mühendislikdergisi
Dicle Üniversitesi Mühendislik Fakültesi
Cilt: 8, Sayı: 4, 767-779
Eylül 2017

Yeni bir veri önişleme metodu: k-Harmonik kümeleme tabanlı öznitelik ağırlıklandırma

Musa Peker *

Muğla Sıtkı Koçman Üniversitesi, Bilişim Sistemleri Mühendisliği Bölümü, Muğla

Makale Gönderme Tarihi: 13.03.2017 Makale Kabul Tarihi: 09.05.2017

Öz

Veri madenciliği, büyük ölçekli veriler arasından önceden bilinmeyen, nitelikli bilgilerin keşfedilmesi sürecidir. Bu süreç farklı adımlardan oluşmaktadır. Bu sürecin ilk adımı verilerin toplanması ve ön işleme aşamasıdır. Sınıflandırma algoritmalarının performansını arttırmak için giriş verilerine, ön işleme yöntemleri uygulanmaktadır. Veri ön işleme yöntemleri içerisinde yer alan öznitelik ağırlıklandırma yöntemleri, veri madenciliği alanında büyük önem taşımaktadır. Bu çalışmada yeni bir öznitelik ağırlıklandırma yöntemi geliştirilmiştir. Önerilen ağırlıklandırma yöntemi k-harmonik ortalamalar algoritması kullanılarak geliştirilmiştir. Bu ağırlıklandırma yöntemi ile doğrusal olarak ayrılamayan veri setini, doğrusal ayrılabilir veri setine dönüştürmek hedeflenmiştir. Ağırlıklandırma işleminden sonra elde edilen öznitelikler üç farklı sınıflandırma algoritması ile sınıflandırıldı. Geliştirilen algoritma medikal veri setlerine uygulanarak başarısı değerlendirilmiştir. Veri seti olarak doğrusal olarak ayrılabilir olmayan dağılıma sahip veri setleri tercih edilmiştir. Önerilen yöntemin başarısını test etmek için, sınıflandırma doğruluğu, duyarlılık, özgüllük, ortalama mutlak hata, ortalama karesel hata karekökü, kappa değeri ve ROC eğrisinin altında kalan alan (AUC) değerlerinden yararlanılmıştır. Deneysel sonuçlar, önerilen yöntemin literatürdeki mevcut yöntemlere göre daha iyi sonuçlar verdiğini göstermiştir. Medikal tanı için önerilen sistem yararlı bir tıbbi karar destek aracı olarak hizmet verebilir.

Anahtar Kelimeler: *Öznitelik ağırlıklandırma; k-harmonik ortalama tabanlı öznitelik ağırlıklandırma; Veri madenciliği; Veri ön işleme.*

*Yazışmaların yapılacağı yazar: Musa PEKER. musa@*mu.edu.tr*; Tel: (252) 211 56 71

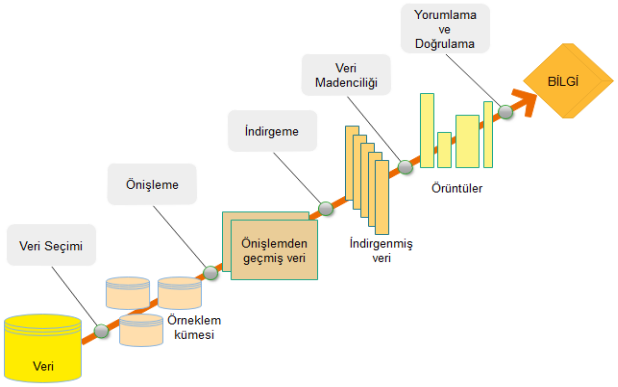
Giriş

Veri madenciliği, veri tabanlarından önceden bilinmeyen, nitelikli bilgilerin keşfedilmesi sürecidir (Olson ve Delen, 2008). Veri madenciliği çeşitli aşamalardan oluşmaktadır. Bu aşamalar Şekil 1'de sunulmuştur. Veri seçimi aşamasından sonra ön işleme aşaması gelmektedir. Bu aşamada normalizasyon, öznitelik seçimi, eksik verilerin tamamlanması ve öznitelik ağırlıklandırma gibi farklı yöntemler uygulanmaktadır. Bu veri ön işleme yöntemleri genellikle sınıflandırma algoritmalarının performansını arttırmak için giriş verilerine uygulanmaktadır. Bu yöntemlerden biri olan öznitelik ağırlıklandırma ile doğrusal olarak ayrılamayan bir veri seti, doğrusal olarak ayrılabilir veri setine dönüştürülmektedir (Polat ve Gunes 2006).

Öznitelik ağırlıklandırma yöntemleri genellikle kümeleme yöntemleri kullanılarak geliştirilmektedir. Kümeleme yöntemi, veriler arasındaki benzerlik kriterine göre veriyi gruplara bölme işlemidir. Kümeleme sonucunda elde edilen kümelerin kendi içlerinde homojen, kendi aralarında ise heterojen bir yapıda olmaları beklenir. Literatürde yaygın olarak kullanılan kümeleme yöntemleri şunlardır: k-ortalama (MacQueen, 1967), k-harmonik ortalama (Zhang ve Hsu 1999), k-medoids (Kaufman ve Rousseeuw 1987) ve bulanık c-ortalamalar (Bezdek, 1981). Literatürde farklı kümeleme yöntemleri kullanılarak geliştirilen öznitelik ağırlıklandırma yöntemleri mevcuttur. Polat ve Güneş (2006), veri ön işlemi yöntemi olarak k-en yakın komşu algoritması tabanlı bir özellik ağırlıklandırma yöntemi geliştirmiştir. Geliştirilen bu yöntem tıbbi veri setlerine uygulanmıştır. Tahir ve diğerleri (2007), k-NN ve Tabu arama algoritmalarını içeren melez bir öznitelik ağırlıklandırma yöntemi önerdi. Sun (2007), RELIEF algoritmasına dayanan ve tekrarlamalı RELIEF olarak isimlendirilen bir ağırlıklandırma yöntemi önerdi. Polat ve Durduran (2011) eksiltici kümeleme temelli ağırlıklandırma isimli yeni bir ağırlıklandırma yöntemi önerdi. Önerilen yöntem trafik kazalarını tespit etmek için uygulanmıştır. Çalışmada etkili sonuçlar elde edilmiştir. Ünal ve diğerleri (2014), bulanık c-ortalama tabanlı bir ağırlıklandırma yöntemi önerdi. Önerilen yöntem ile sınıflandırma performansı önemli ölçüde arttırılmıştır. Genel olarak ağırlıklandırma yöntemleri ile başarılı sonuçların elde edildiği görülmektedir.

Literatürde özellikle k-ortalamalar algoritması kullanılarak geliştirilen ağırlıklandırma yöntemleriyle iyi sonuçlar elde edilmiştir (Güneş ve diğerleri 2010, Polat ve Durduran 2012). K-ortalamalar yöntemi etkili olmakla beraber bazı dezavantajları da olan bir yöntemdir. En önemli dezavantajı ise başlangıç aşamasında merkez değerlere karşı olan duyarlılığıdır. K-ortalamalar algoritmasındaki bu problemden dolayı alternatif yöntemler geliştirilmiştir. Bu yöntemlerden birisi de k-harmonik ortalama (KHO) kümeleme algoritmasıdır. KHO, her bir veri noktasından merkezlere, mesafelerin harmonik ortalamasını kullanır. K-ortalamalar algoritmasından farklı olarak, KHO algoritması başlangıç nokta seçimine karşı duyarsızdır. Bu yönüyle KHO algoritmasının öznitelik ağırlıklandırma amaçlı kullanılması durumunda, etkili sonuçlar vereceği öngörülmüştür. Önerilen yöntem k-harmonik ortalama tabanlı öznitelik ağırlıklandırma (khmFW) olarak isimlendirilmiştir. khmFW algoritması ile doğrusal olarak ayrılamayan veri setini, doğrusal ayrılabilir veri setine dönüştürmek hedeflenmiştir. Önerilen yöntem, bu amacı gerçekleştirmek için birbirine daha yakın veri noktalarını bir araya getirmektedir.

Araştırmacılar geliştirdikleri yöntemlerin etkinliğini test etmek için ortak kullanıma açık olan veri setleri üzerinde deneyler yapmaktadır. Bu amaçla kullanılan veri tabanlarından biri de UCI makine öğrenme deposudur (Bache ve Lichman 2013). Bu çalışmada önerilen yöntem bu veri tabanından alınan kalp hastalığı veri seti ve BUPA karaciğer bozuklukları veri setleri üzerinde test edildi. Bu veri setleri doğrusal olarak ayrılabilir olmayan dağılıma sahiptir.



Şekil 1. Veri madenciliği aşamaları

Makalenin organizasyonu şu şekildedir. Metot başlıklı bölümde, bu çalışmada kullanılan yöntemler hakkında bilgiler verilmiştir. Ayrıca önerilen yöntemin yapısı ve çalışma şekli hakkında bilgiler bu bölümde verilmiştir. Deneysel Tasarım başlıklı bölümde, veriler hakkında bilgiler ve değerlendirme aşamasında kullanılan performans değerlendirme ölçütleri hakkında bilgiler sunulmuştur. Bulgular ve Tartışma bölümde, deneysel sonuçlar ve tartışma bölümü sunulmuştur. Ayrıca bu bölümde literatürdeki mevcut yöntemlerle karşılaştırmalı analizler yapılmıştır. Elde edilen sonuçlar hakkında genel bir bilgi Sonuçlar bölümünde sunulmuştur.

Metot

K-harmonik ortalama algoritması

K-ortalama algoritmasından farklı olarak, KHO, başlangıçtaki çözüme dayanmaz. K-ortalama kümeleme algoritması veri noktalarının her birine eşit ağırlık verir. KHO algoritması ise bir harmonik ortalama ile her bir veri noktasına dinamik ağırlık verir. Harmonik ortalama, merkeze yakın olmayan bir veri noktasına büyük bir ağırlık atarken, merkeze yakın veri noktasına küçük bir ağırlık atar.

Bu nedenle, KHO algoritması k-ortalamalar kümeleme algoritmasına göre başlangıç değerlerine daha az duyarlıdır. K-ortalamalar algoritması ile karşılaştırıldığında, belli durumlarda kümeleme sonuçlarının kalitesini artırır (Zhang ve Hsu 1999). KHO algoritmasının formülizasyonu aşağıda verilmiştir (Zhang ve Hsu 1999).

$X = (x_1, \dots, x_n)$, veri kümesi; $C = (c_1, \dots, c_k)$, küme merkezleri; $m(c_j|x_i)$, c_j merkezine ait veri noktası orantısını tanımlayan üyelik fonksiyonu; $w(x_i)$: sonraki iterasyonlarda x_i'nin, merkez parametrelerini tekrar hesaplamada ne kadar etkisi olduğunu tanımlayan ağırlık fonksiyonudur. KHO algoritması aşağıda verilmiştir.

1. Başlangıç merkezleri rastgele seçilir.
2. Amaç fonksiyonu değeri Denklem 1 kullanılarak hesaplanır.

$$KHO(X, C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \qquad (1)$$

burada p bir girdi parametresidir.

3. Veri noktası x_i'nin, merkez değeri c_j ye üyeliği Denklem 2 kullanılarak hesaplanır.

$$(c_j/x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}, \qquad (2)$$
$$m(c_j/x_i) \in [0, 1]$$

4. Veri noktası x_i'nin ağırlık değeri $w(x_i)$, Denklem 3 kullanılarak hesaplanır.

$$w(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^{k} \|x_i - c_j\|^{-p}\right)^2} \quad (3)$$

5. Üyelik ve ağırlık değerlerine bağlı olarak yeni merkez değeri Denklem 4 kullanılarak hesaplanır.

$$c_j = \frac{\sum_{i=1}^{N} m(c_j/x_i).w(x_i).x_i}{\sum_{i=1}^{N} m(c_j/x_i).w(x_i)} \quad (4)$$

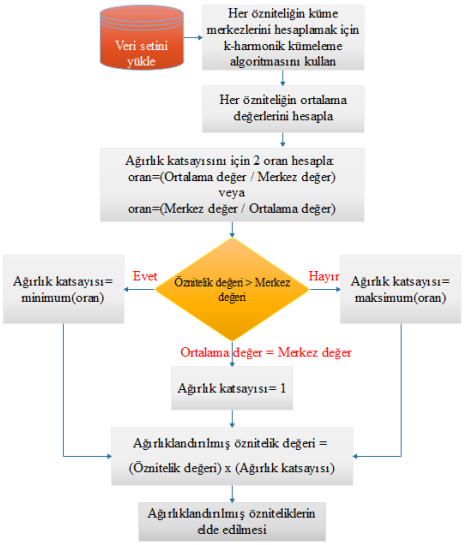
6. 2-5 adımları, durdurma kriterleri sağlanana kadar tekrar eder. Durdurma kriteri, önceden belirlenmiş bir iterasyon sayısına ulaşma veya amaç fonksiyonunun $(KHO(X,C))$ önemli ölçüde değişmesidir.

7. x_i'yi, küme j'ye en büyük $m(c_j|x_i)$ değeri ile atayın.

K-harmonik ortalama tabanlı öznitelik ağırlıklandırma (khmFW)

Bu çalışmada KHO tabanlı yeni bir öznitelik ağırlıklandırma (khmFW) yöntemi geliştirilmiştir. khmFW yöntemi aşağıdaki gibi çalışır: Öncelikle küme merkezleri, k-harmonik ortalama yöntemiyle bulunur. Sonrasında kümelere göre özniteliklerin ortalama değerleri hesaplanır. Sonraki aşamada iki oran elde edilir. Bu oranlardan ilki merkez değer/ortalama değerdir. Diğer oran ise ortalama değer/merkez değerdir. Veri kümesindeki her veri bu oranlardan biriyle çarpılır. Eğer veri değeri merkez değerden büyükse, küçük değere sahip oran ile çarpılır. Eğer veri değeri merkez değerden küçükse, büyük değere sahip oran ile çarpılır. Eğer veri değeri merkez değere eşit ise 1 ile çarpılır. Sonuç olarak ağırlıklandırılan verinin merkez değere yaklaşması sağlanır. Şekil 2'de, khmFW yönteminin akış şeması verilmiştir.

Önerilen yöntemin daha iyi anlaşılması için Şekil 3'de m örnek ve n tane özniteliğe sahip bir veri seti üzerinde khmFW yönteminin çalışma mantığı açıklanmıştır. Örnekte, f_1 özniteliğine ait değerlerin ağırlıklandırması sunulmuştur. Şekilde $f_1, f_2, f_3, \ldots, f_n$ öznitelik isimlerini ifade etmektedir. 1 numaralı bölümde öznitelikler ve değerleri verilmiştir. 2 numaralı bölümde ise KHO kümeleme yönteminin uygulanması sonucunda, sınıflara ayrılma işlemi yapılmaktadır.



Şekil. 2. khmFW yöntemi ile öznitelik ağırlıklandırma

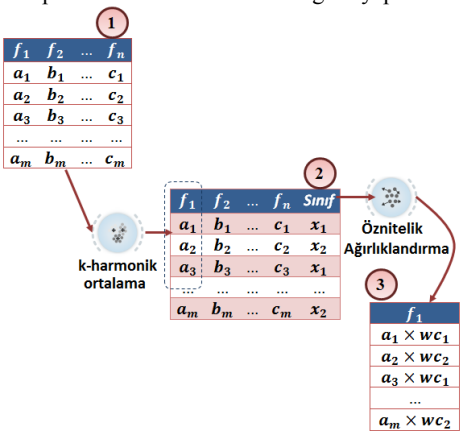
khmFW ile ağırlıklandırma katsayıları hesaplanmış ve bulunan değerler öznitelik değerleri ile çarpılmıştır. 3 numaralı alanda özniteliğin yeni değerleri sunulmaktadır. $wc1$ ve $wc2$ ağırlıklandırma katsayıları aşağıdaki gibi hesaplanmaktadır. Örnek olarak x_1 ve x_2 sınıfları verilsin. x_1 sınıfına bağlı olarak ağırlıklandırma katsayısının $(wc1)$ hesaplanması Denklem 5 ve 6'ya göre yapılır.

$$k_1 = \frac{\sum_{x=1}^{y} a_x}{y} \quad (5)$$

$$wc1 = \frac{k_1}{l_1} \; veya \; wc1 = \frac{l_1}{k_1} \; veya \; wc1 = 1 \quad (6)$$

Burada $x = 1,2,\ldots,y$, f_1 özniteliğine ait değerlerdir. Örneğin Şekil 3'de 2 numaralı alanda x_1 sınıfına bağlı f_1 özniteliğine ait değerler $a_1$ve a_3'tür. y, x_1 sınıfında ilgili öznitelikte kaç değer olduğunu gösterir. k_1 ortalama öznitelik değeridir. l_1, KHO kümeleme yönteminin uygulanması sonucunda x_1 sınıfında ilgili özniteliğe ait merkez değerini ifade eder.

$wc1$ ağırlıklandırma katsayısıdır. Eğer öznitelik değeri merkez değerden küçük ise değeri büyük olan $wc1$, Eğer öznitelik değeri merkez değerden büyük ise değeri küçük olan $wc1$ kullanılır. Eğer öznitelik değeri merkez değere eşit ise $wc1 = 1$ denklemi kullanılır. x_2 sınıfına bağlı olarak ağırlıklandırma katsayısının ($wc2$) hesaplanması Denklem 7 ve 8'e göre yapılır.



Şekil. 3. khmFW ile bir özniteliğe (f_1) ait değerlerin ağırlıklandırılması

$$k_2 = \frac{\sum_{x=1}^{b} a_x}{b} \quad (7)$$

$$wc2 = \frac{k_2}{l_2} \; ve \; wc2 = \frac{l_2}{k_2} \; ve \; wc2 = 1 \quad (8)$$

Burada $x = 1,2,\dots,b$, f_2 özniteliğine ait değerlerdir. Örneğin Şekil 3'de 2 numaralı alanda x_2 sınıfına bağlı f_1 özniteliğine ait değerler $a_2$ve a_m'dir. b, x_2 sınıfında ilgili öznitelikte kaç değer olduğunu gösterir. k_2 ortalama öznitelik değeridir. l_2, KHO algoritmasının uygulanması sonucunda x_2 sınıfında ilgili özniteliğe ait merkez değerini ifade eder. $wc2$ ağırlıklandırma katsayısıdır. Eğer öznitelik değeri merkez değerden küçük ise değeri büyük olan $wc2$, Eğer öznitelik değeri merkez değerden büyük ise değeri küçük olan $wc2$ kullanılır. Eğer öznitelik değeri merkez değere eşit ise $wc2 = 1$ denklemi kullanılır.

Sınıflandırma algoritmaları

a) Lojistik Regresyon

Lojistik regresyon yönteminin amacı, bağımlı değişkenler ile bağımsız değişken arasındaki ilişkiyi araştırmaktır. Diğer bir ifadeyle amaç, uygun sayıda değişken ile açıklayıcı değişkeni ve sonuç değişkenler arasındaki bağıntıyı tanımlayan kabul edilebilir bir model oluşturmaktır. Bu yöntemde bağımlı değişkenin sürekli olması gibi bir varsayım bulunmamaktadır, genellikle bağımlı değişkenin iki veya daha çok değer aldığı durumlarda kullanılmaktadır (Freedman, 2009).

b) Karar Ağacı

Karar ağaçları, sınıflandırma problemlerinde sıkça kullanılır. Bunun nedeni, yapılandırılması ve anlaşılmasının diğer algoritmalara nazaran daha kolay olmasıdır. Karar ağaçları, sınıflandırma aşamasında ardışık bir yaklaşım kullanır. İlk aşamada ağaç oluşturulur. İkinci aşamada ise bu ağaca veriler tek tek uygulanarak sınıflandırma işlemi yapılır. Karar ağaçları için geliştirilen birçok algoritma bulunmaktadır. Bilinen algoritmalar ID3, C4.5 ve C5' dir. Entropi değerine bağlı olarak hesaplanan C4.5 algoritması Quinlan (1993) tarafından geliştirilmiştir.

c) İleri Beslemeli Yapay Sinir Ağları (İBYSA)

Yapay sinir ağları, insan beyninin bilgi işleme yönteminden yararlanarak geliştirilmiş bir matematiksel sistemdir (Alpaydın, 2014). Bu çalışmada sinir ağı modellerinden ileri beslemeli sinir ağı kullanılmıştır. İleri beslemeli yapay sinir ağlarında temel olarak 3 farklı katman bulunur. Bu katmanlar sırasıyla; ağa sunulan verileri tutan giriş katmanı, hesaplamaların yapıldığı, hedeflenen sonuca göre kendisini eğiten gizli katman ve çıkış değerlerini gösteren çıkış katmanıdır.

Deneysel Tasarım

Veri Setleri

Bu çalışmada önerilen yöntemin başarısını değerlendirmek için 2 farklı veri seti üzerinde çalışmalar yapılmıştır. Veriler UCI makine öğrenmesi veri tabanından alınmıştır (Bache ve Lichman 2013). Bu veri setleri kalp hastalığı ve karaciğer bozuklukları ile ilgili verilerden oluşur. Bu veri kümeleri ile ilgili özet bilgiler aşağıda verilmiştir.

Statlog kalp hastalığı veri seti 270 adet veriden oluşmaktadır (Bache ve Lichman 2013). Veriler sağlıklı kişilerden ve kalp rahatsızlığı olan hastalardan alınmıştır. Bu verilerin 120 tanesi sağlıklı kişilere, 150 tanesi ise hastalara aittir. Bu veri setindeki öznitelikler sırasıyla şu şekildedir. Yaş, cinsiyet, dinlenme halinde kan basıncı, göğüs ağrısı tipi (toplam dört değerden oluşur: 1, 2, 3 ve 4), serum kolestrol (mg/dl), dinlenme halinde elektrokardiyografi (EKG) sonuçları (toplam üç değer: 0,1 ve 2), açlık kan şekeri (iki değer: 0 ve 1), Oldpeak: ST (maksimum=6.2, minimum=0), Ulaşılan maksimum kalp atış oranı, Flourosopy ile boyanmış ana damarların sayısı (toplam 4 değer: 0,1,2 ve 3), peak egzersizin ST parçasının eğimi (toplam 3 değer: 1,2 ve 3), egzersizin neden olduğu anjin (iki değer: 0 ve 1), Thal; 3=normal; 6=fixeddefect; 7=reversabledefect (toplam 3 değer).

BUPA karaciğer bozuklukları veri seti 6 öznitelik ve iki sınıftan oluşan 345 örnek içerir (Bache ve Lichman 2013). Bu verilerin 200 tanesi karaciğer bozukluğu olmayan sağlıklı kişilere aittir. Kalan 145 tane veri ise karaciğer bozukluğuna sahip hasta bireylere aittir. Veri seti 6 öznitelikten oluşmaktadır. Toplanan veri örneklerinin ilk 5 özniteliği kan testi sonuçları olup, son öznitelik günlük alkol tüketimini içerir. Veri kümesinde bulunan öznitelikler sırasıyla şu şekildedir: mcv: ortalama eritrosit hacmi, sgpt: alaminaminotransferaz, alkphos: alkalin fosfataz, gammagt: gamma-glutamiltransferaz, sgot: aspartataminotransferaz, içecekler: Bir günde içilen alkollü içeceklerin yarım litre eşdeğer sayısıdır.

Değerlendirme Ölçütleri

Bütün deneyler Intel(R) Core™ i7-2670QM (2.2 GHz) ve 8 GB RAM özelliklerine sahip bir bilgisayarda MATLAB platformunda gerçekleştirildi. Bütün deneylerde eğitim ve test verileri 10-kere çapraz doğrulama kullanılarak belirlendi. Sonuçların güvenirliliği için deneyler 10 defa tekrar edilmiş ve elde edilen değerlerin ortalamaları hesaplanmıştır.

Önerilen yönteminin başarı performası, 7 farklı performans değerlendirme kriteri ile test edilmiştir. Bunlar sırasıyla; sınıflandırma doğruluğu (DO), duyarlılık oranı (SE), özgüllük oranı (SP), ortalama karesel hata (OKH), ortalama karesel hata karekökü (RMSE), kappa değeri ve ROC eğrisinin altında kalan alan (AUC) değerleridir. Testin hasta ve sağlam olarak toplam doğru tanı oranına sınıflandırma doğruluğu denir. Testin, gerçek hastalar içerisinden hastaları ayırma yeteneği, Duyarlılık olarak ifade edilmektedir. Testin, gerçek sağlam kişiler içinden sağlam kişileri ayırma yeteneği ise Özgüllük olarak ifade edilmektedir. Denklem 9, 10 ve 11'de bu 3 ölçütün denklemleri sunulmuştur.

$$DO = \frac{DP + DN}{DP + YP + YN + DN} x100 \quad (9)$$

$$SE = \frac{DP}{DP + YN} x100\% \quad (10)$$

$$SP = \frac{DN}{YP + DN} x100\% \quad (11)$$

Burada, Doğru Negatif (DN) sağlıklı verilerin doğru sınıflandırılma sayısını, Doğru Pozitif (DP) hastalıklı verilerin doğru sınıflandırılma sayısını, Yanlış Pozitif (YP) hastalıklı verilerin yanlış sınıflandırılma sayısını ve Yanlış Negatif (YN) sağlıklı verilerin yanlış sınıflandırma sayısını ifade etmektedir.

OKH, gerçek ve tahmin edilen veri değerlerinin farkının, gerçek değere bölündükten sonra her bir sonuç için yüzde olarak toplanmasıyla elde edilen değerlendirme ölçütüdür. OKH'nın sıfır

değerine yakın olması yöntemin başarılı olduğunu ve yöntem ile elde edilen sonucun istenen değere kuvvetli bir biçimde yaklaştığını ifade eder.

RMSE, model tahminleri ile ölçüm değerleri arasındaki hata oranını belirlemek amacıyla kullanılan bir değerlendirme ölçütüdür. RMSE değerinin sıfır değerine yaklaşması modelin tahmin kabiliyetinin artması anlamına gelir. RMSE, Denklem 12'de görüldüğü gibi hesaplanmaktadır.

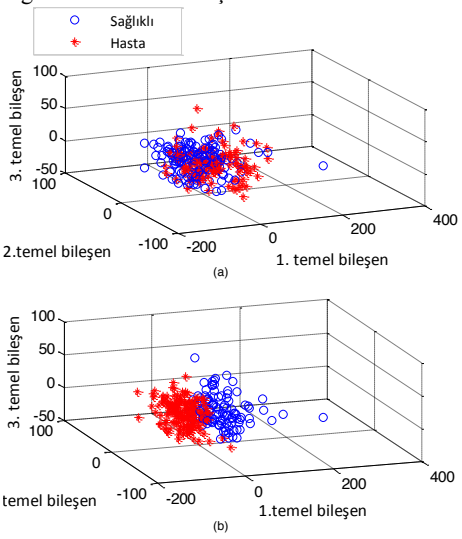
$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(u-y)^2} \quad (12)$$

Burada N kayıt sayısı, y değeri k zamanında gerçekleşen değeri, u değeri k zamanı için tahmin edilen değeri ifade etmektedir.

Performans değerlendirme kapsamında, Kappa istatistik değeri de kullanılmıştır. Bu değer iki ya da ikiden çok değerlendiricinin yapmış oldukları değerlendirmeler arasındaki uyumu hesaplamada kullanılmaktadır (Peker, 2016). ROC eğrileri, testler arasında güvenilir bir karşılaştırma olanağı sağlaması açısından sıklıkla tercih edilmektedir. Bir tanı testi için ROC altında kalan alan (AUC), 0.50 ile 1.00 arasında değerler almaktadır. Bu alanın büyük olması, tanı testinin iyi bir ayrım yeteneğine sahip olduğunu gösterir.

Bulgular ve Tartışma

Önerilen yöntemin 2 farklı veri setine uygulanması ile elde edilen sonuçlar aşağıda sunulmaktadır. Öncelikle veri setlerindeki öznitelikler khmFW yöntemi ile ağırlıklandırıldı. Şekil 4 ve Şekil 5'de her veri tabanı için orijinal ve ağırlıklı örneklerin iki sınıfa dağılımı görülmektedir. Bu şekillerde temel bileşen analizi algoritmasının uygulanmasından sonra elde edilen en iyi üç temel bileşen ile oluşturulan orijinal ve ağırlıklandırılmış örneklerin iki sınıfa 3 boyutlu dağılımı sunulmuştur. Şekillerde de görüldüğü gibi orijinal veri setinin farklılaştırma yeteneği, khmFW yöntemi ile önemli ölçüde geliştirilmiştir. Sonraki aşamada elde edilen ağırlıklı öznitelikler farklı sınıflandırma yöntemleri ile sınıflandırılmıştır. Sınıflandırma aşamasında İBYSA, lojistik regresyon ve karar ağaçları algoritmaları kullanılmıştır. İBYSA algoritmasında iterasyon sayısı 1000 olarak belirlenmiştir. Yapılan deneylerde iyi sonuçlar verdiği için öğrenme katsayısı ve momentum katsayısı değerleri sırasıyla 0.5 ve 0.1 olarak belirlenmiştir. Gizli katman nöron sayısı 15 olarak belirlenmiş ve gizli katman aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılmıştır. Sinir ağında eğitim algoritması olarak geri yayılımlı algoritma tercih edilmiştir. Lojistik regresyon algoritmasında düzenleme yöntemi olarak Ridge (L2) yöntemi kullanılmıştır. Karar ağacı olarak C4.5 algoritması kullanılmıştır.



Şekil. 4. Kalp hastalığı veri setine öznitelik khmFW algoritmasının uygulanması

Önerilen öznitelik ağırlıklandırma yönteminin başarısını değerlendirmek için farklı açılardan birçok deney yapılmıştır. Kalp hastalığı veri seti için elde edilen sonuçlar Tablo 1'de sunulmaktadır. Tabloda görüldüğü gibi, İBYSA, lojistik regresyon ve karar ağaçları yöntemlerinin uygulanması ile sırasıyla %88.51, %89.62 ve %83.33 sınıflandırma doğruluğu elde edilmiştir. Orjinal veri setine aynı yöntemlerin uygulanması ile sırasıyla %80.74, %83.7 ve

%76.6 sınıflandırma doğruluğu elde edilmiştir. Sonuç olarak hem ağırlıklandırılmış verilerle hem de orjinal verilerle Lojistik Regresyon yöntemi daha iyi sonuçlar vermiştir. Ağırlıklandırılmış verilerle elde edilen sonuçların ise her durumda orjinal verilerle elde edilen sonuçlardan daha iyi olduğu görülmektedir. Doğruluk oranı dışındaki diğer performans değerlendirme ölçütlerinde de ağırlıklı verilerle daha iyi sonuçların elde edildiği görülmektedir.

BUPA karaciğer bozuklukları veri seti için elde edilen sonuçlar Tablo 2'de sunulmaktadır. Bu veri seti için, İBYSA, lojistik regresyon ve karar ağaçları yöntemlerinin uygulanması ile sırasıyla %85.79, %80.86 ve %83.18 sınıflandırma doğruluğu elde edilmiştir. Orjinal veri setine aynı yöntemlerin uygulanması ile sırasıyla %69.27, %68.11 ve %68.69 sınıflandırma doğruluğu elde edilmiştir. Ağırlıklı verilerle elde edilen sonuçların ise her durumda orjinal verilerle elde edilen sonuçlardan daha iyi olduğu görülmektedir. Doğruluk oranı dışındaki diğer performans değerlendirme ölçütlerinde de ağırlıklı verilerle daha iyi sonuçların elde edildiği görülmektedir. Sonuç olarak hem ağırlıklı verilerle hem de orjinal verilerle İBYSA yöntemi daha iyi sonuçlar vermiştir.

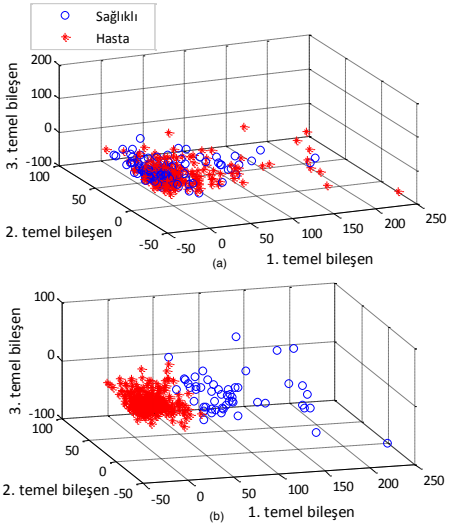
Önerilen yöntem ile elde edilen sonuçların, daha önce yapılan çalışmalarla karşılaştırmalı analizi Tablo 3 ve Tablo 4'de sunulmaktadır. Bu tabloda doğruluk oranı dışında özgüllük ve duyarlılık değerleri de karşılaştırılmıştır. Kalp hastalığı veri seti için karşılaştırmalı analiz Tablo 3'de sunulmaktadır. Bu tablo incelendiğinde, diğer araştırmacılar tarafından genel olarak %82-%88 aralığında değişen doğruluk değerleri elde edilmiştir. Önerilen metot ile aynı veri seti için %89.62 sınıflandırma doğruluğu elde edilmiştir. Karaciğer bozuklukları veri seti için karşılaştırmalı analiz Tablo 4'de sunulmaktadır. Tablo incelendiğinde diğer araştırmacılar tarafından genel olarak %60-%85 aralığında değişen doğruluk değerleri bulunmuşken, bu çalışmada %85.79 sınıflandırma doğruluğu elde edilmiştir. Ayrıca her iki veri seti üzerinde elde edilen sonuçlar incelendiğinde doğruluk oranı dışındaki diğer başarı değerlendirme ölçütlerinde de diğer çalışmalara göre daha iyi sonuçların elde edildiği görülmektedir.

Tablo 1. Kalp hastalığı veri seti için elde edilen sonuçlar

ÖZNİTELİKLER	ÖLÇÜTLER	İBYSA	LOJİSTİK REGRESYON	KARAR AĞACI
Orijinal Öznitelikler	Doğruluk	80.74	83.70	76.66
	OKH	0.194	0.224	0.274
	RMSE	0.399	0.352	0.461
	Kappa	0.610	0.668	0.527
	AUC	0.875	0.900	0.744
Ağırlıklı Öznitelikler	Doğruluk	88.51	89.62	83.33
	OKH	0.144	0.106	0.166
	RMSE	0.327	0.321	0.408
	Kappa	0.766	0.790	0.658
	AUC	0.896	0.891	0.826

Tablo 2. Karaciğer bozuklukları veri seti için elde edilen sonuçlar

ÖZNİTELİKLER	ÖLÇÜTLER	İBYSA	LOJİSTİK REGRESYON	KARAR AĞACI
Orijinal Öznitelikler	Doğruluk	69.27	68.11	68.69
	OKH	0.340	0.415	0.367
	RMSE	0.484	0.458	0.502
	Kappa	0.358	0.329	0.340
	AUC	0.721	0.718	0.665
Ağırlıklı Öznitelikler	Doğruluk	85.79	80.86	83.18
	OKH	0.222	0.191	0.190
	RMSE	0.317	0.437	0.379
	Kappa	0.707	0.612	0.651
	AUC	0.934	0.810	0.874



Şekil. 5. BUPA veri setine khmFW algoritmasının uygulanması

Deney aşamasında son olarak önerilen öznitelik ağırlıklandırma yönteminin etkisini daha iyi gözlemlemek için literatürdeki mevcut çalışmalarda kullanılan yöntemlere khmFW ağırlıklandırma yöntemi de dâhil edilerek sonuçlardaki değişimler gözlemlendi. Kalp hastalığı veri seti için 3 çalışmada kullanılan sınıflandırma algoritmaları ile deneyler yapıldı. Kahramanlı ve Allahverdi (2008), bulanık yapay sinir algoritmasını kullandı. Subbulakshmi ve ark. (2012) kalp hastalığı verilerini aşırı öğrenme makineleri ile sınıflandırdı. Mantas ve Abellán (2014) çalışmasında Credal C4.5 olarak isimlendirdikleri kesin olmayan olasılıklara dayalı bir karar ağacı yapısı kullandılar. Karaciğer bozuklukları veri seti için 2 çalışmada kullanılan sınıflandırma algoritmaları ile deneyler yapıldı. Savitha ve ark. (2012) sınıflandırıcı olarak tamamen karmaşık değerlikli radyal tabanlı sinir ağı kullandılar. Mantas ve Abellán (2014) daha önce de belirtildiği gibi Credal C4.5 olarak isimlendirilen bir karar ağacı yöntemi ile verileri sınıflandırdı. Elde edilen sonuçlar Tablo 5'de sunulmaktadır. Tabloda Yöntem 1 olarak belirtilen kısım yazarların önerdiği yöntemle elde ettikleri sonuçlardır. Yöntem 2 ile belirtilen ise öznitelik ağırlıklandırma yöntemi uygulandıktan sonra elde edilen sonuçtur. Tabloda görüldüğü gibi öznitelik ağırlıklandırma yönteminin sonuçlara etkisi olumlu olmuştur. Özellikle de karaciğer bozuklukları veri setinde doğruluk oranı önemli ölçüde artmıştır.

Tablo 5. Literatürdeki mevcut yöntemlere khmFW yönteminin uygulanması

VERİ SETİ	YAZARLAR	YÖNTEM 1	YÖNTEM 2
Kalp Hastalığı	Kahramanli ve Allahverdi (2008)	86.80	89.50
	Subbulakshmi ve ark. (2012)	87.50	88.80
	Mantas ve Abellán (2014)	80.33	84.25
Karaciğer Bozuklukları	Savitha ve ark. (2012)	74.60	80.21
	Mantas ve Abellán (2014)	64.53	72.65

Sonuçlar

Bu çalışmada sınıflandırmada başarı performansının iyileştirilmesine yönelik olarak yeni bir veri ön işleme yöntemi önerilmektedir. Bu kapsamda 2 veri seti üzerinde önerilen yöntemin etkisi araştırıldı. Önerilen yöntemde veri setlerindeki özniteliklerin varyansını azaltmak için khmFW bir öznitelik ağırlıklandırma yöntemi olarak önerilmiştir. Sınıflandırma algoritması olarak İBYSA, Lojistik Regresyon ve Karar Ağacı algoritmaları kullanılmıştır.

Önerilen yöntem ile kalp hastalığı veri seti ve BUPA karaciğer bozuklukları veri setleri için sırasıyla %89.62 ve %85.79 doğruluk oranları elde edilmiştir. Bu sonuçlar literatürde aynı veriler üzerinde yapılan birçok çalışmadan daha yüksektir. Önerilen ağırlıklandırma yöntemi ile veri madenciliği işlemlerinde sınıflandırma algoritmasının başarısının arttırılması mümkün olmuştur. Bu ağırlıklandırma yöntemi ile farklı problemler için daha iyi çözümlerin elde edilmesi sağlanabilir.

Tablo 3. Kalp hastalığı veri seti için literatürdeki mevcut yöntemlerle karşılaştırmalı analiz

YAZARLAR	YÖNTEM	BAŞARI DEĞERLENDİRME
Duch ve ark. (2001)	k-en yakın komşu algoritması, k=28, 7 öznitelik (10-kere çapraz doğrulama) k-en yakın komşu algoritması, k=28, Manhattan (10-kere çapraz doğrulama)	DO: 84.60–85.60 DO: 82.20–83.40
Sahan ve ark. (2005)	Öznitelik ağırlıklı yapay bağışıklık sistemi (10-kere çapraz doğrulama)	DO: 82.59
Ozsen and Gunes (2008)	Melez benzerlik ölçütü ile yapay bağışıklık sistemi (10-kere çapraz doğrulama)	DO: 83.95
Kahramanli ve Allahverdi (2008)	Bulanık yapay sinir ağı (10-kere çapraz doğrulama)	DO: 86.80; SE: 93; SP: 78.5
Ozsen ve ark. (2009)	Çekirdek tabanlı yapay bağışıklık sistemi (5-kere çapraz doğrulama)	DO: 85.93
Tian ve ark. (2009)	Eliptik tabanlı sinir ağı (%50-%25-%25 eğitim-doğrulama-test verileri)	DO: 82.45
Polat ve Gunes (2009)	RBF çekirdek f-skor öznitelik seçme ve en küçük kareler destek vektör makineleri (%50-%50 eğitim-test verileri)	DO: 83.70; SE: 83.92; SP: 83.54
Subbulakshmi ve ark. (2012)	Aşırı öğrenme makineleri (%70-%30 eğitim-test verileri)	DO: 87.50
Ahmadi ve ark. (2012)	Geliştirilmiş melez genetik algoritma – Çok katmanlı algılayıcı sinir ağı (%75- %25 eğitim-test verileri)	DO: 86.30; SE: 84.5; SP: 88.2
Mantas ve Abellán (2014)	Kesin olmayan olasılıklara dayalı karar ağacı (Credal C4.5) (10-kere çapraz doğrulama)	DO: 80.33
Yang ve ark. (2013)	Bulanık sınıf – etiket destek vektör makineleri (y_i - SVM) ve bulanık destek vektör makineleri (F-SVM)	DO: 85.19
Önerilen yöntem	khmFW + Lojistik Regresyon	DO: 89.62; SE: 94.56; SP: 88.4

Tablo 4. Karaciğer bozuklukları veri seti için literatürdeki mevcut yöntemlerle karşılaştırmalı analiz

YAZARLAR	YÖNTEM	BAŞARI DEĞERLENDİRME
Lee ve Mangasarian (2001)	İndirgenmiş destek vektör makineleri (10-kere çapraz doğrulama)	DO: 70.33
Polat ve ark. (2007)	Bulanık yapay bağışıklık sistemi (10-kere çapraz doğrulama)	DO: 83.38; SE: 80.95; SP: 78.57
Al-Obeidat ve ark. (2011)	Parçacık sürü optimizasyonu tabanlı melez bir yöntem (PSOPRO) (10-kere çapraz doğrulama)	DO: 69.31
Savitha ve ark. (2012)	Tamamen karmaşık değerlikli Radyal tabanlı sinir ağı (10-kere çapraz doğrulama)	DO: 74.60
Yang ve ark. (2013)	Bulanık sınıf – etiket destek vektör makineleri (y_i - SVM) ve bulanık destek vektör makineleri (F-SVM)	DO: 74.78
López ve ark. (2014)	Mahalanobis destek vektör makineleri	DO: 72.17
Mantas ve Abellán (2014)	Kesin olmayan olasılıklara dayalı karar ağacı (Credal C4.5) (10-kere çapraz doğrulama)	DO: 64.53
Ozsen ve Yucelbas (2015)	Elips-Yapay bağışıklık sistemi (5-kere çapraz doğrulama)	DO: 85.59; SE: 90.10; SP: 81.05
Ozsen ve Gunes (2008)	Melez benzerlik ölçütü ile yapay bağışıklık sistemi (10-kere çapraz doğrulama)	DO: 60.57
Li ve ark. (2011)	Bulanık tabanlı lineer olmayan dönüştürme metodu + Destek Vektör Makineleri	DO: 70.85
Önerilen yöntem	khmFW + İBYSA	DO: 85.79; SE: 91.20; SP: 82.04

Kaynaklar

Ahmad, F., Isa, N. A. M., Hussain, Z., ve Osman, M. K. (2013). Intelligent medical disease diagnosis using improved hybrid genetic algorithm - multilayer perceptron network *Journal of Medical Systems*, 37, 2, 9934.

Al-Obeidat, F., Belacela, N., Carretero, J.A. ve Mahanti, P., (2011). An evolutionary framework using particle swarm optimization for classification method PROAFTN, *Applied Soft Computing*, 11, 8, 4971-4980.

Alpaydın, E., (2014). *Introduction to machine learning*, MIT press.

Bache, K. ve Lichman, M., (2013). UCI machine learning repository. http://archive.ics.uci.edu/ ml, (18.05.2016)

Bezdek, J.C., 1981. *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum Press.

Duch, W., Adamczak, R. ve Grabczewski, K., (2001). A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, *IEEE Transactions on Neural Networks*, 12, 2, 277-306.

Güneş, S., Polat, K., ve Yosunkaya, Ş., (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting, *Expert Systems with Applications*, 37, 12, 7922-7928.

Freedman, D.A., (2009). *Statistical Models: Theory and Practice*, Cambridge University Press., 128-129.

Kahramanli, H. ve Allahverdi, N., (2008). Design of a hybrid system for the diabetes and heart diseases, *Expert Systems with Applications*, 35, 1-2, 82-89.

Kaufman, L. ve Rousseeuw, P., (1987). *Clustering by means of medoids*, North-Holland.

Lee, Y.J. ve Mangasarian, O.L., (2001). SSVM: A smooth support vector machine for classification, *Computational Optimization and Applications*, 20, 1, 5-22.

Li, D.C., Liu, C.W. ve Hu, S.C., (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical datasets, *Artificial Intelligence in Medicine*, 52, 1, 45-52.

López, F.M., Puertas, S.M. ve Arriaza, J.T., (2014). Training of support vector machine with the use of multivariate normalization, *Applied Soft Computing*, 24, 1105-1111.

MacQueen, J.B., (1967). Some methods for classification and analysis of multivariate observations, *Proceedings*, 5th Berkeley Symposium on Mathematical Statistics and Probability, 281-297.

Mantas, C.J. ve Abellán, J., (2014). Credal-C4. 5: Decision tree based on imprecise probabilities to classify noisy data, *Expert Systems with Applications*, 41, 10, 4625-4637.

Olson, D. L., ve Delen, D., (2008). *Advanced data mining techniques*, Springer Science & Business Media.

Ozsen, S. ve Yucelbas, C., (2015). On the evolution of ellipsoidal recognition regions in artificial immune systems, *Applied Soft Computing*, 31, 210-222.

Ozsen, S. ve Gunes, S., (2008). Effect of feature-type in selecting distance measure for an artificial immune system as a pattern recognizer, *Digital Signal Processing*, 18, 4, 635-645.

Ozsen, S., Gunes, S., Kara, S. ve Latifoglu, F., (2009). Use of kernel functions in artificial immune systems for the nonlinear classification problems, *IEEE Transactions on Information Technology in Biomedicine*, 13, 4, 621-628.

Peker, M., (2016). An efficient sleep scoring system based on EEG signal using complex-valued machine learning algorithms, *Neurocomputing*, 207, 165-177.

Polat, K. ve Gunes, S., (2006). A hybrid medical decision making system based on principles component analysis, k-NN based weighted pre-processing and adaptive neuro-fuzzy inference system, *Digital Signal Processing*, 16, 6, 913-921.

Polat, K., Sahan, S., Kodaz, H. ve Gunes, S., (2007). Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism, *Expert Systems with Applications*, 32, 1, 172-183.

Polat, K. ve Durduran, S.S., (2011). Subtractive clustering attribute weighting (SCAW) to discriminate the traffic accidents on Konya–Afyonkarahisar highway in Turkey with the help of GIS: A case study, *Advances in Engineering Software*, 42, 7, 491-500.

Polat, K. ve Gunes, S., (2009). A new feature selection method on classification of medical datasets: Kernel f-score feature selection, *Expert Systems with Applications*, 36, 7, 10367-10373.

Polat, K. ve Durduran, S.S., (2012). Automatic determination of traffic accidents based on KMC-based attribute weighting, *Neural Computing and Applications*, 21, 6, 1271-1279.

Quinlan, L., 1993. *C4.5: Programs for Machine Learning,* 1st ed., Morgan Kaufmann, San Franscisco, 70-80.

Sahan, S., Polat, K., Kodaz, H. ve Gunes, S., (2005). The medical applications of attribute weighted artificial immune system (AWAIS): Diagnosis of heart and diabetes diseases, *Lecture Notes in Computer Science*, 3627, 456-468.

Savitha, R., Suresh, S., Sundararajan, N. ve Kim, H.J., (2012). A fully complex-valued radial basis function classifier for real-valued classification problems, *Neurocomputing*, 78, 1, 104-110.

Subbulakshmi, C.V., Deepa, S.N., ve Malathi, N., (2012). Extreme learning machine for two category data classification, *Proceedings*, In 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 458-461.

Sun, Y., (2007). Iterative RELIEF for feature weighting: algorithms, theories, and applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 6, 1035-1051.

Tahir, M.A., Bouridane, A. ve Kurugollu, F., (2007). Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier, *Pattern Recognition Letters*, 28, 4, 438-446.

Tian, J., Li, M. ve Chen, F., (2009). A hybrid classification algorithm based on coevolutionary EBFNN and domain covering method, *Neural Computing and Applications*, 18, 3, 293-308.

Unal, Y., Polat, K. ve Kocer, H.E., (2014). Pairwise FCM based feature weighting for improved classification of vertebral column disorders, *Computers in Biology and Medicine*, 46, 61-70.

Yang, C.Y., Chou, J.J. ve Lian, F.L., (2013). Robust classifier learning with fuzzy class labels for large-margin support vector machines, *Neurocomputing*, 99, 1-14.

Zhang, B. ve Hsu, M., (1999). *K-harmonic means: a data clustering algorithm*, Technical Report, Hewlett-Packard Labs, HPL-1999-124.

A Novel Data Pre-processing Method: K-Harmonic Means based Feature Weighting

Extended abstract

Data mining is the analysis of data for relationships that have not previously been discovered (Olson & Delen, 2008). Data mining involves various stages. One of the most important is the data pre-processing stage. In this stage, different methods such as normalization, feature selection, completion of missing data and feature weighting are applied. Data pre-processing methods are applied to input data in order to improve the performance of the classification algorithms. Feature weighting methods included in data pre-processing methods are of great importance in the field of machine learning. With weighting methods, a linearly inseparable dataset is converted into linearly separable datasets (Polat & Gunes, 2006).

Feature weighting algorithms are usually developed by using clustering methods. In the literature, researchers are presented different methods on this subject. In particular, successful results have been obtained with k-means clustering based on weighting methods (Güneş & Ark, 2010, Polat & Durduran, 2012). The k-means method is an effective method although it also has some disadvantages. The most important disadvantage is its sensitivity to the central values in the initial stages. Alternative methods have been developed because of this problem associated with the k-means algorithm. One of these methods is use of the k-harmonic means clustering algorithm. The algorithm uses the harmonic mean of the distance to the centre from each data point. Unlike the k-means clustering algorithm, the k-harmonic means clustering algorithm is insensitive to the selection of the starting point. This being the case, it is expected that it will provide effective results when used for feature weighting. The recommended weighting method is entitled khmFW and, in conjunction with the khmFW algorithm, its aim is to convert the linearly inseparable dataset into linearly separable datasets. In order to achieve this, the recommended weighting method aggregates data points which are closer to each other.

In order to test the effectiveness of a computer-aided medical diagnostic system which they developed, researchers are conducting experiments on datasets that are open to common use. One of the databases used for this purpose is the UCI machine learning repository [3]. The proposed system used in this study has been tested on a heart disease dataset and a BUPA liver disorders dataset obtained from this database. These datasets have a distribution which cannot be separated linearly.

The features obtained after the feature weighting process were then classified using a different classification algorithm method. In order to test the performance of the proposed method, classification accuracy, specificity, sensitivity, mean absolute error, root mean square error, kappa value and AUC values under the ROC curve have been used. The experimental results show that the proposed method gave better results compared to existing methods described in the literature. The proposed system can serve as a useful medical decision support tool for medical diagnosis.

With the proposed method, accuracy rates of 89.62% and 85.79% were obtained for the heart disease data set and the BUPA liver disorders data set, respectively. These results provide better results than many studies on the same data in the literature. With the proposed weighting method, it is possible to increase the success of the classification algorithm in data mining operations. With this weighting method, better solutions can be obtained for different problems.

Keywords: *Feature weighting; k-harmonic means based feature weighting; Data mining; Data pre-processing.*

mühendislikdergi