Gazi Üniv. Müh. Mim. Fak. Der.
Cilt 31, No 2, 323-330, 2016
DOI: 10.17341/gummfd.50403

Journal of the Faculty of Engineering and Architecture of Gazi University
Vol 31, No 2, 323-330, 2016

CENTROID SINIFLAYICILAR YARDIMIYLA MEME KANSERİ TEŞHİSİ

Hidayet TAKCI
Cumhuriyet Üniversitesi, Bilgisayar Mühendisliği Bölümü, 58140, Sivas
htakci@cumhuriyet.edu.tr

(Geliş/Received: 06.01.2015; Kabul/Accepted: 01.04.2016)

ÖZET

Meme kanseri kadınlarda en sık görülen kanser türüdür. Artan meme kanseri vakaları nedeniyle meme kanserinde erken teşhis eskisinden daha önemli hale gelmiştir. Erken teşhis için yaşa bağlı birçok yöntem bulunmakla birlikte en sık kullanılan yöntem mamografidir. Bununla birlikte mamografi görüntülerini yorumlamada radyologların yaşadığı görüş ayrılıkları erken teşhis konusunda daha güvenilir sonuçlar veren bilgisayar destekli karar verme mekanizmalarını bir seçenek haline getirmiştir. Bu kapsamda; destek vektör makineleri, yapay sinir ağları ve karar ağaçları gibi makine öğrenmesi yöntemleri bilgisayar destekli karar vermede bugüne kadar kullanılmıştır. Bu çalışmada diğerlerinden farklı olarak centroid tabanlı sınıflayıcılar erken teşhis için ele alınmıştır. Bu tercihin en önemli nedeni centroid sınıflayıcıların karmaşıklığı düşük fakat performansı yüksek sınıflayıcılar olmasıdır. Centroid sınıflayıcılar meme kanseriyle ilgili; Wisconsin, Diagnostic ve Prognostic veri setleri üzerinde test edilmiştir. Centroid sınıflayıcılar ve diğer makine öğrenmesi yöntemleri arasında karşılaştırmalar yapılmış ve sonuçlar doğruluk ve zaman açısından raporlanmıştır. En yüksek sınıflandırma doğruluğunu %99,04 değeriyle Euclidian tabanlı centroid sınıflayıcı vermiştir. Centroid sınıflayıcılar hız açısından da diğer sınıflayıcılara üstünlük sağlamıştır.

Anahtar Kelimeler: Makine öğrenmesi, sınıflandırma, meme kanseri teşhisi, centroid tabanlı sınıflayıcılar, Manhattan uzaklık ölçümü, Euclidian uzaklık ölçümü

DIAGNOSIS OF BREAST CANCER BY THE HELP OF CENTROID BASED CLASSIFIERS

ABSTRACT

Breast cancer is a common cancer type among women. With its increasing incidence early diagnosis has become more important. There are a variety of age-dependent methods for early diagnosis of breast cancer but mammography is the most used method. However, the radiologists show considerable variability in how they interpret a mammogram. Therefore, there is need computer-aided decision-making mechanisms for more reliable results. In this scope various machine learning techniques such as support vector machines, multi layer perceptron and decision trees have been used to early diagnosis in recent years. In this study, centroid-based classifiers are examined for the early diagnosis of breast cancer. The most important reason for this preference is centroid classifiers have low complexity and high performance. Experiments were evaluated on Wisconsin, Diagnostic and Prognostic Dataset. Comparisons between centroid classifiers and the orher classifiers have been done and the results have been presented in terms of accuracy and speed. The highest classification accuracy obtained in the experiments is 99.04%. This classfication rate belongs to the centroid based classifier using the Euclidian measurement. Also, centroid classifiers outperform the other classifiers in terms of classification speed.

Keywords: Machine learning, classification, breast cancer diagnosis, centroid based classifiers, Manhattan distance, Euclidian distance

1. GİRİŞ (INTRODUCTION)

Meme kanseri; kökenini meme dokusundan alan ve erken teşhis edilebildiğinde tedavi şansı yüksek olan bir kanser türüdür [1, 2]. Hastalığın erken teşhisi için hastanın yaşına bağlı çok sayıda yöntem vardır. Örneğin, yirmili yaşlarda hasta kendi kendine meme kontrolü ile hastalığını teşhis edebilirken daha ileri yaşlarda doktor muayenesi zorunludur. Meme dokusunda meydana gelen sertleşme ve şişme durumunda ise klinik meme muayenesi gereklidir. Klinik muayene; ultrasonografi, ince iğne biyopsisi ve mamografi teknikleriyle yapılır [3]. Bir diğer teşhis yöntemi klinik yöntemlerden daha hızlı ve onlardan daha güvenilir olan bilgisayar destekli teşhistir. Bilgisayar destekli teşhiste kişinin bulgularına göre meme kanseri olup olmadığı bir karar destek sistemi yardımıyla tespit edilir. Bugüne kadar sıklıkla; destek vektör makineleri, yapay sinir ağları ve karar ağaçları gibi yöntemler karar destek sistemleri içerisinde kullanılmıştır. Bilgisayar destekli teşhis aslında bir sınıflandırma problemi olup etkili bir sınıflayıcı olan centroid tabanlı sınıflayıcılar da meme kanseri teşhisi için kullanılabilecektir. Bugüne kadar sıklıkla bilgi alma (information retrieval) alanında tercih edilen centroid sınıflayıcılar bu çalışmada bilgisayar destekli teşhis için kullanılmıştır. Centroid sınıflayıcıların özünü adına centroid denen, her biri bir sınıfı sunmak için kullanılan ve genellikle sınıfa ait örneklerin bir ortalamasından elde edilen merkezi bir değer oluşturur. Hem eğitim hem de test aşamasında hızlı ve düşük maliyetli sınıflayıcılardır. Centroid sınıflayıcılar özellikle bilgi alma etki alanında destek vektör makineleri kadar başarılı sonuçlar vermiş ve sınıflandırma hızı açısından da destek vektör makineleri dâhil bütün sınıflayıcıları geçmiştir [4]. Yüksek performansa sahip olan centroid sınıflayıcıların çalışma prensibi oldukça basittir. Centroid sınıflayıcılara göre sınıflandırmada önce sınıfı bilinmeyen bir kayıt her bir sınıfın centroid vektörü ile uzaklık veya benzerlik bilgisine göre karşılaştırılır, bu karşılaştırma sonunda kayıt hangi sınıfa daha yakın veya berzerse o sınıfa atanır. Centoid tabanlı sınıflandırmada kullanılan benzerlik veya uzaklık ölçümleri önemli olup doğrudan sınıflandırma performansına etki eder. Bu çalışmada Manhattan ve Euclidian uzaklık ölçümü ile Cosine benzerlik ölçümü meme kanseri teşhisinde kullanılacak ve performansa etkisi gözlemlenecektir. Çalışmamızda öncelikli olarak meme kanseri veri setleri üzerinde centroid sınıflayıcıların kendi aralarında karşılaştırmaları yapılacak ardından diğer sınıflayıcılar ile sınıflandırma doğruluğu ve hız açısından karşılaştırmalar yapılacaktır. Böylece centroid sınıflayıcıların meme kanseri için uygunluğu gösterilmeye çalışılacaktır.

2. MEVCUT DURUM (CURRENT SITUATION)

Meme kanseri erken teşhisinde bugüne kadar çok sayıda makine öğrenmesi yaklaşımı denenmiş ve başarılı sonuçlar alınmıştır. Abdelghani ve Guven [5] C4.5, Naive Bayes ve Back-Propagated Neural Network algoritmalarını meme kanseri hastalarının hayatta kalma oranlarını tespit etmek amacıyla kullanmışlar ve yaptıkları deneylerde C4.5 algoritmasının Naive Bayes ve Back-Propagated Neural Network algoritmalarından daha başarılı sonuç verdiğini görmüşlerdir. Pena-Reyes ve Sipper [6] tarafından yapılan çalışmada fuzzy logic ve genetik algoritmalar yardımıyla meme kanseri teşhisi %97,36 gibi bir oranla yapılabilmiştir. Setiono tarafından yapılan çalışmada yapay sinir ağları tabanlı bir algoritma kullanılmış ve %98.10'luk bir doğru tanıma oranı elde edilmiştir [7]. Abonyi ve Szeifert, denetimli bulanık küme tekniği yardımıyla meme kanseri teşhisinde %95,57 gibi bir doğru tanıma oranı elde etmiştir [8].

Übeyli tarafından yapılan çalışmada destek vektör makineleri, olasılıksal yapay sinir ağı, tekrarlı yapay sinir ağı ve çok katmanlı algılayıcı kullanılmış ve sırayla; %99,54, %98,61, % 98,15 ve %97,40 gibi doğru tanıma sonuçları elde etmiştir [9]. Akay çalışmasında destek vektör makinesi tabanlı bir metodu özellik seçimi yöntemiyle birlikte kullanmış ve %99,51 gibi bir doğru teşhis oranı elde etmiştir [10]. Bir başka çalışmada Peng, Yang ve Jiang [11] filtre ve sarma (filter and wrapper) yöntemi kullanarak %99,50 gibi bir doğru teşhis oranı elde etmişlerdir. Konuyla ilgili yakın zamanda yapılan çalışmalardan birinde Kaya [12] kanser verilerini melez bir yöntem ile ele almış ve %100 gibi bir doğru tanıma oranı elde etmiştir. Yazar çalışmasında kaba kümeler teorisi (rough set theory) yardımıyla gereksiz nitelikleri veri setinden çıkarmış ve kalan nitelik değerleriyle sınıflandırma işlemini yerine getirmiştir. Delen ve arkadaşları [13] bir çalışmada meme kanseri teşhisi için yapay sinir ağları, karar ağacı ve lojistik regresyon tekniklerini SEER veri seti üzerinde kullanmışlardır. Karar ağacı için elde edilen sonuç %93,6; yapay sinir ağı için %91,2 ve lojistik regresyon için %89,2 olmuştur.

Bu çalışmalar göstermektedir ki meme kanseri teşhisinde makine öğrenmesi yaklaşımları klinik yöntemlere güçlü bir alternatif oluşturmaktadır ve daha iyi sonuçlar için yeni çalışmalara ihtiyaç vardır. Bu kapsamda daha önce meme kanseri erken teşhisinde kullanılmamış olan centroid tabanlı sınıflayıcılar bu çalışmada kullanılacaktır. Centroid sınıflayıcıların bu alanda kullanılan yüksek maliyetli sınıflayıcılara nispetle en önemli avantajı doğrusal zamanlı ve düşük maliyetli bir sınıflayıcı olmasıdır.

3. YÖNTEM (THE METHOD)

Meme kanseri erken teşhisinde daha yüksek teşhis doğrulukları ve daha hızlı teşhis için yeni çalışmalara ihtiyaç vardır. Bu ihtiyaçlar dolayısıyla meme kanseri erken teşhisinde daha önce kullanılmamış performansı yüksek bir makine öğrenmesi yaklaşımı kullanılacaktır.

3.1 Centroid Tabanlı Sınıflayıcılar (Centroid Based Classifiers)

Centroid tabanlı sınıflayıcılar, vektör uzayı modeli tabanlı, hem eğitim hem de test maliyeti düşük, etkili sınıflayıcılardır. Centroid sınıflayıcıların temel prensibi her bir sınıfın centroid adı verilen bir vektör ile sunulmasıdır [14]. Centroid vektörü sınıflayıcının eğitimi aşamasında sınıfa ait eğitim örneklerinden elde edilir. Centroid vektörü elde edilmesinde farklı yöntemler kullanılmakla birlikte en sık tercih edileni ortalama yöntemidir ve bu yöntem çalışmamızda da kullanılmıştır.

Ortalama yöntemine göre hesaplamada; her biri $\vec{d}$ ile sunulabilecek S adet eğitim örneğinden elde edilen centroid değeri $\vec{C}$ Eşitlik 1'deki gibi sunulabilir:

$$\vec{C} = \frac{1}{|S|}\sum_{d \in S} \vec{d} \quad (1)$$

Centroid tabanlı sınıflandırma, sınıfı bilinmeyen kayıtların centroid vektörlerine olan uzaklık veya benzerliğine dayalı olarak yapılır. Bu çalışmada uzaklık ölçümü için Manhattan ve Euclidian, benzerlik ölçümü için ise Cosine benzerlik yöntemi kullanılmıştır. İki nokta arasındaki Manhattan uzaklığı noktalar arasındaki yatay ve dikey uzaklıkların toplamından elde edilir. Euclidian uzaklığı ise iki nokta arasındaki Pisagor uzaklığından elde edilir. Her iki uzaklık metodu da Minkowski adı verilen bir uzaklık ölçümünün farklı formlarıdır. $X = (x_1, x_2, \dots, x_n)$sınıfı bilinmeyen bir veri elemanı ve $C = (c_1, c_2, \dots, c_n)$ bir centroid vektörü iken Minkowski uzaklığı Eşitlik 2'deki gibidir.

$$(\sum_{i=1}^{n} |x_i - c_i|^p)^{1/p} \quad (2)$$

Minkowski uzaklığı Euclidian uzayda bir ölçüm olup Euclidian ve Manhattan uzaklıklarının genelleştirilmiş formu olarak düşünülebilir [15]. Bu çalışmada kullanılacak bir diğer ölçüm yöntemi Cosine benzerlik ölçümü olup bu yöntem vektörler arasındaki açısal benzerliğe göre işlem yapmayı destekler. $X = (x_1, x_2, \dots, x_n)$ ve $C = (c_1, c_2, \dots, c_n)$ R^n iken X ve C vektörleri arasındaki uzaklık ve benzerlik ölçümleri Tablo 1'deki gibi olacaktır.

Meme kanseri teşhisinde centroid tabanlı sınıflayıcılar kullanılırken; önce meme kanseri veri setinin bir kısmı eğitim diğer kısmı test verisi olarak ikiye ayrılır. Eğitim verisi yardımıyla Malignant ve Benign sınıfları için centroid vektörleri elde edilir. Ardından test verileri sırayla üç farklı centroid sınıflayıcıya göre (Manhattan tabanlı, Euclidian tabanlı ve Cosine tabanlı) işleme tabi tutulur. Böylece test verilerinin bir hastaya mı yoksa sağlıklı bir kişiye mi ait olduğu farklı ölçüm yöntemleriyle tespit edilmiş olur. Farklı uzaklık ve benzerlik yöntemlerine göre sınıflandırma karar kuralları Tablo 2'deki gibi olacaktır.

Tablo 1. Uzaklık/benzerlik ölçümleri (Distance / similarity metrics)

Ölçüm yöntemi	Denklemi
Manhattan uzaklık	$d_man(X,C) = \sum_{i=1}^{n} \lvert x_i - c_i \rvert$
Euclidian uzaklık	$d_euc(X,C) = \sqrt[2]{\sum_{i=1}^{n} (x_i - c_i)^2}$
Cosine benzerlik	$d_{sim(X,C)} = \frac{\sum_{i=1}^{n}(x_i * c_i)}{\sqrt{\sum_{i=1}^{n}(x_i * x_i)} * \sqrt{\sum_{i=1}^{n}(c_i * c_i)}}$

Tablo 2. Uzaklık/benzerlik ölçümleri için sınıflandırma karar kuralı (Classification decision rule for distance/similarity measures)

Ölçüm yöntemi	Karar kuralı
Manhattan uzaklık ölçümü	$\arg\min_{j=1\dots k}(d_man(X, C_j))$
Euclidian uzaklık ölçümü	$\arg\min_{j=1\dots k}(d_euc(X, C_j))$
Cosine benzerlik ölçümü	$\arg\max_{j=1\dots k}(d_sim(X, C_j))$

Tablo 2'de verilen denklemlerde kullanılan **j** sınıf numarasını, $\mathbf{C_i}$ o sınıfın centroid vektörünü, **X** değeri sınıfı bilinmeyen kayıt verisini ve **y** tahmin edilen sınıf numarasını vermektedir. Uzaklık tabanlı centroid sınıflayıcılar için sınıflandırma kuralı minimum uzaklık, benzerlik tabanlı centroid sınıflayıcılar için ise sınıflandırma kuralı maksimum benzerliktir. Böylece sınıfı bilinmeyen kayıtlar uzaklık veya benzerlik yöntemine uygun olarak kanser veya değil şeklinde sınıflanacaklardır.

3.2 Diğer Sınıflayıcılar (The other Classifiers)

Bu çalışmada kullanacağımız centroid sınıflayıcıların başarısını değerlendirebilmek için meme kanseri erken teşhisinde daha önce kullanılmış olan makine öğrenmesi yöntemleri ile karşılaştırma yapmaya ihtiyacımız vardır. Bu nedenle çalışmamızda centroid sınıflayıcılar ile C4.5, k en yakın komşu (k-NN), destek vektör makinesi (SVM) ve çok katmanlı algılayıcı (MLP) karşılaştırılacaktır. Tercih ettiğimiz yöntemlerden C4.5 algoritması, Quinlan'ın ID3 algoritmasından kaynağını alan bir çeşit karar ağacı tabanlı sınıflayıcıdır. Sınıflayıcı, bilgi kazancı kavramını kullanarak etiketli eğitim verisinin bir

kümesinden karar ağaçları inşa eder. C4.5 sıklıkla tıbbi veri analizinde kullanılan bir algoritmadır. k-NN örüntü tanımada sıklıkla kullanılan, çoğunluk oylamasına dayalı bir sınıflayıcıdır. Komşuların sınıfına göre sınıflandırma yapan k-NN komşuları bulmada benzerlik bilgisi kullanmaktadır [16]. Meme kanseri teşhisinde bir verinin kanser hastasına ait olduğunu söyleyebilmek için en yakınındaki k adet komşusundan çoğunun kanserli örnekler olması gerekmektedir.

SVM, sadece meme kanseri erken teşhisinde değil diğer birçok sınıflandırma probleminde iyi sonuç vermiştir. İlk olarak Vapnik ve arkadaşları tarafından tanıtılan algoritma [17] bugüne kadar birçok uygulamada kullanılmıştır. SVM algoritmasının temel kavramı doğrusal ayrılabilirliktir ve o giriş uzayını bir kernel uzayına eşleştirir, daha sonra bu kernel uzayı üzerinde bir doğrusal model oluşturur. SVM tabanlı sınıflandırma modeli ise sınıfları birbirinden mümkün olan en uzak noktalara ayırmaya çalışan modeli işaret eder.

Bu çalışmada tercih edilen bir diğer algoritma MLP algoritmasıdır. MLP algoritması yapay sinir ağı tabanlıdır ve bir yapay sinir ağı biyolojik sinir ağına dayalı bir matematiksel veya hesaplamalı modeldir. Diğer bir ifadeyle, YSA biyolojik sinir sisteminin bir benzetimidir [18]. YSA'nın kullanım alanlarından birisi de sınıflandırmadır ve bu alanda sıklıkla çok katmanlı algılayıcı tercih edilir. Çalışmamızda da sınıflandırma için MLP tercih edilmiştir. Bir MLP genellikle bir giriş katmanı, bir çıkış katmanı ve en az bir adet gizli katman kullanır.

4. DENEYLER (EXPERIMENTS)

Bu bölümde önce kullanılan veri setleri ve performans ölçütleri sunulacak ardından yöntemlerin doğruluk ve hız açısından karşılaştırmaları raporlanacaktır. Deneylerin başında centroid tabanlı sınıflayıcılar kendi arasında karşılaştırılacak ardından centroid tabanlı sınıflayıcılar ile diğer sınıflayıcılar karşılaştırılacaktır. Sınıflayıcılar hem sınıflandırma doğruluğu hem de ROC analizi yardımıyla karşılaştırılacaktır. Bölümün sonunda ise hız açısından değerlendirmeler yer alacaktır.

4.1 Veri Seti ve Performans Ölçümleri (Data Set and Performance Measures)

Deneysel çalışmalarda orijinal Wisconsin Meme Kanseri Veri seti (WBCD), Wisconsin Diagnostic Meme Kanseri Veri seti (WDBC) ve Wisconsin Prognostic Meme Kanseri Veri seti (WPBC) kullanılmıştır. Orijinal veri seti Wisconsin Üniversitesi hastanelerinden Dr. William H. Wolberg tarafından elde edilmiş olup veri setinde 16'sı kayıp değerler içeren 683'ü tam toplam 699 adet örnek içermektedir. Her bir kayıt 11 farklı nitelik değeri ile sunulmaktadır. Bu niteliklerden ikisi hasta numarası ve sınıf etiketidir. Meme kanseri verisinde sınıf etiketi için iki değer bulunur: iyi huylu (kanserli olmayan) için 2 değeri ve habis (kanserli) için 4 değeri. Geriye kalan nitelikler meme kütlesinin ince iğne aspirasyonu yardımıyla elde edilmiş resminden hesap edilen değerlerdir. Bu değerler; yığın kalınlığı, hücre boyutu homojenliği, hücre şekli homojenliği, marjinal yapışma, epitel hücre boyutu, çıplak çekirdekler, yumuşak kromatin, normal çekirdekçik ve mitoz ismi verilen niteliklerle 1-10 arası değerlerle hücre çekirdeği karakteristiklerini açıklar [20, 21]. Diagnostic veri seti ve prognostic veri seti de orijinal veri seti gibi meme kitlesinin ince iğne aspirasyonundan elde edilen resmin sayısallaştırılmasına dayalı bilgiler içerir. Bununla birlikte Diagnostic veri seti toplam 32 adet özellikten meydana gelir. Bu özellikler; ID, Diagnosis (M=malignant, B=benign) ve 30 adet gerçek değerli özelliktir. Veri setinde toplam 357 benign, 212 adet malignant kayıt verisi bulunur. Kayıp değer içeren kayıt bulunmaz. Prognostic veri seti invaziv meme kanseri verileri içerir. Veride yer alan 30 özellik meme kitlesinin ince iğne aspirasyonundan elde edilen resmin sayısallaştırılmasından elde edilmiştir. Bu özellikler görüntüde bulunan hücre çekirdeğinin karakteristiklerini açıklar. Toplam 198 adet kayıt bulunur. Her kayıt 34 özellik ile (ID, sonuç(outcome), 32 gerçek değerli giriş özelliği) sunulur. 151 nonrecur ve 47 recur kayıt vardır. 4 adet kayıp veri vardır.

Gerek centroid tabanlı sınıflayıcılarla gerekse diğer sınıflayıcılarla çalışırken verinin %90'ı eğitim %10'u test olarak kullanılmış ve bütün deneylerde 10-katlı çapraz doğrulama çalıştırılmıştır. Sınıflayıcıları karşılaştırırken Tablo 3'de detayları verilen doğruluk ölçümü kullanılmıştır.

Tablo 3. Katıştırma matrisi ve doğruluk değeri (Confusion matrix and accuracy)

a: TP (gerçek pozitif)
b: FN (hatalı negatif)
c: FP (hatalı pozitif)
d: TN (gerçek negatif)

	Tahmin edilen	
Gerçek	İyi huylu	Kötü huylu
İyi huylu	A	B
Kötü huylu	C	D

$$accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FN+FP+TN}$$

4.2 Centroid Sınıflayıcıların Karşılaştırılması (Comparison of Centroid Classifiers)

Bu bölümde centroid tabanlı sınıflayıcılar kendi aralarında doğruluk oranına göre karşılaştırılmıştır. Deney amacına uygun olarak Manhattan uzaklık yöntemi, Euclidian uzaklık yöntemi ve Cosine benzerlik yöntemi üç farklı veri seti üzerinde kullanılmış ve elde edilen doğru tanıma oranları Tablo 4'de sunulmuştur.

Tablo 4. Centroid sınıflayıcılar için doğru tanıma oranları (Accuracy rates for centroid classifiers)

Sınıflayıcı	Doğruluk oranı		
	WBCD	WDBC	WPBC
Manhattan tabanlı centroid sınıflayıcı	%98,56	%93,57	%59,32
Euclidian tabanlı centroid sınıflayıcı	%99,04	%92,40	%59,32
Cosine tabanlı centroid sınıflayıcı	%91,35	%87,72	%76,27

Yapılan deneyler sonunda orijinal veri setinde en iyi sonucu Euclidian tabanlı sınıflayıcı, Diagnostic veri setinde Manhattan tabanlı centroid sınıflayıcı, prognostic veri setinde ise Cosine tabanlı sınıflayıcı vermiştir. Bu sonucun elde edilmesinde doğrusal ayrılabilirlik önemli bir faktör olup orijinal veri seti ve Diagnostic veri seti daha doğrusal ayrılabilir değer içerdiklerinden uzaklık tabanlı yöntemler daha iyi sonuç vermiştir. Bu deneyler bize veri setine bağlı olarak farklı centroid sınıflayıcıların daha iyi sonuç verebileceğini göstermiştir. Bu arada meme kanseri veri setinde centroid sınıflayıcılar ile elde edilebilen en yüksek sınıflandırma doğruluğu %99,04 olmuştur.

4.3 Centroid Tabanlı Sınıflayıcılar ile Diğer Sınıflayıcıların Karşılaştırılması (Comparison of Centroid Based Classifiers and Other Classifiers)

Centroid tabanlı sınıflayıcıların ardından aynı veri setleri üzerinde; C4.5, k-NN, SVM ve MLP algoritmaları çalıştırılmış ve elde edilen sonuçlar karşılaştırılmıştır. Bu kısımda elde edilen sınıflandırma sonuçları optimize edilmiş parametrelerle çalıştırılan makine öğrenmesi yöntemlerine aittir. Optimize edilmiş parametre değerleri eğitim veri setinin %10'luk validasyon amaçlı bölümü kullanılarak elde edilmiştir. Bu değerlere çok sayıda deney sonrasında ulaşılmıştır. Parametre optimizasyonu çalışmaları ve karar verilen parametre değerleri şu şekilde olmuştur. C4.5 algoritması için iki önemli parametre minimum yaprak sayısı (minimum leaf size) ve güvenilirlik seviyesidir (confidence level). Minimum yaprak sayısı için sırayla 3, 5 ve 7 değerleri kullanılmış ve her bir değer 0,05, 0,10, 0,15. 0,20, 0,25 ve 0,50 güvenilirlik seviyeleri için test edilmiştir. Bu eşleştirmeler sonucunda en iyi sonuç minimum yaprak sayısı 5 ve güvenilirlik seviyesi 0,15 iken elde edilmiştir. k-NN algoritması için en önemli parametre değeri komşu sayısı yani k değeri olup k için 1 ile10 arasındaki değerler sırayla test edilmiştir. Bu işlemler sırasında uzaklık ölçümü olarak Euclidian ölçümü kullanılmıştır. Komşu sayısı parametresi için en optimum değer 5 olarak bulunmuştur. Bir diğer sınıflayıcı SVM olup SVM için en önemli parametre kernel tipidir. Kernel tipi olarak karşımızdaki seçenekler Linear, Polynomial, Radial Basis ve Sigmoid fonksiyonudur. Yapılan denemeler sonrasında en iyi sonucu veren kernel tipi Linear kernel olmuştur. Bunun nedeni verinin doğrusal karakterde olmasıdır. Ayrıca kernel derecesi 1 olarak tespit edilmiştir. Parametre optimizasyonunda son çalışılan algoritma MLP olmuştur. MLP için parametreler üç temel kategori şeklinde ele alınabilir. Bunlardan sinir ağının yapısı (network) kısmında karşımıza gizli katmandaki nöron sayısı çıkmaktadır. Yaptığımız denemelerde 5 ile 15 arasında bütün değerlerle çalışma yapılmış veri setimiz için en uygun nöron sayısı olarak 12 bulunmuştur. Bir diğer parametre kategorisi öğrenim (learning) olup burada iki değer öğrenim oranı ve öğrenim yapılan veri setinin eğitim veri setindeki oranıdır. Öğrenim oranı için sırayla; 0,01, 0,05, 0,10, 0,15, 0,25 ve 0,50 değerleri test edilmiş ve en iyi sonuç 0,05 değeriyle elde edilmiştir. Bu kategorideki bir diğer parametre kullanılan validasyon kümesinin eğitim seti üzerindeki oranıdır. Yapılan testlerde 0.05, 0,10, 0,15 ve 0,20 değerleri test edilmiş ve en iyi sonuç 0,10 için elde edilmiştir. MLP ile ilgili son parametre kategorisi durma kuralı (stopping rule) olup bu kategoride maksimum iterasyon sayısı ve hata oranı eşiğidir. Bu değerler de sırayla 100 ve 0,01 olarak tespit edilmiştir. Dört sınıflayıcı için de optimize edilmiş parametre değerleri elde edildikten sonra ilgili değerlerle sınıflandırma deneyleri yapılmıştır. Deneyler Tanagra [19] adı verilen bir makine öğrenmesi aracı yardımıyla yerine getirilmiş ve deneysel sonuçlar elde edilmiştir, sonuçlar Tablo 5'de yer almaktadır.

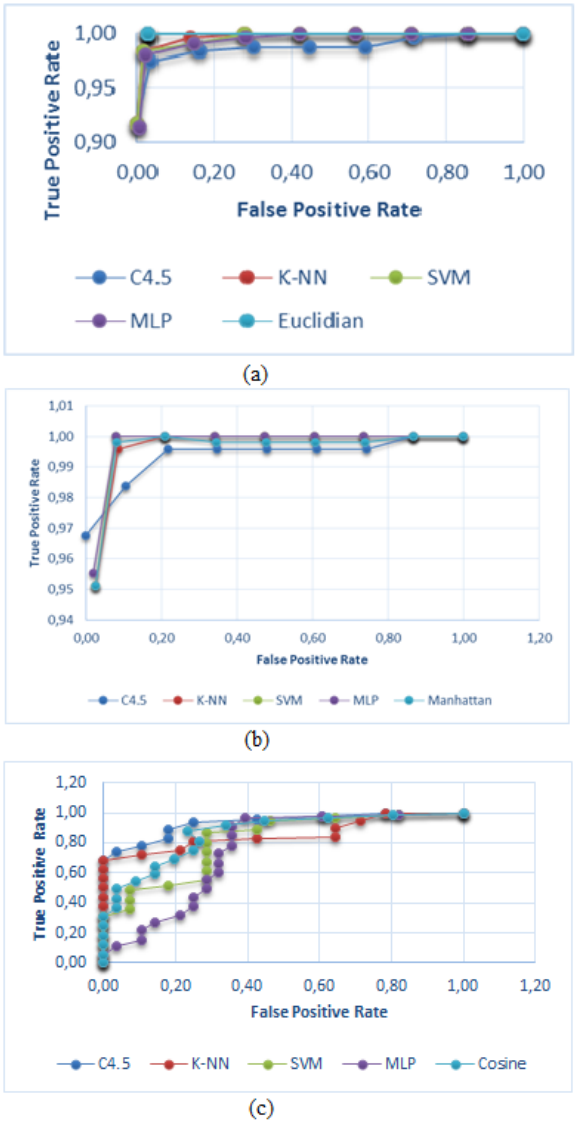
Tablo 5. Centroid sınıflayıcılar ve diğer sınıflayıcılar için doğru tanıma oranları (Accuracy rates for centroid classifiers and the other classifiers)

Sınıflayıcı	Doğruluk oranı		
	WBCD	WDBC	WPBC
C4.5	%95,74	%92,32	%64,74
k-NN	%97,35	%96,25	%71,05
SVM	%96,91	%97,50	%75,26
MLP	%97,50	%96,96	%68,95
Manhattan based centroid classifier	%98,56	%93,57	%59,32
Euclidian based centroid classifier	%99,04	%92,40	%59,32
Cosine based centroid classifier	%91,35	%87,72	%76,27

Centroid tabanlı sınıflayıcılar ile diğer sınıflayıcılar karşılaştırıldığında orijinal veri setinde Euclidian tabanlı tabanlı centroid sınıflayıcının diğer sınıflayıcılardan daha üstün sonuç verdiği görülmüştür. Bununla birlikte bütün sınıflayıcılar arasında en düşük sonucu ise yine bir centroid tabanlı sınıflayıcı olan Cosine tabanlı centroid sınıflayıcı vermiştir. Diagnostic veri setinde ise SVM algoritması centroid sınıflayıcılar dâhil bütün sınıflayıcılara karşı üstünlük sağlamıştır. Bu veri setinde hemen hemen en kötü sonuçlar centroid sınıflayıcılara aittir. Son veri seti olan Prognostic veri setinde en iyi sonucu şaşırtıcı şekilde Cosine tabanlı centroid sınıflayıcı vermiştir. Diğer iki veri setinde en kötü sonucu veren Cosine tabanlı centroid sınıflayıcıyı Prognostic veri setinde SVM algoritması takip etmiştir. Bütün veri setleri için sonuçlar dikkate alındığında Centroid sınıflayıcılar ve SVM algoritmasını sırayla; k-NN, MLP ve C4.5 algoritmaları takip etmiştir.

4.4 ROC Analizi Yardımıyla Sınıflayıcıların Karşılaştırılması (Comparison of the Classifiers with ROC Analysis)

Bir ROC eğrisi yatay eksende hatalı pozitif oranı, dikey eksende gerçek pozitif oranı olan, eğitim verisinin farklı boyutları için elde edilen ve sınıflayıcı başarısını gösteren bir eğridir. Eğri ile yatay eksen arasında kalan alan ne kadar büyükse sınıflayıcının o kadar başarılı olduğu kabul edilir. ROC analizi her üç veri seti üzerinde de yapılmış olup her bir verisetinde bir centroid sınıflayıcı ile birlikte MLP, SVM, K-NN ve C4.5 sınıflayıcıları kullanılmıştır.



Şekil 1. ROC analizi sonuçları (a) Wisconsin veriseti üzerinde Euclidian tabanlı centroid sınıflayıcı ve diğerleri (b) Diagnostic veriseti üzerinde Manhattan tabanlı centroid sınıflayıcı ve diğerleri (c) Prognostic veriseti üzerinde Cosine tabanlı centroid sınıflayıcı ve diğerleri (ROC analysis results (a) Euclidian based centroid classifier and the others on Wisconsin dataset (b) Manhattan based centroid classifier and the others on Diagnostic dataset (c) Cosine based centroid classifier and the others on Prognostic dataset)

Şekil 1'de görüleceği üzere analiz sonucuna göre orijinal veri setinde en iyi sonucu Euclidian tabanlı centroid sınıflayıcı vermiştir. İkinci sırada k-NN algoritması gelmiştir. Euclidian tabanlı centroid sınıflayıcı ile k-NN sınıflayıcı ikisi de geometrik uzaklığa dayalı sınıflayıcılardır, her iki sınıflayıcının da iyi sonuçlar vermesi meme kanseri orijinal veri setinde uzaklığa dayalı yöntemlerin başarılı olduğu anlamına gelmektedir. Orijinal veri setinden farklı olarak Diagnostic veri seti ve Prognostic veri seti için sonuçlar farklıdır. Örneğin, Diagnostic veri seti için en iyi sonucu SVM ve MLP vermiştir. Orijinal veri setinde en iyi sonucu centroid tabanlı sınıflayıcı vermesine rağmen burada hemen hemen en kötü sonuçlar ona aittir. Belki daha önemli fark Prognostic veri setinde ortaya çıkmıştır. Orijinal veri setinde en düşük sınıflandırma sonucunu veren Cosine tabanlı centroid sınıflayıcı Prognostic veri setinde en iyi sonucu veren yöntem olmuştur. Bu durum doğru tanıma sonuçlarında olduğu kadar ROC analizi sonuçlarında da görülmektedir.

4.5 Performans Açısından Yöntemlerin Karşılaştırılması (Comparisons of the Methods in Terms of Performance)

Sınıflayıcıları performans açısından ele almanın bir yöntemi de işlem maliyetlerinin karşılaştırılmasıdır. İşlem maliyeti açısından karşılaştırmalarda önce büyük O notasyonu sonra da işlem süreleri kullanılmıştır.

C4.5 algoritması ağaç veri yapısına dayalı bir algoritmadır ve n eleman sayısı olmak üzere C4.5 algoritması için zaman açısından karmaşıklık değeri O(nlogn)'dir [22]. K-NN algoritması eğitim açısından karmaşık olmamakla birlikte test açısından O(nm) karmaşıklığına sahiptir [23]. Burada n değeri veri setindeki eleman sayısı, m ise her bir elemanın nitelik sayısını vermektedir. Centroid sınıflayıcı gibi doğrusal bir sınıflayıcı olmasına rağmen daha fazla karmaşıklığa sahiptir aradaki fark sınıf adedi ile eleman sayısı arasındaki farktan kaynaklanmaktadır. SVM algoritması da karmaşıklığı yüksek sınıflayıcılardan biridir. Optimize edilmediğinde $O(n^3)$ civarında olan algoritma karmaşıklığı en iyileme sonrası $O(n^2)$ değerine düşmektedir [24]. MLP de SVM gibi karmaşıklığı yüksek bir sınıflayıcıdır. MLP algoritması için karmaşıklığa etki eden faktörler giriş düğüm sayısı, gizli katmandaki nöron sayısı ve iterasyon sayısıdır.

Asimptotik olarak karmaşıklık ölçümü bazı nondeterministik algoritmalar için sorunlu olduğundan sınıflayıcıların aynı makine üzerinde işlem süreleri açısından karşılaştırılması daha uygun olacaktır. C4.5, k-NN, SVM ve MLP algoritmaları ile birlikte Centroid sınıflayıcılar işlem süreleri açısından test edildiğinde aşağıdaki işlem sürelerine ulaşılmıştır.

Tablo 6. Sınıflayıcıların sınıflandırma süresi açısından karşılaştırması (Comparison of the classifiers in terms of classification time)

Sınıflayıcı	İşlem süresi (ms)		
	WBCD	WDBC	WPBC
C4.5	1952	1906	1808
k-NN	6911	7731	6481
SVM	7129	7994	8144
MLP	1320	1195	1264
Manhattan Based Centroid Classifier	25	28	22
Euclidian Based Centroid Classifier	116	130	99
Cosine Based Centroid Classifier	224	251	188

Tablo 6'da bulunan değerler Intel Core i3 işlemcili, 4 GB hafizaya sahip bir PC ortamında elde edilmiştir. Değerlerden görüleceği üzere centroid sınıflayıcılar düşük maliyetli ve hızlı sınıflayıcılar olup bu durum önerimizin için en büyük avantajıdır.

4.6 İlişkili Çalışmalar ve Bizim Çalışmamız (Related Works and our Study)

Meme kanseri erken teşhisinde bugüne kadar birçok yöntem kullanılmış olup çalışmamızda onlara ek olarak centroid sınıflayıcılar kullanılmıştır. Orijinal Wisconsin veri setini referans alarak yapılan karşılaştırma Tablo 7'de yer almaktadır.

Tablo 7. Çalışmamız ve diğer çalışmaların karşılaştırması (Comparison between our study and the other studies)

Yazar(lar)	Kullanılan yöntem(ler)	Doğru teşhis oranı
Pena-Reyes ve Sipper [6]	Fuzzy logic ve genetik algoritmalar	%97,36
Setiono [7]	Yapay sinir ağları tabanlı bir algoritma	%98,10
Abonyi ve Szeifert [8]	Denetimli bulanık küme tekniği	%95,57
Übeyli [9]	Destek vektör makineleri Olasılıksal yapay sinir ağı Tekrarlı yapay sinir ağı Çok katmanlı algılayıcı	%99,54 %98,61 %98,15 %97,40
Akay [10]	Destek vektör makinesi	%99,51
Peng, Yang ve Jiang [11]	Filter and wrapper	%99,50
Kaya [12]	Rough set theory kullanan melez yöntem	%100,00
Delen ve arkadaşları [13]	Yapay sinir ağları Karar ağaçları Lojistik regresyon teknikleri	%93,60 %91,20 %89,20
Yöntemimiz	Euclidian tabanlı centroid sınıflayıcı Manhattan tabanlı centroid sınıflayıcı Cosine tabanlı centroid sınıflayıcı	%99,04 %98,56 %91,35

Değerlerden görüleceği üzere kullandığımız centroid tabanlı yöntem diğer yöntemler arasında yer bulabilecek doğruluk değerlerine sahiptir.

4. SONUÇLAR (CONCLUSIONS)

Meme kanseri erken teşhisi hayat kurtarıcı bir role sahip olup erken teşhis için bugüne kadar başta mamografi olmak üzere çeşitli yöntemler kullanılmıştır. Hala kullanılmakta olan mamografi yönteminde belki en önemli sorun radyologların mamogram görüntülerini yorumlarken ortaya koydukları yorum farklarıdır. Bu durum bilgisayar desteli karar vermeyi gündeme getirmiştir. Bilgisayar destekli karar verme daha güvenilir ve daha hızlı olması yönüyle tercih edilen bir yöntem olmuştur. Bugüne kadar bilgisayar destekli karar vermede; destek vektör makineleri, yapay sinir ağları ve karar ağaçları yoğun olarak kullanılmıştır. Bu çalışmada ise daha önce meme kanseri erken teşhisinde kullanılmamış olan centroid sınıflayıcılar tercih edilmiştir. Centroid tabanlı sınıflayıcılar performansı yüksek sınıflayıcılar olup bu çalışmada farklı uzaklık ve benzerlik yöntemlerine göre üç farklı centroid sınıflayıcı ortaya konmuştur. Bu sınıflayıcılar meme kanseri veri setleri üzerinde test edilmiş ve böylece hem en iyi centroid sınıflayıcı bulunmaya çalışılmış hem de meme kanseri veri setlerinde ne kadar doğru tanıma yapılabildiği ortaya konmuştur. Üç farklı centroid sınıflayıcı arasından Euclidian tabanlı centroid sınıflayıcı %99,04 değeriyle orijinal Wisconsin veri setinde diğer sınıflayıcıları geçerek en iyi sonucu vermiştir. Aynı şekilde Cosine tabanlı centroid sınıflayıcı %76,27 tanıma oranıyla Prognostic veri setinde en iyi sınıflandırma başarısını vermiştir. Bununla birlikte Diagnostic veri setinde centroid tabanlı sınıflayıcılar diğer sınıflayıcıların gerisinde kalmıştır. Deneylerde centroid sınıflayıcılar; C4.5, SVM, k-NN ve MLP gibi yöntemlerle karşılaştırılmıştır. Sonuçlar hem doğruluk ölçümüyle hem de ROC analiziyle elde edilmiştir. Ayrıca sınıflayıcılar işlem hızı açısından da karşılaştırılmış ve işlem hızı açısından karşılaştırmada centroid tabanlı sınıflayıcıların diğerlerinden belirgin derecede hızlı olduğu görülmüştür. Farklı veri setleri üzerinde yapılan deneyler bize bazı gözlem bilgileri sunmuştur. En başta her veri setinde bir centroid türü değil veri setine göre farklı centroid türleri başarılı olabilmektedir. İkincisi centroid sınıflayıcıların sınıflandırma prensibi ile verinin doğrusal ayrılabilir olup olmaması gibi.

Bu çalışmada elde edilen bulgular doğrultusunda centroid tabanlı sınıflayıcılar diğer makine öğrenmesi yöntemleri kadar iyi sonuç vermekte hız açısından da onlardan iyi sonuç vermektedir. Bu durum da bize göstermiştir ki düşük işlem maliyeti ve yüksek tanıma doğruluklarına sahip centroid Sınıflayıcılar diğer sınıflayıcılar gibi meme kanseri teşhisinde kullanılabilir sınıflayıcılardır.

Devam eden çalışmalarımızda başta parametre optimizasyonu olmak üzere, veriler üzerinde ölçekleme ve özellik seçimi gibi çalışmalar yapılarak önerimizin literatürde daha sağlam şekilde yer alması sağlanmaya çalışılacaktır.

KAYNAKLAR (REFERENCES)

1. Sariego, J., "Breast cancer in the young patient", **The American Surgeon**, Cilt 76, No 12, 1397-1401, 2010.
2. Florescu, A., Amir, E., Bouganim, N., Clemons, M., "Immune therapy for breast cancer in 2010-hype or hope?", **Curr Oncol**, Cilt 18, No 1, e9-e18, 2011.
3. İnternet: Breast cancer, http://en.wikipedia.org/wiki/Breast_cancer, 2014.
4. Takci, H. ve Güngör, T., "A High Performance Centroid-based Classification Approach for Language Identification", **Pattern Recogn Lett**, Cilt 33, No 16, 2077-2084, 2012.
5. Bellaachia, A. ve Erhan G., "Predicting Breast Cancer Survivability using Data Mining Techniques", **Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining**, April 22, 2006.
6. Pena-Reyes, C. A. ve Sipper, M. A. "Fuzzy-genetic approach to breast cancer diagnosis", **Artif Intell Med**, Cilt 17, No 2, 131-155, 1999.
7. Setiono, R., "Generating concise and accurate classification rules for breast cancer diagnosis", **Artif Intell Med**, Cilt 18, No 3, 205-217, 2010.
8. Abonyi, J. ve Szeifert, F., "Supervised fuzzy clustering for the identification of fuzzy classifiers", **Pattern Recogn Lett**, Cilt 14, No 24, 2195-2207, 2003.
9. Übeyli, E. D., "Implementing automated diagnostic systems for breast cancer detection", **Expert Syst Appl**, Cilt 33, No 4, 1054-1062, 2007.
10. Akay, M. F., "Support vector machines combined with feature selection for breast cancer diagnosis", **Expert Syst Appl**, Cilt 36, No 2, 3240-3247, 2009.
11. Peng, L., Yang, B., ve Jiang, J., "A novel feature selection approach for biomedical data classification", **J. Biomed Inform,** Cilt 179, No 1, 809-819, 2009.
12. Kaya, Y., "A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: RS + ELM", **Turk J. Elec. Eng. & Comp. Sci**, Cilt 21, 2079-2091, 2013.
13. Delen, D., Walker G. ve Kadam A., "Predicting breast cancer survivability: a comparison of three data mining methods", **Artif Intell Med**, Cilt 34, 113-127, June 2005.
14. Han, E. H. ve Karypis, G., "Centroid-based Document Classification: Analysis and Experimental Results", **Lect Notes Artif Int**, Cilt 1910 of LNCS Series, 424-431, 2000.
15. İnternet: Manhattan and Euclidian distance, http://en.wikipedia.org/wiki/Minkowski_distance, 2014.
16. Christobel, A. ve Sivaprakasam, Y., "An Empirical Comparison of Data Mining Classification Methods", **International Journal of Computer Information Systems**, Cilt 3, No 2, 2011.
17. Vapnik, V.N., **The Nature of Statistical Learning Theory**, Springer-Verlag, New York, 1995.
18. Shantakumar B. P. ve Kumaraswamy Y.S., "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", **European Journal of Scientific Research**, Vol 31, No 4, 2009.
19. Rakotomalala R., "TANAGRA: a free software for research and academic purposes", **Proceedings of EGC'2005, RNTI-E-3**, Cilt 2, 697-702, 2005.
20. Newton, C., **Machine Learning Techniques for Medical Analysis**, Master Tezi, University of Queensland, 2001.
21. William, H., Wolberg, M.D., W. Nick Street, Dennis, M. Heisey, Olvi, L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine needle aspirates", **Western Surgical Association meeting in Palm Desert**, California, November 14, 1994.
22. İnternet: Machine Learning in Real World: C4.5, http://www.sts.tu-harburg.de/teaching/ss-09/ml-sose-09/03-Decision-Tree-c45.pdf, 2015.
23. İnternet: A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm, https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/, 2010.
24. İnternet: What is the computational complexity of an SVM?, https://www.quora.com/What-is-the-computational-complexity-of-an-SVM, 2010.