrilerle uymaktansa DVM'yi çekirdek fonksiyonu ile başka bir uzaya taşıyarak daha tutarlı bir ayrım sağlanmış olunur. Sekil 4'de çekirdek fonksiyonlarının üst boyuta taşınması gösterilmektedir (Kecman, 2001).



Şekil 4. Çekirdek fonksiyonun üst boyuta taşınması

2.2.5. Rastgele Orman Algoritması

Rastgele Orman Algoritması, birçok karar ağacı yapısını kullanarak sınıflandırmada başarı oranını yüksek seviye ile yakalayan bir algoritmadır (Breiman, 2001).

Rastgele orman algoritmasında, diğer karar ağaçlarına benzer şekilde, dallanma kriterlerinin belirlenmesi ve uygun bir budama yönteminin seçilmesi önemlidir. Dallanma kriterlerinin belirlenmesinde *Gini Katsayısı* yöntemi kullanılmaktadır (Mather, 2005).

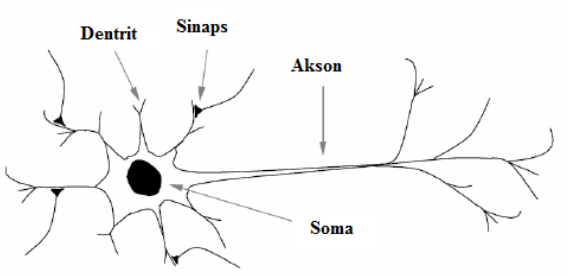
Algoritmada ağaç yapısının oluşturulması için her bir düğümde kullanılacak örneklerin sayısı ve oluşturulacak ağaç sayısının belirlenmesi gerekir. Sınıflandırma sırasında karar ormanı, kullanıcı tarafından belirlenen K adet ağaçtan oluşturulur. Yeni bir nesne sınıflandırılacağı zaman bu K adet karar ağacı tarafından işleme tabi tutulur ve her ağaçtan elde edilen oranlar içerisinde en yüksek olanı seçilerek sınıf belirlenmesi yapılır (Pal, 2005; Çölkesen, 2009).

Rastgele ormanda kullanılacak veri seti için eğitim ve test verisi önceden belirlenmemişse, veri setindeki sınıf oranları dikkate alınarak tüm veri setinin 2/3'si eğitim (inBag) ve 1/3'i test verisi (OutOfBag, OOB) olarak kullanılır. Karar ormanını oluşturacak K tane karar ağacı için, gene K adet bootstrap tekniği kullanılarak örneklem oluşturulur ve her bir örneklem için inBag ve OOB verisi ayrılır. Tüm ağaçlar ayrılan OOB verisi ile test edilerek hata oranı hesaplanır ve ardından bu hata oranlarının ortalaması alınarak karar ormanının OOB hatası hesaplanır. Hesaplanan OOB hata oranına göre tüm ağaçlara bir ağırlık verilir. Hata oranı ve ağaca verilen ağırlık değeri ters orantılıdır. Hata oranı en yüksek olan karar ağacı en düşük ağırlığı, hata oranı en düşük olan karar ağacı ise en yüksek ağırlığı alır. Belirlenen ağırlıklara göre tüm ağaçlar sınıflandırma işlemi için bir oylama işleminden geçirilir. En yüksek oyu alan ağaç sınıf tahmini olarak belirlenmiş olur (Akman vd., 2011).

2.2.6. Yapay Sinir Ağları

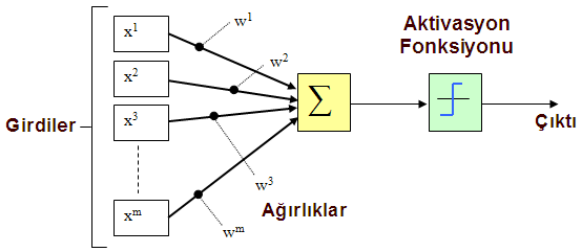
Yapay sinir ağları; insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleridir. Bu yetenekleri geleneksel programlama yöntemleri ile gerçekleştirmek oldukça zordur. Bu sebeple yapay sinir ağlarının, programlanması çok zor veya mümkün olmayan olaylar için geliştirilmiş bilgisayar tabanlı, adaptif bir bilim dalı olduğu söylenebilir (Öztemel, 2012).

Şekil 5' de insanın sinir hücresinin biyolojik yapısı gösterilmektedir.



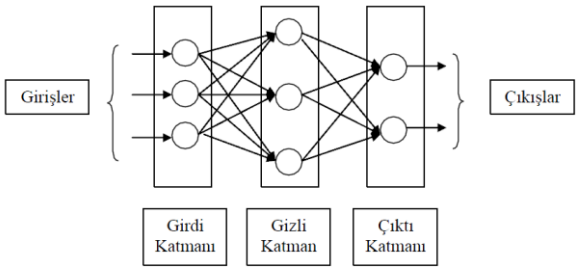
Şekil 5. Biyolojik sinir hücre yapısı

Yapay sinir ağları, insanın biyolojik sinir hücrelerinden yola çıkılarak ortaya konulan bir makine öğrenmesi modelidir. İlk modelleri algılayıcı (perceptron) adını alan, beyin sinir hücrelerine benzer bir yapıdadır. Sonraki gelişmeler ve modeller de temelde aynı yapıdan beslenir. Şekil 6' da, bir algılayıcıya ait, girdiler, bu girdileri işleyen bir transfer fonksiyonu, maliyet fonksiyonu ve sistem sonucunda üretilen çıktı değeri/değerleri vardır. Biyolojik modelin matematiksel gösterime çevrilmiş hali olarak da ifade edilebilir (Han ve Kamber, 2006; Gershenson, 2014).



Şekil 6. YSA'ya ait bir algılayıcının temel yapısı

Şekil 7' de bir YSA' nın temel yapısı gösterilmektedir.



Şekil 7. YSA'nın temel yapısı

Eğiticili öğrenme yaklaşımı kapsamında, eğitim için kullanılacak veri kümesi belirlenir ve yapay sinir ağına uygulanır. Ağ öğrenmeye başladığında çeşitli kontrol mekanizmaları ile öğrenme oranı, hata oranı, denetlenir ve belirli bir noktada işlem sonuçlanır. Kontrol mekanizmaları, ağ iterasyon sayısı, öğrenme oranı, hata eşik değeri gibi elemanlardan oluşmaktadır. İterasyon sayısında, ileri beslemeli algoritmanın, mevcut veri kümesi üzerine kaç sefer uygulanacağı belirlenebilir. Öğrenme oranı, ağın hangi hızda öğrenmeyi gerçekleştireceğini ifade eder ve genelde 0,2-0,4 arası bir değer alması önerilir (Han ve Kamber, 2006; Gershenson, 2014).

Algılayıcı, girdi ve çıktıları olan, yapay sinir ağlarının temel işlemci birimidir. x_j: j = 1,….,d girdi birimlerini gösterir. x_0 her zaman 1 değerini alan ek girdidir. w_j, x_j girdi biriminin ağırlığı, y de çıktı birimidir. y çıktı birimi en basit durumda girdilerin ağırlıklı toplamları olarak hesaplanır. Eşitlik 4 ile bir algılayıcıya ait giriş ve çıkışların değer hesaplaması yapılır (Alpaydın, 2010).

$$y = \sum_{j=1}^{d} w_j x_j + w_0 \quad (4)$$

YSA, eğitildikten sonra eğitim sırasında karşılaşmadığı örnekler için de ilgili tepkiyi üretme yeteneğine sahiptir. Eğitilmiş bir ağa, girişin tamamı değil sadece bir kısmı verilse bile ağ, hafızadan bu girişe en yakın olanını seçerek tam bir giriş verisi alıyormuş gibi kabul eder ve buna uygun bir çıkış değeri üretir. YSA' ya, eksik, bozuk veya daha önce hiç karşılaşmadığı şekilde veri girilse bile, ağ kabul edilebilir en uygun çıkışı üretecektir. Bu özellik ağın genelleştirme özelliğidir (Tebelskis, 1995).

3. DENEY SONUÇLARI

Bu çalışmada makine öğrenmesi sınıflandırma algoritmalarından NB, k-EYK, CN2 ve DVM algoritmaları kullanılarak; Dermatology, Echocardio, Breast Cancer, Heart Disease,

Parkinson, Diabetes, Thyroid, Pima ve Bupa biyomedikal veri kümeleri üzerinde sınıflandırma işlemi gerçekleştirilmiştir. Deneyler, Intel i7 2.4 GHz işlemci, 8 GB RAM özelliklerine sahip Windows 7 64 bit işletim sistemi kurulu bilgisayarda yapılmıştır.

Bütün algoritmalarda çapraz doğrulama sayısı 10 olarak ayarlanmıştır. NB algoritmasında olasılığı 0 olan durumlarda *Laplace Doğrulama* yöntemi kullanılmıştır. k-EYK algoritmasında *k* komşu değeri 7 seçilmiştir. DVM algoritmasında kullanılan kernel fonksiyonlarından *Sigmoid* fonksiyon kullanılmıştır. YSA' da gizli (ara) katmandaki nöron sayısı her bir veri kümesi için, veri kümesinin sınıf özelliğinin sahip olduğu değerlerin sayısı kadar seçilmiştir. Gene YSA için maksimum iterasyon sayısı 500 alınmıştır. RO algoritması uygulanırken ormandaki optimum ağaç sayısı her veri kümesi için farklı çıkmıştır. Yapılan testlerde her bir veri kümesi için bu optimum ağaç sayısına kadar algoritmanın başarımı sürekli artmıştır; ancak optimum sayıyı geçtikten sonra ya başarımda değişme görülmemiş ya da başarım oranının daha düşük değerlere indiği görülmüştür. Ayrıca YSA, k-EYK ve DVM algoritmaları için veriler üzerinde normalizasyon işlemi uygulanmıştır. Bupa ve Thyroid veri kümelerindeki tüm özellikler sürekli değerler olduğu için veriler üzerinde ayrıştırma işlemi yapılarak kategorik hale dönüştürülmüştür. Echocardio veri kümesindeki verilerin yaklaşık olarak %53' ü kayıp veri içermektedir. Aynı şekilde Diabetes veri kümesinde de yaklaşık olarak %48 oranında kayıp veri bulunmaktadır. Bu kayıp veriler, veri ön işleme tekniklerinden en yüksek frekans yöntemi ile belirlenerek veri kümesi güncellenmiştir.

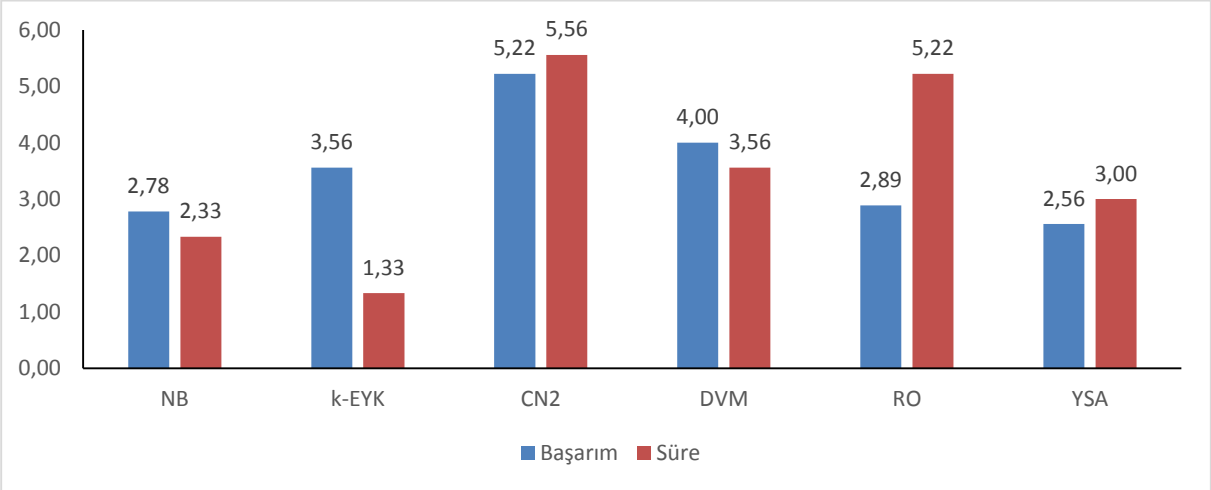
Gerçekleştirilen deneylere sınıflandırma algoritmaların başarımları açısından bakarsak RO, YSA, k-EYK ve NB algoritmalarının CN2 ve DVM algoritmalarına göre daha iyi sonuçlar verdiği görülür. Bu dört algoritmadan özellikle RO ve YSA algoritmalarının başarım değerleri birbirine çok yakındır ve kendilerinden sonra gelen algoritmadan %1.35 daha iyidir. Çalışma süreleri açısından bakarsak k-EYK, YSA ve NB algoritmalarının diğer algoritmalardan özellikle de CN2 ve RO algoritmalarından çok daha hızlı çalıştığı görülür. YSA algoritmasının en iyi başarıma sahip algoritmalardan bir tanesi olması ve en hızlı çalışan ikinci algoritma olması YSA algoritmasını diğer algoritmalardan bir adım öne çıkarmaktadır. k-EYK algoritmasının en iyi üçüncü başarıma sahip olması dezavantaj gibi gözükürken çalışma zamanının çok kısa olması bir avantajıdır. CN2 algoritmasının uzun çalışma zamanına ve düşük başarıma sahip olması, bu algoritmanın deney sonuçlarına göre en kötü algoritma olduğunu göstermektedir. Yapılan deneylere ait sonuçlar Çizelge 2'de gösterilmektedir.

4. DENEY SONUÇLARININ İSTATİSTİKSEL OLARAK KARŞILAŞTIRILMASI

Deney sonuçlarının detaylı bir analizinin yapılabilmesi için kullanılan algoritmaların başarımları ve çalışma süreleri istatistiksel olarak karşılaştırılmıştır. Biyomedikal veri kümeleri için en başarılı sınıflandırma algoritmasını belirlemek için algoritmalar öncelikle deney çalışmalarında kullanılan her veri kümesi için başarımları ve çalışma sürelerine göre sıralanmışlardır. Algoritmaları başarıma göre kıyaslarken herhangi bir biyomedikal veri kümesinde başarımı en yüksek olan algoritmadan en düşük olan algoritmaya kadar sırasıyla 1'den 6'ya kadar olan değerler verilmiştir. Çalışma süresine göre kıyaslamak için ise çalışma süresi en düşük olan algoritmadan en yüksek olan algoritmaya kadar sırasıyla 1'den 6'ya kadar olan değerler verilmiştir. En son ise her bir biyomedikal veri kümesi için verilen değerlerin ortalaması alınarak ortalama başarım ve ortalama çalışma süreleri bulunmuştur. Bulunan değerler şekil 8'de gösterilmektedir.

Çizelge 2. Biyomedikal veri kümeleri için makine öğrenmesi sınıflandırma algoritmalarının başarımları ve çalışma süreleri

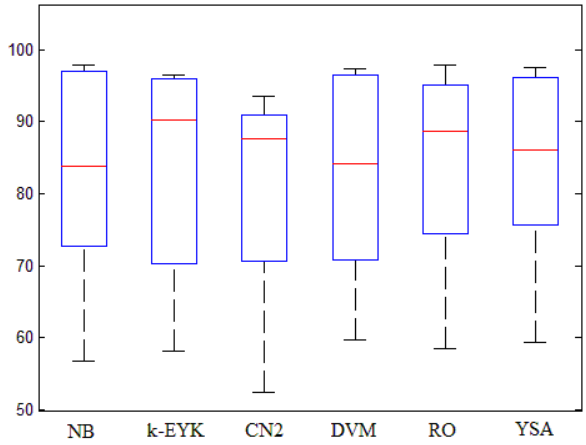
Biyomedikal Veri kümesi	Başarım (%)						Süre (sn)					
	NB	k-EYK	CN2	DVM	RO	YSA	NB	k-EYK	CN2	DVM	RO	YSA
Dermatology	**97.82**	96.43	90.16	97.27	97.81	97.55	**0.11**	0.31	3.80	1.52	7.56	0.92
Echocardio	89.29	**90.22**	87.91	86.48	88.68	88.74	0.09	**0.04**	0.22	0.07	0.08	0.1
Breast Cancer	**97.42**	95.85	93.41	96.71	95.42	97.14	**0.36**	0.45	1.82	0.99	1.78	0.48
Heart Disease	56.74	58.09	52.47	**59.72**	58.45	59.28	0.4	**0.1**	4	1	1.2	0.9
Parkinson	77.92	**91.79**	87.66	84.11	89.79	86.13	2.3	**0.2**	3.2	0.6	4.3	0.5
Diabetes	83.79	79.96	78.04	81.33	**85.71**	83.25	0.32	**0.09**	1.1	0.42	0.8	0.4
Thyroid	**96.75**	96.30	93.51	96.30	94.94	95.84	0.3	**0.19**	0.8	0.36	0.72	0.49
Pima	74.99	73.16	72.00	74.35	76.43	**77.73**	0.6	0.5	9.7	1.76	11.2	**0.4**
Bupa	66.05	61.40	66.68	60.31	68.13	**69.29**	0.1	**0.08**	2.2	0.6	2.3	0.31
Ortalama	82.31	82.58	80.20	81.84	**83.93**	83.88	0.51	**0.22**	2.98	0.81	3.33	0.50



Şekil 8. Deneysel sonuçların başarım ve süre yönünden istatistiksel karşılaştırılması

Elde edilen istatistiksel karşılaştırma değerlerine göre YSA algoritması ortalama en iyi başarım değerine ve k-EYK algoritması da ortalama en iyi çalışma süresine sahip algoritmalardır. NB algoritması ise hem ortalama başarım değerinde hem de ortalama çalışma süresinde ikinci sırada yer almıştır. Bu grafiğe bakarak YSA ve NB algoritmalarının biyomedikal veri kümeleri için hem başarım değerleri hem de çalışma süreleri açısından iyi sonuçlar verdiği ve diğer algoritmalara göre tercih edilebileceği söylenebilir. Deney sonuçlarına göre başarımda birinci sırada yer alan RO algoritması ise burada 2.89 değeri ile üçüncü sırada yer almaktadır. Bu değere bakarak RO algoritmasının yarışmalarda genellikle ikinci ve çoğunlukla üçüncü geldiği yorumu yapılabilir. Böylelikle deneysel açıdan başarılı olan RO algoritmasına istatistiksel açıdan bakıldığında aslında o kadarda başarılı bir algoritma olmadığı söylenebilir. Bunun yanı sıra CN2 algoritması 5.22 başarım değeri ve 5.56 çalışma süresi değere ile diğer bütün algoritmalardan daha kötüdür. Böylelikle CN2 algoritması deneysel çalışmalarda da olduğu gibi sınıflandırma işlemi için en son tercih edilebilecek algoritmadır.

Son olarak algoritmalar istatistiksel olarak kutu grafiği yöntemi ile karşılaştırılmıştır. Kutu grafiği yöntemi ile algoritmaların sınıflandırma başarımları merkezsel konumları gözlenmiştir. Bulunan değerler Şekil 9'da gösterilmektedir.



Şekil 9. Algoritmaların başarım değerlerinin kutu grafiği yöntemi ile karşılaştırılması

Şekil 9'da baktığımızda YSA algoritmasının kutu grafiği boyu diğer algoritmaların kutu grafik boylarına göre daha kısadır, bıyıkların kutuya olan uzaklıkları yakın sayılabilir ve medyan değeri kutunun ortasına yakın bir yerde oluşmuştur. Bunun yanında kutunun alt sınırı diğer algoritmalardan daha yukarıdadır ve üst sınırı ise en yükseğe yakın bir konumdadır. Bu bilgiler doğrultusunda YSA algoritmasının değişik özellikte olan veri kümelerine için diğer algoritmalara göre daha kararlı sonuçlar verdiği görülmüştür. Böylelikle bu çalışmada kullanılmayan diğer küçük ve orta ölçekli biyomedikal veri kümeleri için YSA algoritmasının başarılı sonuçlar vereceği söylenilebilir. CN2 algoritmasının kutu grafiğine baktığımızda özellikle alt bıyığın kutudan çok uzakta olduğu ve medyan değerinin ise ortasına uzak olduğu görülmektedir. Böylelikle CN2 algoritmasının değişik özellikteki veri kümeleri için farklı sonuçlar verdiği ve kararsız bir tutum sergilediği söylenilebilir. Ayrıca kutunun üst ve alt sınırlarının da diğer algoritmalara göre daha düşük seviyede olması bu algoritmanın diğer algoritmalara göre daha başarısız bir algoritma olduğunun göstergesidir.

5. SONUÇ VE ÖNERİLER

Makine öğrenmesi algoritmalarının sağlık sektörü alanındaki uygulamaları her geçen gün biraz daha artmaktadır. Sağlık sektöründe doktorlar tarafından yapılan teşhis ve tedavilerde daha doğru sonuçlar elde etmek, insan kaynaklı hataları engellemek ve doktorun kararına yardımcı olmak amacıyla makine öğrenmesi tabanlı karar destek sistemleri kullanılmaktadır. Bunun yanı sıra çok sayıda hastanın bilgilerini barındıran biyomedikal veri kümelerinin analizi, istatistiki bilgilerin çıkarılması ve bilimsel çalışmalarda kullanılabilmesi için hızlı bir şekilde yorumlanması gerekmektedir. Bu yorumlama işleminde sadece ve sadece bilgisayar ortamındaki makine öğrenmesi tabanlı yazılımlar tarafından gerçekleştirilebilir. Bu çalışmada makine öğrenmesi algoritmalarından k-EYK, NB, CN2, RO, YSA ve DVM kullanılmıştır. Yapılan deneysel ve istatistiksel çalışmalara göre YSA algoritmasının yüksek başarımda sonuçlar verdiği ve küçük ve orta ölçekli veri kümeleri için k-EYK algoritmasının daha hızlı çalıştığı görülmüştür.

Bu çalışma ilerde yapılacak olan çalışmaların için bir başlangıçtır. Sonraki çalışmalarda bu çalışmada kullanılan makine öğrenmesi algoritmalarının iyileştirilmesi veya hibrit kullanımlarının gerçekleştirilmesi düşünülmektedir. Böylelikle de sınıflandırma başarımlarının arttırılabileceği veya çalışma sürelerinin azaltılabileceği düşünülmektedir.

KAYNAKLAR

Akman M., Genç Y., Aankarali H. (2011): "Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama", Türkiye Klinikleri Biyoistatistik Dergisi, Cilt 3, No. 1, s.36–48.

Alpaydın E. (2010) : "Introduction to Machine Learning", The MIT Press Cambridge, Massachusetts London, England.

Altman N. S. (2007): "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", The American Statiscian, Cilt 46, No. 3, s.175–185.

Bermejo P., Gámez J. A., Puerta J. M. (2011) : "Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets", Expert Systems with Applications, Cilt 38, No. 3, s.2072–2080.

Breiman L. (2001) : "Random Forests", Machine Learning, Cilt 45, No. 1, s.5–32.

Clark P., Niblett T. (1989): "The CN2 Induction Algorithm", Machine Learning, Cilt 3, No. 4, s.261–283.

Çölkesen I. (2009): " Uzaktan Algılamada İLeri Sınıflandırma Tekniklerinin Karşılaştırılması ve Analizi", Gebze Yüksek Teknoloji Entitüsü, Jeodezi ve Fotogrametri Mühendisliği, Y.Lisans Tezi.

Cortes C., Vapnik V. (1995): "Support-vector networks", Machine Learning, Cilt 20, No. 3, s.273.

DTREG, "Support Vector Machines", http://www.dtreg.com/svm.htm, Erişim Tarihi: 10.12.2014.

Farid D. M., Li Z., Mofizur R. C., Hossain M.A., Strachan R. (2014): "Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-Class Classification Tasks", Expert Systems with Applications, Cilt 41, No. 4, s.1937–1946.

Fu K., Qu J., Chai Y., Dong Y. (2014): "Classification of Seizure Based on the Time-Frequency Image of EEG Signals Using HHT and SVM", Biomedical Signal Processing and Control, Cilt 13, s.15–22.

Gershenson C., "Artificial Neural Networks for Beginners", http://arxiv.org/ftp/cs /papers/0308/0308031.pdf, Erişim Tarihi: 10.12.2014.

Hacibeyoğlu M., Arslan A., Kahramanlı S. (2011): "Improving Classification Accuracy with Discretization on Datasets Including Continuous Valued Features", World Academy of Science, Engineering and Technology, Cilt 5, No. 6, s.497–500.

He B., Shi Y., Wan Q., Zhao X. (2014): "Prediction of Customer Attrition of Commercial Banks based on SVM Model", Procedia Computer Science, Cilt 31, s.423–430.

Huang J., Lu J., Ling C.X. (2003): "Comparing Naive Bayes, decision Trees, and SVM with AUC and Accuracy", Third IEEE International Conference on Data Mining, s.553–556.

Han J., Kamber M. (2006): "Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers.

Kaya E., O. Fındık, Babaoğlu İ., Arslan A. (2011): "Effect of Discretization Method on the Diagnosis of Parkinson's Disease", International Journal of Innovative Computing, Information and Control , Cilt 7, No. 8, s.4669–4678.

Kecman V. (2001): "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models", The MIT Press, Cambridge, MA.

Küçük H., Tepe C., Eminoğlu I. (2013): "Classification of EMG Signals by k-Nearest Neighbor Algorithm and Support Vector Machine Methods", In 2013 21st Signal Processing and Communications Applications Conference (SIU) IEEE, s.1–4.

Mather P. M. (2005): "Computer Processing of Remotely-Sensed Images".

McCulloch W. S., Pitts W. (1990): "A Logical Calculus of the Ideas Immanent in Bervous Activity", Bullettin of Mathematical Biophysics, Cilt 5, s.115-133.

Nitze I., Barrett B. ve Cawkwell F. (2015): "Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series.", International Journal of Applied Earth Observation and Geoinformation, Cilt.34, s.136–146.

Pal M. (2005): "Random forest classifier for remote sensing classification", International Journal of Remote Sensing, Cilt.26, No.1, s.217–222.

Rajendra U.A., Subbanna P.B., Iyengar S.S. Rao A. ve Dua S. (2003): "Classification of heart rate data using artificial neural network and fuzzy equivalence relation", Pattern Recognition, Cilt.36, No.1, s.61–68.

Tebelskis J. (1995): "Speech Recognition using Neural Networks", Carnegie Mellon University Pittsburgh, Pennsylvania.

UCI, http://archive.ics.uci.edu/ml/, Erişim Tarihi: 10.12.2014