Weka k-Mean Algoritması ile Konjestif Kalp Yetmezliği Hastalığı Teşhisi

Diagnosis of Disease with Congestive Heart Failure Using k-Means Algorithm in WEKA

Seda Güler

***Özetçe*— Bu çalışmada konjestif kalp yetmezliği olan hastaların kontrol grubundan ayırt edilmesi için kalp hızı değişkenliği verileri WEKA yazılımı kullanılarak k-Ortalama kümeleme algoritması başarımı incelenmiştir. Kalp hızı değişkenliği ölçümleri 29 adet konjestif kalp yetmezliği rahatsızlığı bulunan hastadan ve kontrol grubunda yer alan 54 kişiden elde edildikten sonra WEKA yazılımı aracılığıyla k-Ortalama kümeleme algoritmasına uygulanmıştır. Sonuç olarak, sadece dört kümenin kullanıldığı durum için en yüksek %98,79 başarıma ulaşıldığı tespit edilmiştir. Veri madenciliği alanında oldukça yüksek bir kullanım alanına sahip olan WEKA yazılımında sunulan seçenekler hakkında bilgi de verilmiştir.**

Anahtar Kelimeler — WEKA, Kalp Yetmezliği, k-Mean

Abstract: In this study, the performance of the k-means clustering algorithm was investigated by using heart rate variability WEKA software to distinguish the patients with congestive heart failure from the control group. Heart rate variability measures were applied to the k-means clustering algorithm via WEKA software after obtaining from 29 patients in 29 congestive heart failure patients and control group. As a result, it was determined that the highest 98.79% achievement was achieved when only four cones were used. Information about the options offered in the WEKA software, which has a very high usage area in the field of data mining, is also given.

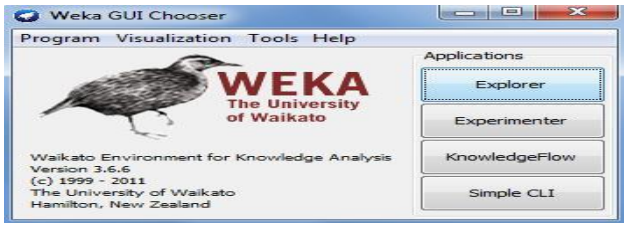
Keywords — WEKA, Heart Failure, k-Means

I. GİRİŞ

Günümüzde veriyi bilgiye dönüştürme işlemi Veri Madenciliği olarak adlandırılmaktadır. Ölçüm cihazlarının artmasına paralel olarak veri sayısı ve türleri artmaktadır. Veri toplama araçları ve veri tabanı teknolojilerindeki gelişmeler, bilgi depolarında çok miktarda bilginin depolanmasını ve çözümlenmesini gerektirmektedir. Birçok kaynaktan elde edilen veriler içerisinde saklı bulunan bilgiyi bulma işlemine veri madenciliği denilmektedir (Kudyba, 2004). Bu işlemleri yapmak için birçok program kullanılmaktadır. Açık kaynak kodlu programlar arasında WEKA, ARTool, RapidMiner, C4.5, Orange, KNIME ve R sayılabilir. Literatürde WEKA ile yapılmış birçok çalışma bulunmaktadır.Bu çalışmada amaç, WEKA'yı kullanarak sık görülen kalp rahatsızlığı olan (Hunt vd., 2005) Konjestif Kalp Yetmezliği hastalarının teşhis edilmesidir. KKY olarak kısaltılan Konjestif kalp yetmezliği, organizmanın ihtiyaçlarını karşılayacak yeterli kardiyak debinin kalp tarafından sağlanamaması durumudur. Kalp, gerekli durumlarda yedek kapasitesini kullanarak debisini %200-600 oranında artırır. Kalbin yedek kapasitesinin aşılması veya artan debi ihtiyacını karşılayamaması durumunda KKY görülür (Jovic ve Bogunovic, 2011). Bu çalışmada www.physionet.org internet adresinden ücretsiz olarak erişilebilen Physionet veritabanından veriler kullanılmıştır. WEKA'daki kümeleme algoritmalarından k-ortalama algoritması ile bu veriler üzerinden çalışma yürütülmüştür. Diğer bölümde WEKA yazılımı hakkında bilgi verilmiştir.

II. WEKA

WEKA, bilgisayar bilimlerinin önemli konularından birisi olan makine öğrenmesi konusunda kullanılan paketlerden birisinin ismidir. Waikato üniversitesinde açık kaynak kodlu olarak java dili üzerinde geliştirilmiştir. İsmi de buradan gelir ve Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden oluşur. WEKA çeşitli veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları ve görselleştirme araçlarından oluşur. Aynı zamanda yeni makine öğrenme algoritmaları geliştirmek için de uygundur. Program çalıştırıldığında Şekil 1'deki kullanıcı ara yüzü ekrana gelir. Bu ara yüz ekranı "Program", "Visualization", "Tools" ve "Help" menülerinden oluşan ana menü ve "Explorer", "Experimenter", "Knowledge Flow" ve "Simple CLI" kısımlarından oluşan "Applications" bölümlerini içerir.



Şekil 1. WEKA kullanıcı ara yüzü

"Applications" bölümünde yer alan "Explorer" seçeneği mevcut veri üzerinde yapılabilecek uygulamaları içeren genel bir grafiksel kullanıcı ara yüzü içerir. "Exprimenter" seçeneği ise bir veya daha çok veri kümesi üzerinde bir veya daha çok algoritmanın uygulanabilmesine olanak sağlayan bir kullanıcı ara yüzüdür. "Knowledge Flow" seçeneği ise Matlab içindeki Simulink veya National Instruments firmasına LabVIEW programı gibi sürükle bırak özelliğine sahip olan "Explorer" penceresi gibi çalışmaktadır. Kullanıcı tercihine bağlı olarak "Explorer" ya da "Knowledge Flow" seçeneklerini kullanabilir. Son seçenek olan "Simple CLI" ise komut ekranı aracılığıyla işlem yapmayı sağlar. Ayrıca WEKA'ya ve kullanım dokümanlarına http://www.cs.waikato.ac.nz/ml/weka internet adresinden ulaşılabilir .

III. VERİSETİ

Bu çalışmada konjestif kalp yetmezliği hastaların normal gruptan ayırt edilmesini sağlayacak bir çalışma yapılacaktır. Bu çalışmada kullanılan veriler www.physionet.org_internet adresinden erişilebilen veri tabanlarından elde edilmiştir. Kullanılan veri tabanlarındaki EKG kayıtları 24 saatlik olmasına rağmen, her bir kayıttan içinde bozuk ritim vuruları bulunmayan 5 dakikalık bir dilim kullanılmıştır. Bu çalışmada yararlanılan veri tabanları şunlardır:

- "Congestive Heart Failure RR Interval Database" (chf2db) veritabanı: yaşları 34 ile 79 arasında değişen 29 adet hastadan elde edilmiş 24 saat süreli EKG kaydı
- "Normal Sinus Rhythm RR Interval Database" (nsr2db) veritabanı: yaşları 24 ile 76 arasında değişen 54 adet 24 saat süreli normal EKG kaydı

A. KALP HIZI DEĞİŞKENLİĞİ ANALİZİ

KHD analizi için NN olarak adlandırılan normal-normal aralıkların analizi yapılmaktadır. KHD çalışmalarında, hasta bilgisi (yaş), zaman dizisi analizi (ortalama, standart sapma, 20 ve 50 ms'den fazla değişim olan verilerin sayısı ve oranı, vb.), frekans alanı analizi (çeşitli frekans aralıklarındaki spektral güç miktarları) ve doğrusal olmayan yöntemlerle elde edilen sonuçlar (Poincare ölçümlerinin yanı sıra sembolik, yaklaşık ve örnek entropisi gibi) kullanılmaktadır.

Frekans ölçümleri için çoğunlukla hızlı Fourier dönüşümü (FFT) yöntemini kullanan Welch periyodogram yöntemi kullanılmaktadır:

$$P(w)=\frac{1}{N}\sum_{j=0}^{N-1}\left|x(t_j)e^{-iwt_j}\right|^2 \qquad (1)$$

N zaman dizisi verisinin uzunluğudur.Bu yöntem kullanılarak sadece zamanda eşit aralıklarla örneklenmiş veriler üzerinden güç spektral yoğunluğu (GSY) hesaplanabilir. Bu yüzden elde edilen KHD'nin 4 Hz örnekleme hızında kübik interpolasyon metodu ile yeniden örneklendirilmesi ve analiz durağanlığı sağlamak için eğilim yok etmenin kullanılması gerekir.

Frekans ölçümlerinde, GSY üzerindeki farklı frekans aralıklarındaki güçler ve tepe frekansları hesaplanarak incelenir. KHD analizinde yaygın olarak üç frekans bandı kullanılmaktadır: VLF(0–0,033 Hz), VLF(0,033–0,15 Hz) ve HF(0,15–0,4 Hz). Bu araştırmada, klasik GSY yöntemlerine alternatif olarak geliştirilen Lomb periyodogram yöntemi ile elde edilen frekans alanı ölçümleri de kullanılmaktadır.

$$P(w_n)=\frac{1}{2\sigma^2}\left\{\frac{\left[\sum_{j=0}^{N-1}(x(t_j)-\bar{x})\cos\left(w_n(t_j-\tau)\right)\right]^2}{\sum_{j=0}^{N-1}\cos^2\left(w_n(t_j-\tau)\right)}+\frac{\left[\sum_{j=0}^{N-1}(x(t_j)-\bar{x})\sin\left(w_n(t_j-\tau)\right)\right]^2}{\sum_{j=0}^{N-1}\sin^2\left(w_n(t_j-\tau)\right)}\right\}$$

burada

$$\tau\equiv\frac{1}{2w}\tan^{-1}\left(\frac{\sum_{j=1}^{N}\sin(wt_j)}{\sum_{j=1}^{N}\cos(wt_j)}\right) \qquad (3)$$

ve P(w) hesaplamasını tüm t_j örnekleme zamanlarından bağımsız hale getiren ortalama bir değerdir. Bu yöntemle yeniden örnekleme ölçümleri elde edilebilmektedir.

Bu araştırmada kullanılan ve 5 dakikalık zaman aralığı için tanımlanan standart frekans alanı KHD ölçümleri Çizelge 1'de listelenmiştir.Frekans alanı ölçümleri Welch periyodogram ve Lomb periyodogram yöntemleri kullanılarak her ikisi için de ayrı ayrı hesaplanmıştır.

Dalgacık analizi bir sinyalin zaman ve ölçek (veya frekans) boyutlarının birlikte incelenmesine olanak tanır. Bu yüzden, dalgacıkların özellikle RR aralıklarının analizinde çok kullanışlı olduğu düşünülür. Bu analiz yönteminde de, frekans alanı ölçümlerinin hesaplanışı gibi 4 Hz ile yeniden örneklenmiş KHD verisi üzerinde çalışılmaktadır. Bu çalışmada, KHD analizinde kullanıldığı daha önce Quian (2001) tarafından rapor edilen Daubechies-4 ana dalgacığı kullanılarak 7 seviyeli dalgacık dönüşümü metodu kullanılmıştır (İşler ve Kuntalp, 2007).

Çizelge 1. Çalışmada kullanılan frekans alanı standart KHD ölçümleri.

VLF	VLF frekans bandı toplam gücü
LF	LF frekans bandı toplam gücü
HF	HF frekans bandı toplam gücü
LFHF	LF/HF frekans bantları güçleri oranı
NLF	LF / (LF + HF) oranı (Normalize LF gücü)
NHF	HF / (LF + HF) oranı (Normalize HF

Dalgacık entropisi, işaretin düzensizlik derecesinin ölçüsü olarak ortaya çıkar ve işaretle ilişkili temeli oluşturan dinamik bir süreç hakkında önemli bilgiler içerir (Quian, 2001). Shannon entropisi metodu olasılık dağılımlarının analizinde ve karşılaştırılmasında önemli bir ölçüttür.

Bu tanımdan yola çıkılarak, dalgacık entropisi aşağıdaki gibi tanımlanmaktadır:

$$S = -\sum_{j<0} p_j \ln[p_j] \qquad (4)$$

j dalgacık analizi katsayısının indisini ve p ise dalgacık analizi katsayısının olasılığını gösterir. C_j dalgacık dönüşümü katsayılarını göstermek üzere p_j şu şekilde hesaplanır:

$$p_j = \frac{C_j^2}{\sum_k C_k^2} \qquad j,k = 1,2,\ldots,N \qquad (5)$$

Burada, N frekans boyutunda incelenen nokta sayısını (yani, frekans çözünürlüğünü) verir. Burada her bir frekans alanı ölçümlerine karşılık gelen 6 adet dalgacık entropisi değeri birer öznitelik olarak çalışmaya dahil edilmiştir. Bu çalışmada, KHD analizi kullanılarak elde edilen toplam 59 adet öznitelik kullanılmıştır.

B. k-Ortalama Kümeleme Algoritması

Bu algoritma giriş uzayını k adet merkezle ifade etmeye çalışan bir yöntemdir. Merkezlere ilk değer ataması rastgele olarak yapıldıktan sonra merkez değerlerinin güncellenmesi için iki farklı yöntem kullanılır. Birinci yöntemde (batch metodu) giriş kümesindeki her bir örneğin hangi merkeze yakın olduğu hesaplanır. Aynı merkeze yakın olan örneklerin ortalaması alınarak merkezin değeri güncellenmiş olur. Durma koşulu sağlanana kadar bu işlem tekrar edilir. İkinci yöntemde (online metodu) giriş kümesinden bir örnek seçilir ve bu örneğin merkezlere olan uzaklığına bakılır. Örneğin en yakın olduğu merkez bulunarak bu merkezin değeri güncellenir. Her bir örnek için bu işlem tekrarlanır. Merkez değeri güncellenirken merkezle örnek arasındaki mesafe değeri her adımda azalan bir öğrenme katsayısıyla çarpılarak kullanılır. Bu sayede ilk adımlarda merkezlerin yer değiştirmesi büyük miktarlarda olurken zamanla yer değiştirme azalır ve merkezler yakınsar. Durma şartı sağlanıncaya kadar her örnek için bu işlem tekrarlanır. Örneklerin işlenme sırası her adımda aynı olacağı gibi tamamen rastgele de olabilir. Bu çalışmada birinci yaklaşım tercih edilmiştir.

C. Değerlendirme

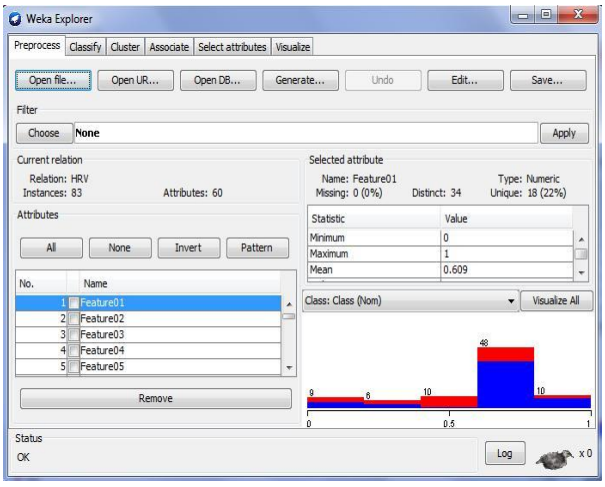
Başarımların değerlendirilmesinde kullanılan genel başarım ölçütü şu şekilde verilmektedir:

$$\text{Genel Başarım} = \frac{TP + TN}{TP + FN + TN + FP} \qquad (6)$$

Burada, TP sınıflandırıcının hasta olarak etiketlediği ve gerçekten hasta olanların sayısını, TN sınıflandırıcının sağlam olarak etiketlediği ve
gerçekten sağlam olanların sayısını, FN sınıflandırıcının sağlam olarak etiketlediği ve gerçekte hasta olanların sayısını, FP ise sınıflandırıcının hasta olarak etiketlediği ve gerçekte sağlam olanların sayısını verir. Böylece, her bir sınıflandırıcının tüm sağlam ve hasta grubunda doğru olarak verdiği tüm kararlar dikkate alınmıştır.

IV. SONUÇLAR

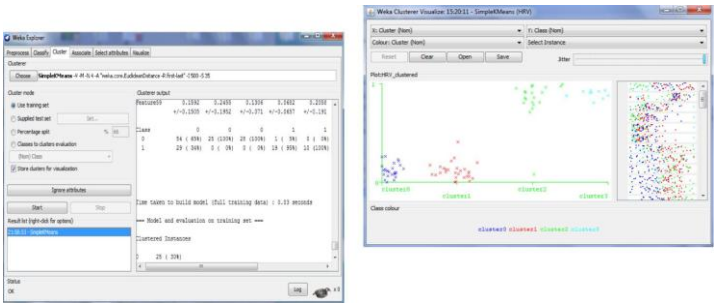
Bu çalışmada elde edilen 59 adet öznitelikten ve kayıt alınan kişinin hasta olup olmadığı bilgisinden oluşan öznitelikler kayıt edildi. Daha sonra bu veri setinin WEKA'nın desteklediği dosya biçimlerinden biri olan ARFF uzantılı dosyaya kayıt edilmesi sağlandı. WEKA yazılımı kullanılarak "Explorer" ekranından "Open file" komutu ile bu veri dosyası hafızaya alındı. Veri yüklendikten sonra oluşan ekran görüntüsü Şekil 2'de gösterilmiştir.



Şekil 2. Verinin yüklenmesi sonucu oluşan ara yüz görüntüsü

Bu ekrandaki "Current relation" başlığı altında yüklenen verinin ismi, kaç tane örneğe ve özniteliğe sahip olduğu görülür. Verilerden sol tarafta seçili olan özniteliğe ait en küçük, en büyük, standart sapma ve ortalama değer bilgileri

ekranda sağ taraftaki "Selected attribute" başlığı altında görülür. Seçilen özniteliğe ait dağılım sağ alt köşedeki grafikte görüntülenir. Bu çalışmada kullanılan veri setine herhangi bir ön işlem uygulanmamıştır. Daha sonra "Cluster" seçeneği seçilip veriye uygulanacak kümeleme algoritması "Clusterer" başlığındaki "Choose" butonu tıklanarak k-Ortalama kümeleme algoritması için "SimpleKMeans" seçilir. Bu seçimden sonra oluşan ekran görüntüsü Şekil 3'deki gibidir. Seçilen algoritmanın üstüne tıklanarak algoritmayla ilgili ayarlar yapılmış olup "Choose" butonunun sağındaki metin kutusu içinde bu ayarlar görülmektedir. Bu çalışmada öznitelik seçimi yapılmamış olup eğitim için verilen tüm veriler aynı zamanda test için kullanılmıştır. "Start" aracılığıyla çalıştırılan deney için elde edilen sonuçlar aşağıdaki şekildeki "Clusterer output" başlığı altında görülmektedir.



Şekil 3. K-ortalama kümeleme algoritması ile kümeleme ve En yüksek başarımın elde edildiği 4 adet küme için görselleştirme ekranı

Çalışmada küme sayısını belirleyen k değeri2'den 10'a kadar alınarak yukarıdaki işlemler farklı başlangıç değerleri (seed) için tekrarlanmıştır. Başlangıç değerinin etkisinin gözlenebilmesi için her k değeri için 1'den 50'ye kadar seed değerleri denenmiştir. Böylece çalışma toplam 450 kez tekrarlanmıştır. Her bir k değeri için elde edilen en yüksek başarım ve bu başarım için uygulanan seed değerleri çizelge ***2***'de özetlenmiştir. Bu sonuçlara göre KKY hastalarının teşhisinde k=4 değeri için maksimum başarım %98,79 olarak elde edilmiştir. Böylece 83 kayıttan 82 tanesi doğru olarak tanımlanabilmiştir. WEKA yazılımı "Visualise" sekmesiyle en yüksek başarımın elde edildiği k=4 değeri için sonuçlar görsel olarak elde edilmiştir (Şekil 3). Bu şekle göre normal

(0) sınıfı için Cluster0 ve Cluster1 kümeleri ve hasta
(1) sınıfı için Cluster2 ve Cluster3 kümeleri atanmıştır.
Ayrıca sadece 1 adet sağlam kişinin Cluster2 kümesine dahil edildiği (dolayısıyla hasta olarak belirlendiği) görülmektedir.

Çizelge 2. Farklı küme sayısı ve başlangıç değerleri için kümeleyici başarımları

Küme Sayısı (k)	Başarım (%)	Seed (Başlangıç)
2	92,77	3
3	93,97	25
4	98,79	25
5	97,59	5
6	97,59	5
7	96,38	3
8	96,38	5
9	96,38	15
10	96,38	15

Kaynaklar

[1] Hunt, S.A., Abraham, W.T., Chin, M.H., Feldman, A.M., Francis, G.S., Ganiats, T.G., vd. 2005. ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult. Circulation, 112, 154–235.

[2] Kudyba, S. 2004. Managing Data Mining. CyberTech Publishing, 146–163.

[3] Jovic, A., Bogunovic, N. 2011. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. Artificial Intelligence in Medicine, 51, 175–186.

[4] *http://bilgisayarkavramlari.sadievrenseker.com*

[5] İşler, Y., Kuntalp, M. 2007. Combining Classical HRV Indices with Wavelet Entropy Measures Improves to Performance in Diagnosing Congestive Heart Failure. Computers in Biology and Medicine, 37(10), 1502–1510.

[6] Quian Quiraga, R., Rosso, O.A., Başar, E., Schürmann, M. 2001. Wavelet entropy in event-related potentials: a new method shows ordering of EEG oscillations. Biological Cybernetics, 84(4), 291– 299.

[7] *http://www.aysebilge.com/*

[8] *www.physionet.org*

[9] *http://ugurozdemir.org*

[10] *www.medikalakademi.com.tr*

[11] *http://www.cs.waikato.ac.nz/ml/weka*