SÜLEYMAN DEMİREL ÜNİVERSİTESİ TEKNİK BİLİMLER MESLEK YÜKSEKOKULU *SULEYMAN DEMIREL UNIVERSITY TECHNICAL SCIENCES VOCATIONAL SCHOOL*
SDU TEKNİK BİLİMLER DERGİSİ
SDU JOURNAL OF TECHNICAL SCIENCES

YIL : 2012
CİLT : 2
SAYI : 4
SAYFA : 21-29

WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi

Yalçın İŞLER[1*], Ali NARİN[1]

[1]*Zonguldak Karaelmas Üniversitesi, Elektrik ve Elektronik Mühendisliği Bölümü, Zonguldak.*

Özet: Bu çalışmada, konjestif kalp yetmezliği (KKY) olan hastaların kontrol grubundan ayırt edilmesi için kalp hızı değişkenliği (KHD) analizi verileri üzerinde WEKA yazılımı kullanılarak k-Ortalama kümeleme algoritması başarımı incelenmiştir. KHD ölçümleri 29 adet KKY 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. Ayrıca, veri madenciliği alanında oldukça yüksek bir kullanım alanına sahip olan ve ücretsiz olarak sunulan WEKA yazılımında sunulan seçenekler hakkında bilgi de verilmiştir.

Anahtar Kelimeler: WEKA, Konjestif Kalp Yetmezliği, Kalp Hızı Değişkenliği, k-Ortalama

Diagnosis of the Patients with Congestive Heart Failure using k-Means Algorithm in WEKA Software

Abstract: In this study, the accuracy of k-Means clustering algorithm, implemented in WEKA software, in the analysis of heart rate variability (HRV) that are used in discriminating the patients with congestive heart failure (CHF) from normal subjects is investigated. After being obtained from 29 CHF patients and 54 normal subjects, HRV measures were applied to k-Means clustering algorithm using WEKA software. As a result, the maximum discrimination accuracy of 98.79% was achieved when only four clusters were used. Additionally, information about the choices given in WEKA software, which has been widely used in the data mining field and is free of charge, was also presented.
Keywords: WEKA, Congestive Heart Failure, Heart Rate Variability, k-Means

Giriş

Son yıllarda 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 açık kaynak kodlu ve ticari amaçlı birçok program kullanılmaktadır (Dener vd., 2009). Açık kaynak kodlu programlar arasında WEKA, ARTool, RapidMiner, C4.5, Orange, KNIME ve R, ticari programlar arasında ise SPSS Clementine, SPSS, SAS, Angoss, KXEN, SQL Server ve MATLAB sayılabilir (Danacı vd., 2010).

Literatürde açık kaynak kodlu veri madenciliği programı olan WEKA ile yapılmış birçok çalışma bulunmaktadır. Meme kanseri hücrelerinin teşhis ve tahmini (Danacı vd., 2010), göğüs kanserinin teşhis ve tahmini (Coşkun ve Baykal, 2011), Parkinson hastalığının teşhisi ve tahmini (Özçift, 2011) gibi çalışmalar bulunmaktadır. Sadece "IEEE Xplore" ve "ScienceDirect" veri tabanlarında "WEKA" anahtar kelimesiyle yapılan aramalarda, ilki için 144 akademik çalışma ve diğeri için 1415 makale, 88 kitap ve 6 tane referans çalışması bulunmaktadır. Sadece son beş yıl içinde, bu çalışmanın hazırlandığı 2012 yılı ilk 2 ayı için 121 makale ve 8 kitap, 2011 yılı için 357 makale ve 27 kitap, 2010 yılı için 220 makale (+2 referans çalışması) ve 3 kitap, 2009 yılı için 182 makale (+2 referans çalışması) ve 7 kitap, 2008 yılı için 138 makale ve 5 kitap bulunduğu görülmektedir.

Bu çalışmada hedeflenen amaç, WEKA'yı kullanarak sık görülen kalp rahatsızlığı olan (Hunt vd., 2005) Konjestif Kalp Yetmezliği hastalarının teşhis edilmesidir. Konjestif kalp yetmezliği (KKY), organizmanın metabolik ihtiyaçlarını karşılayacak yeterli kardiyak debinin kalp tarafından sağlanamaması halidir. Kalp, gerekli olduğu durumda yedek kapasitesini kullanarak debisini %200-600 oranında arttırabilir. Kalbin yedek kapasitesinin aşılması veya artan debi ihtiyacını karşılayamaması durumunda KKY görülür (Jovic ve Bogunovic, 2011). Bu çalışmada http://www.physionet.org internet

adresinden ücretsiz olarak erişilebilen ve belki de en çok kullanılan Physionet veritabanlarından (Goldberger vd., 2000) elde edilen veriler kullanılmıştır. WEKA'da hazır sunulan kümeleme algoritmalarından k-ortalama algoritması (Duda vd., 2000) kullanılarak bu veriler üzerinden çalışma yürütülmüştür. Takip eden bölümde WEKA yazılımı hakkında bilgi verilmiştir. Üçüncü bölümde yazılım kullanılarak çalışmanın gerçeklenmesi anlatılmış ve son bölümde ulaşılan sonuçlar verilerek tartışılmıştır.

WEKA

WEKA, Yeni Zellanda'daki Waikato Üniversitesi tarafından geliştirilmiş, makine öğrenimi algoritmalarının bir arada barındıran, işlevsel bir grafik arabirimine sahip, açık kaynak kodlu bir veri madenciliği programıdır (Witten vd., 2011). WEKA çeşitli veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları ve görselleştirme araçları içerir. Algoritmalar veri kümesine doğrudan veya Jawa kodundan çağrılarak uygulanabilir (Patterson vd., 2008; Hall vd., 2009). Aynı zamanda yeni makine öğrenme algoritmaları geliştirmek için de uygundur.

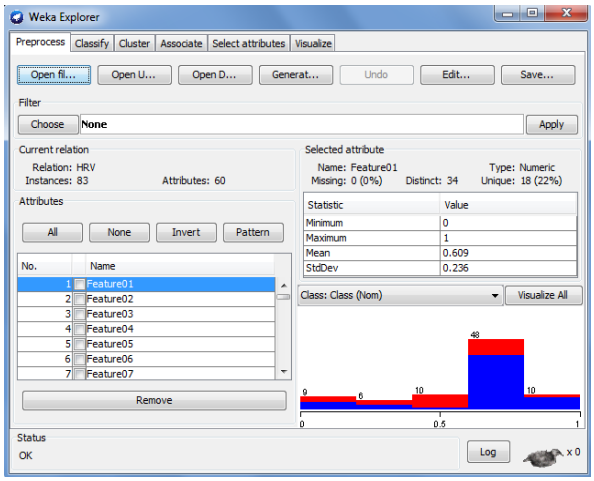
Program çalıştırıldığında Şekil 1'deki kullanıcı ara yüzü ekrana gelir. Bu ara yüz ekranında "Program", "Visualization", "Tools" ve "Help" menülerinden oluşan ana menü ve "Explorer", "Experimenter", "Knowledge Flow" ve "Simple CLI" kısımlarından oluşan "Applications" bölümleri bulunmaktadır.

Şekil 1. WEKA kullanıcı ara yüzü

"Applications" bölümünde yer alan "Explorer" seçeneği mevcut veri üzerinde yapılabilecek uygulamaları içeren genel bir grafiksel kullanıcı ara yüzünü içerir. "Exprimenter" seçeneği ise bir veya daha çok veri kümesi üzerinde bir veya daha çok algoritmanın uygulanabilmesine ve gözlemlenebilmesine 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. Takip eden alt başlıklar bu seçeneklerin tanıtılmasına ayrılmıştır. Ayrıca WEKA programına, kullanım kitabına ve dokümanlarına http://www.cs.waikato.ac.nz/ml/weka internet adresinden ulaşılabilir (Witten vd., 2011).

Explorer

Temel grafiksel kullanıcı ara yüzüdür. Şekil 2'de görüldüğü gibi "Explorer" ekranı önişlem (preprocess), sınıflama (classify), kümeleme (cluster), ilişkilendirme (associate), öznitelik seçimi (select attributes) ve görselleştirme (visualize) bölümlerinden oluşur. Bu ekran ilk açıldığında, üst kısımda bulunan komutlar veriyi yükleyene kadar pasif durumdadır. Veri yüklendikten sonra üst kısımda yer alan bu komutlar aktif hale gelecektir. Ara yüzün en alt kısmında bulunan "Status" bölümünde yapılmakta olan ve o ana kadar gerçekleştirilmiş işlemler hakkında bilgi verilir. Sağ alt köşedeki WEKA kuşu hareket halinde ise o an işlem yapıldığını gösterir ve çarpı işaretinden sonraki sayı ise kaç tane işlemin o ana kadar tamamlandığını gösterir.

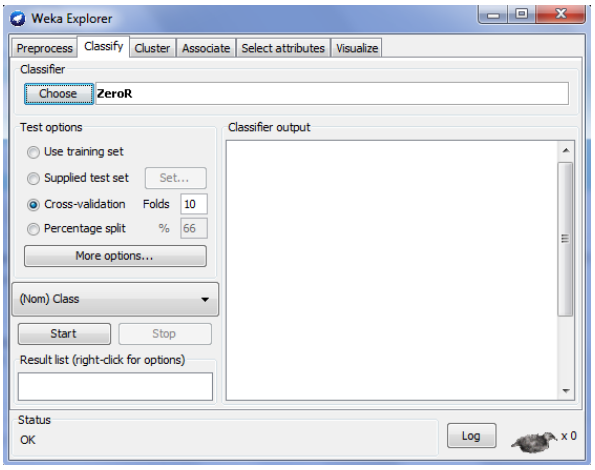


Şekil 2. Explorer kullanıcı ara yüzü

"Preprocess" sekmesinde çalışılacak verinin yüklenmesi, çeşitli filtreleme işlemleri ile verinin istenilen şekle sokulması, kullanılacak değişkenlerin seçilmesi ve basit istatistiklerin okunabilmesi gibi seçenekler sunulmaktadır. Öncelikle, verilerin diskten yüklenmesini sağlayan "Open file", herhangi bir Internet veri tabanından alınmasını sağlayan "Open URL", ağdaki bir veri tabanı sunucusundan alınmasını sağlayan "Open DB" ve verilerin baştan elle girilebilmesini sağlayan "Generate" seçenekleri ile üzerinde çalışılacak veri yüklenmelidir. Programın desteklediği veri biçimleri ARFF, CSV, LibSVM ve C4.5'dir. Veri üzerinde yapılabilecek işlemler arasında verinin nominal değerlere dönüşümü, ayrık değerlerden ayıklanması, yeniden örnekleme ve dalgacık analizi başta olmak üzere birçok seçenek sayılabilir. "Filter" işlemi için "Choose" komutu

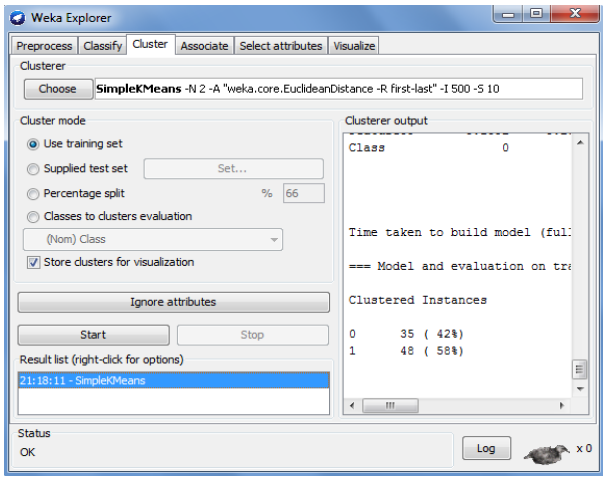
aracılığı ile veri üzerinde gerçekleştirilecek işlem seçilir. Seçilen işlem hakkında daha detaylı bilgi almak ve filtrenin ayarlarını değiştirmek için seçilen filtrenin üzerine tıklanır. Ekrana gelen "GenericObjectEditor" diyalog kutusu kullanılarak filtre ile ilgili gerekli bilgiler alınabilir ve ayarlarında değişiklikler yapılabilir. Üstelik bu ekranın sağ tarafında verinin ismini, sahip olduğu örnek sayısını, öz nitelik (değişken) sayısını, minimum değerini, maksimum değerini, standart sapmasını ve dağılımını gösteren kısımlar bulunmaktadır.

"Classify" sekmesinde çeşitli sınıflandırıcı algoritmalarının gerçeklenmesiyle ilgili seçenekleri gösteren bir kullanıcı ara yüzü ekrana gelir. Şekil 3'deki bu ekranda öncelikle "Classifier" başlığı altındaki "BayesNet", "SMO", "LibSVM", "IBk", "AdaBoostM1", "OneR", "ZeroR" ve "J48" gibi birçok sınıflandırıcı algoritmasından birisi seçilir. Bu ekranda "Test Options" başlığında eğitim setinin tamamının (Use training set), ayrı bir veri setinin (Supplied test set), parçalara ayrılarak birinin (Cross-validation) ve belli bir oranının (Percentage split) test amaçlı kullanılmasına olanak sağlayan seçenekler bulunmaktadır. Sınıflandırıcının veri üzerindeki performansının raporlanması için "Result list" ekranında genel olarak sınıflandırıcıya verilen verinin adı, içerdiği örnek sayısı, öznitelik sayısı, doğru ve yanlış şekilde sınıflandırılan örnek sayısı, doğruluk tablosu, ortalama mutlak hata gibi birçok sonuç gösterilmektedir. Sınıflandırıcı çıkışı (Classifier output) bölümünde çalışma bilgileri, detaylı doğruluk tablosu, hata matrisi ve sınıflandırma ile ilgili özet bilgiler verilmektedir.



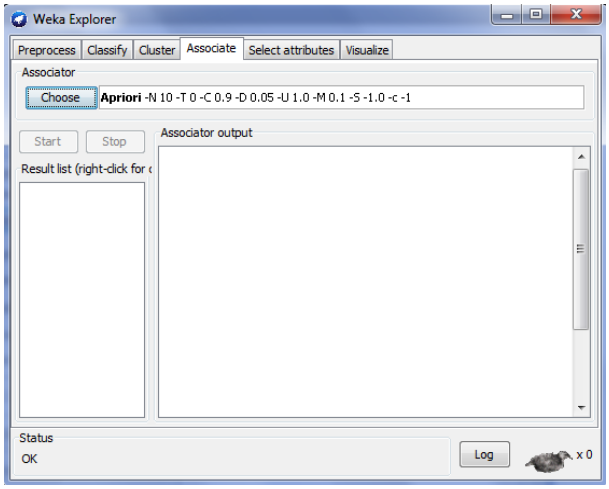
Şekil 3. Classify kullanıcı ara yüzü

"Cluster" sekmesinde ise kümeleme algoritmaları ile ilgili seçenekleri gösteren kullanıcı ara yüzü ekrana gelir. Şekil 4'deki bu ekranda öncelikle "Clusterer" başlığı altındaki "EM", "SimpleKMeans", "OPTICS" ve "HierarchicalClusterer" gibi kümeleme algoritmalarından birisi seçilir. Kümeleme algoritması seçilirken sınıflandırıcı sekmesinde olduğu gibi çeşitli ayar parametreleri seçilebilir ve algoritma hakkında detaylı bilgi alınabilir. Bu ekranda "Clusterer mode" başlığında eğitim setinin tamamının (Use training set), ayrı bir veri setinin (Supplied test set), eğitim setinin belli bir oranının (Percentage split) veya sınıf bilgisinin kümelerin değerlendirilmesinde (Classes to clusters evaluation) test amaçlı kullanılmasına olanak sağlayan seçenekler bulunmaktadır. Ayrıca bazı özniteliklerin göz ardı edilmesi için seçilmesine (Ignore attributes) ve böylece kalan özniteliklerin başarımı ne kadar etkilediğinin gözlenebilmesi de sağlanmaktadır.



Şekil 4. Cluster kullanıcı ara yüzü

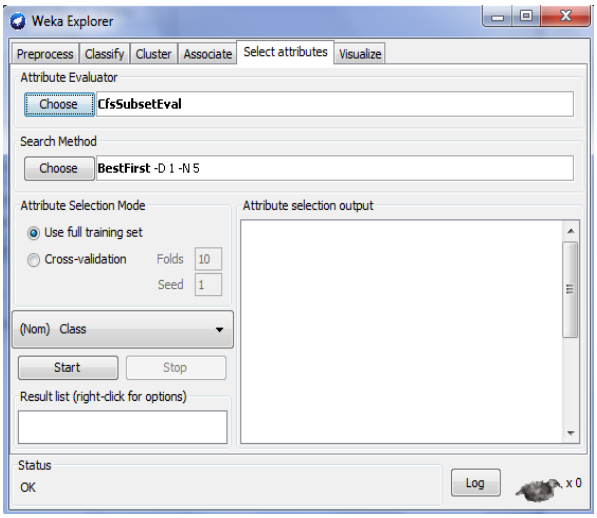
"Associate" sekmesinde ilişkilendirme kurallarının uygulanmasıyla ilgili komutların kullanılabildiği bir kullanıcı ara yüzü ekrana gelir. Şekil 5'deki bu ekranda "Associator" komutu ile ilişkilendirici seçimi ve seçilen ilişkilendiriciye göre çeşitli ayarların yapılabilmesi sağlanır. "Associator output" ve "Result list" bölümlerinde çalıştırılan algoritmaya ait sonuçlar görüntülenmektedir.



Şekil 5. Associate kullanıcı ara yüzü

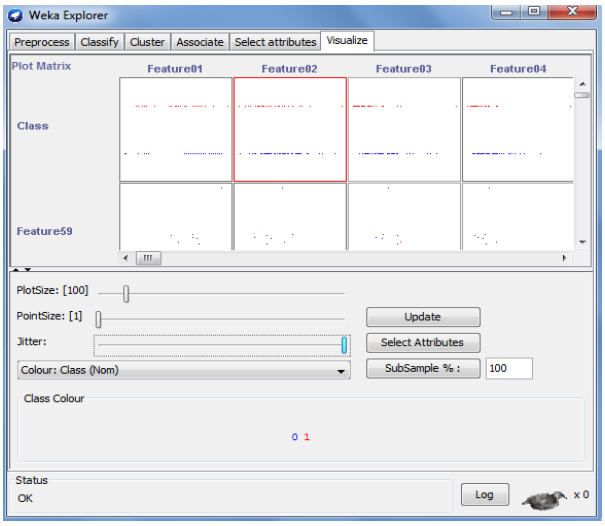
"Select Attributes" sekmesinde öznitelik seçimi yapılabilmektedir. Şekil 6'daki bu ekranda öncelikle "Attribute Evaluator" ve "Search Method" yöntemleri

seçilmelidir. Bunların ilkinde özniteliklerin her bir alt kümesinin uygunluk derecesinin belirlenmesi için kullanılacak değerlendirici yöntemini ve diğeri özniteliklerin içerisinde nasıl arama yapılacağını belirler. "Attribute selection mode" başlığı altında özniteliklerin değerlendirilmesinde eğitim veri kümesinin tamamının (Use full training set) veya çapraz doğrulama (Cross-validation) yöntemi ile seçilenlerin kullanılacağı belirlenir. Sonuç olarak seçilen özniteliklerin listesi "Attribute selection output" başlığı altında görüntülenir.



Şekil 6. Select attributes kullanıcı ara yüzü

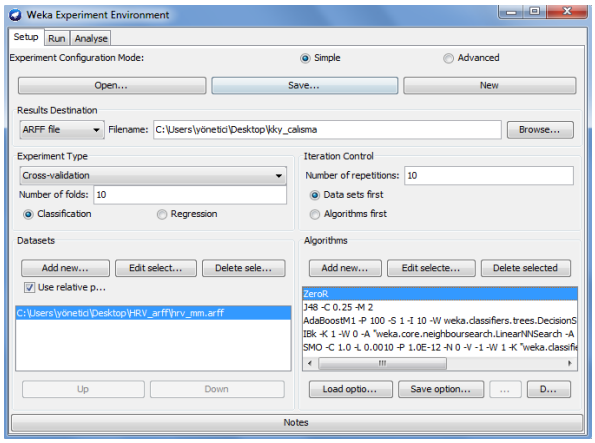
"Visualize" seçeneğinde veri setindeki örneklerin özniteliklere göre nasıl dağıldığı Şekil 7'de gösterildiği gibi iki boyutlu şekiller halinde görüntülenebilir. Bu ekranda isteğe bağlı olarak öznitelik seçimi yapılarak görsel inceleme yapılması da sağlanmaktadır.



Şekil 7. Visualize kullanıcı ara yüzü

Experimenter

Veri madenciliği uygulamalarında, sıklıkla geliştirilen bir algoritmanın birden fazla veri seti üzerindeki başarımlarının gösterilmesi veya birden fazla algoritmanın aynı veri seti üzerindeki başarımlarının kıyaslanması şeklindeki çalışmalara rastlanmaktadır. Bazen her iki durum aynı çalışma içerisinde görülebilmektedir. WEKA yazılımı bu tür çalışmalara destek verebilecek yapıdadır. Böylece tasarlanan çalışmaların daha etkin, daha hızlı ve daha kolay bir şekilde yürütülebilmesi sağlanır. Bu tür çalışmalar WEKA yazılımının "Experimenter" seçeneği ile gerçekleştirilebilir (Şekil 8). Yapılan çalışmalar tekrar kullanılmak üzere diske kaydedilebilir ve yapılandırılmış veya kaydedilmiş bu deneyler komut satırından da çalıştırılabilir. Bu ekranda yapılacak olan deneylerde çalışılması istenen veri tabanı ve çalışılması istenen algoritmalar seçilebildiği "Setup", deneylerin yürütülebildiği "Run" ve yürütülen çalışmanın sonuçlarının incelenebildiği "Analyse" sekmeleri mevcuttur. Bu ekranda sadece sınıflandırma problemleri üzerinden deneyler yürütülebilmekte olup kümeleme çalışmaları için uygun değildir.



Şekil 8. Experimenter kullanıcı ara yüzü

Knowledge Flow

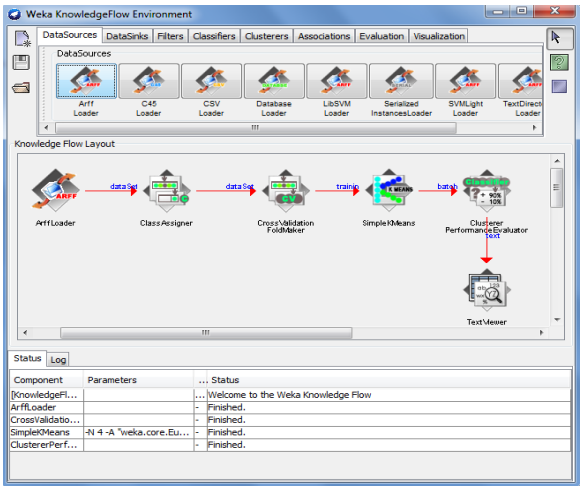
"Explorer"da gerçekleştirilebilen çalışmaların görsel yolla sürükle-bırak mantığıyla tasarlanabilmesi için geliştirilen alternatif çalışma ortamına "Knowledge Flow" denilmektedir (Şekil 9). "Explorer" ekranında gerçekleştirilemeyen bazı ekstra özellikler de bu ekranda mümkün hale gelmektedir. Örneğin "Explorer" verileri toplu olarak işlerken "Knowledge Flow" verileri kademeli ya da toplu olarak işleme özelliğine sahiptir. Bu durum "Explorer" için büyük veriler için sıkıntı oluşturmaktadır. Bu ekranda "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Evaluation" ve "Visualization" sekmeleri mevcuttur. Yapılacak çalışmaya uygun ara işlemler sürükle-bırak yöntemiyle "Knowledge Flow Layout " ekranına bırakılıp işlemler tamamlandıktan sonra "Arff Loader" üzerinde sağ kliklenerek çalışma yürütülür.

Simple CLI

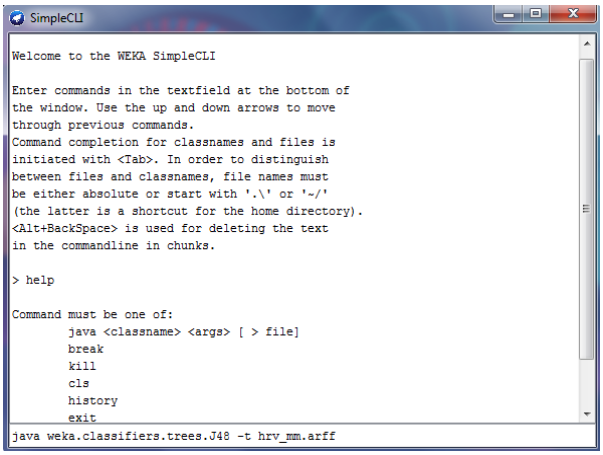
Çalışmaları komutsal olarak yapmaya olanak sağlayan kısımdır (Şekil 10). Örneğin;

```
java weka.classifiers.rules.ZeroR -t weather.arff
java weka.classifiers.trees.J48 -t weather.arff
```

şeklinde yazılarak çalışmalar yapılır.



Şekil 9. Knowledge Flow kullanıcı ara yüzü



Şekil 10. SimpleCLI ekranı

Veri

Bu çalışmada konjestif kalp yetmezliği hastaların normal guruptan ayırt edilmesini sağlayacak WEKA'da bir çalışma yapılacaktır. Bu çalışmada kullanılan veriler http://www.physionet.org Internet adresinden ücretsiz olarak erişilebilen ve biyomedikal işaret işleme alanında çalışan herkesin kullanımına açık olan Physionet veri tabanlarından (Goldberger vd., 2000) 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

Kalp Hızı Değişkenliği Analizi

KHD analizi için NN olarak adlandırılan normal-normal aralıklarının, başka bir deyişle EKG verileri içinde sadece sinüs düğümünün depolarizasyonu sonucu ortaya çıkan QRS yapıları arası ardışık zaman aralıklarının analizi yapılmaktadır (Task Force, 1996). 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. Ayrıntılı bilgi için kaynakçada verilen referanslara bakılabilir (İşler ve Kuntalp, 2007; İşler vd., 2008).

Frekans alanı ölçümleri için genellikle hızlı Fourier dönüşümü (FFT) yöntemini kullanan Welch periyodogram yöntemi kullanılmaktadır (İşler ve Kuntalp, 2007):

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

burada N zaman dizisi verisinin uzunluğudur.

Bu yöntem kullanılarak sadece zamanda eşit aralıklarla örneklenmiş veriler üzerinden güç spektral yoğunluğu (GSY) hesaplanabilir. Bu nedenle elde edilen KHD verisinin 4 Hz örnekleme hızında kübik interpolasyon metodu ile yeniden örneklenmesi (Task Force, 1996) ve analizde durağanlığı sağlamak için eğilim yok etmenin kullanılması gerekmektedir (İşler vd., 2005).

Frekans alanı 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) (Task Force, 1996). Bu araştırmada, klasik GSY yöntemlerine alternatif olarak geliştirilen Lomb periyodogram yöntemi ile elde edilen frekans alanı

ölçümleri de kullanılmaktadır. Bu yöntem ile zaman alanında yeniden örneklemeye ihtiyaç duyulmadan, doğrudan KHD verileri üzerinden, GSY hesaplanabilmektedir (Lomb, 1976):

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

burada

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

ve P(w) hesaplamasını tüm t_j örnekleme zamanlarından bağımsız hale getiren ortalama bir ofset değeridir. Bu yöntemle yeniden örnekleme ve eğilim yok etmenin olumsuz yanlarından etkilenmeksizin frekans alanı ölçümleri elde edilebilmektedir (İşler vd., 2005).

Bu araştırmada kullanılan ve 5 dakikalık zaman aralığı için tanımlanan standart frekans alanı KHD ölçümleri Çizelge 1'de listelenmiştir. Bu ölçümler İşler ve Kuntalp (2007) çalışmasında ayrıntılı olarak tanımlanmış ve standart olarak kullanımları Task Force (1996) tarafından önerilmiştir. 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 tanıdığı gibi yapısından dolayı polinomsal durağansızlıkları da ortadan kaldırır (Quian vd., 2001). Bu yönüyle, dalgacıkların özellikle RR aralıklarının analizinde çok kullanışlı olduğu rapor edilmiştir (Wiklund, 1997; İşler ve Kuntalp, 2007). 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 gücü)

İşaretin dalgacık dönüşümünün ayrıntı katsayılarına uygulanmasıyla hesaplanan 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:

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

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

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

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.

k-Ortalama Kümeleme Algoritması

k ortalama algoritması adından da anlaşılabileceği gibi giriş uzayını k adet merkezle ifade etmeye çalışan bir yöntemdir (Duda vd., 2000; Martis ve Chakraborty, 2011; Kıranyaz vd., 2011). Merkezlere ilk değer ataması rasgele 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 rasgele sırada da olabilir (Duda vd., 2000). Bu çalışmada birinci yaklaşım tercih edilmiştir.

Değerlendirme

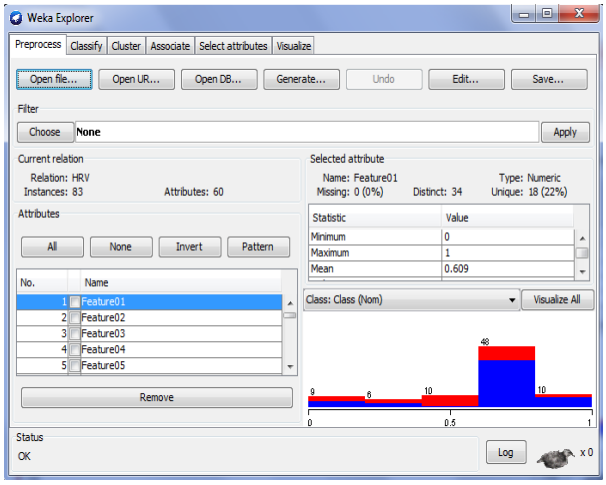
Algoritmaların başarımlarının değerlendirilmesinde kullanılan genel başarım ölçütü şu şekilde verilmektedir (Duda vd., 2000):

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

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.

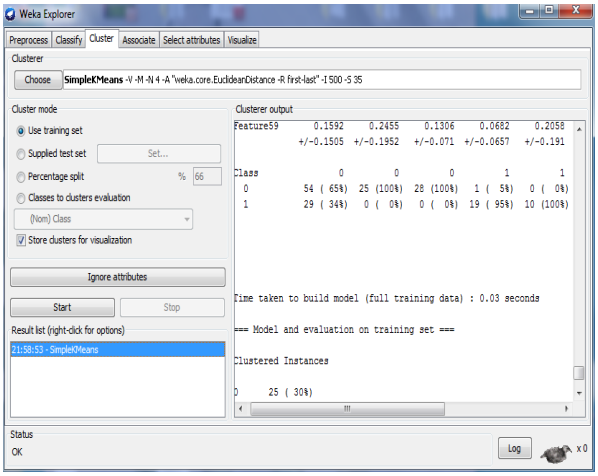
Sonuçlar

Bu çalışmada elde edilen 59 adet öznitelikten ve kayıt alınan kişinin hasta olup olmadığı bilgisinden oluşan öznitelikler diske kayıt edildi. Daha sonra MATLAB ile geliştirilen bir program aracılığıyla, 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 11'de gösterilmiştir.



Şekil 11. 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 (örnekte Feature1) 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 12'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.



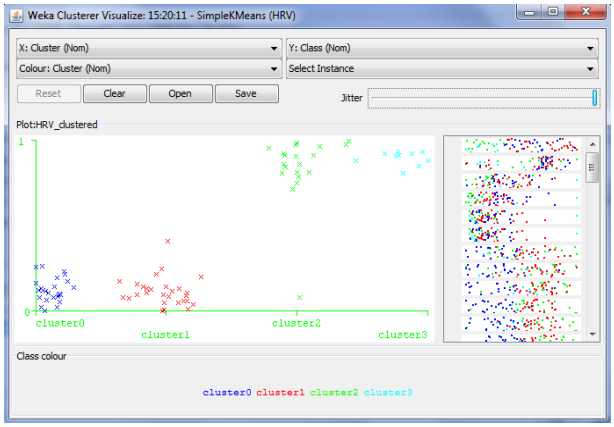
Şekil 12. K-ortalama kümeleme algoritması ile kümeleme

- Çalışmada küme sayısını belirleyen k değeri 2'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 13). 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



Şekil 13. En yüksek başarımın elde edildiği 4 adet küme için görselleştirme ekranı

Tartışma

Günümüzde inanılmaz oranda artan veri miktarından dolayı istenen bilgiye kolayca ulaşabilme ve veriden anlamlı sonuçlar elde edebilme zorlaşmıştır. Bunun daha kolay gerçekleştirilmesi için yazılımlardan faydalanılmaktadır. Bu çalışmada, bu yazılımlardan biri olan WEKA yazılımı tanıtılmış ve konjestif kalp yetmezliği hastalarının teşhisi uygulamasında kullanılmıştır. Diğer yazılımlara göre kolay elde edilebilir ve açık kaynak kodlu olması sebebiyle kullanıcı sayısı fazladır. WEKA, içerisinde bulundurduğu birçok algoritma sayesinde verilerden bilgi çıkarımı yapabilmesine olanak sağlamaktadır.

Çalışmada elde edilen sonuçlara göre KKY hastalığının teşhisinde k ortalama kümeleyicisinin oldukça başarılı olduğu görülmüştür. Bununla birlikte k-ortalama algoritması kullanılırken başlangıç merkez değerlerinin algoritma başarımı etkilediği bilindiğinden, farklı başlangıç değerleri için algoritma tekrarlanmış ve en yüksek başarımların verildiği sonuç tablosunda bu değerler de yazılmıştır. WEKA yazılımı kullanılarak gerçekleştirilecek diğer çalışmalarda da farklı başlangıç değerlerinin denenmesi gereklidir.

Çalışmada öznitelik seçimi yapılmamıştır. Öznitelik seçimi yapılan diğer algoritmaların da eklenmesi ve test kümesinin eğitim kümesi dışından seçimine olanak sağlayan diğer algoritmaların da uygulanmasıyla daha güvenilir sonuçların elde edilebileceği ön görülmektedir.

Kaynaklar

[1] Coşkun, C., Baykal, A. 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim, Malatya.

[2] Danacı, M., Çelik, M., Akaya, A.E. 2010. Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi. Akıllı sistemlerde Yenilikler ve Uygulamaları Sempozyumu (ASYU'2010), 21-24 Haziran, Kayseri, 9–12.

[3] Dener, M., Dörterler, M., Orman, A. 2009. Açık Kaynak Kodlu Veri Madenciliği Programları: Weka'da Örnek Uygulama. Akademik Bilişim, 11-13 Şubat, Şanlıurfa.

[4] Duda, R.O., Hart, P.E., Stork, D.G. 2000. Pattern Classification. 2. Baskı, Wiley, New York.

[5] Goldberger, A.L., Amaral, L.A.N., vd. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation, 101(23), e215–e220.

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. 2009. The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18.

[7] 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.

[8] İşler, Y., Avcu, N., Kocaoğlu, A., Kuntalp, M. 2008. Kalp Hızı Değişkenliği Analizi Frekans Alanı Ölçümlerinde Kullanılan Yöntemlerin Etkilerinin Araştırılması. IEEE 16th Signal Processing and Communications Applications Conference (SIU2008), 20-22 Nisan, Didim / Aydin.

[9] 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.

[10] İşler, Y., Selver, M.A., Kuntalp, M. 2005. Kalp Hızı Değişkenliği Analizinde Eğilim Yok Etmenin Etkileri. II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi (MBGAK'2005) , 17-19 Ekim, İstanbul, 213–219.

[11] 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.

[12]Kıranyaz, S., İnce, T., Pulkkinen, J., Gabbouj, M. 2011. Personalized long-term ECG classification: A systematic approach. Expert Systems with Applications, 38, 3220–3226.

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

[14] Lomb, N.R. 1976. Least-squares frequency analysis of unequally spaced data. Astrophysical and Space Science, 39, 447–462.

[15] Martis, R.J., Chakraborty, C. 2011. Arrhythmia Disease Diagnosis using Neural Network, SVM, and Genetic Algorithm-Optimized k-Means Clustering. Journal of Mechanics in Medicine and Biology, 11(4), 897–915.

[16] Özçift, A. 2011. Biyomedikal verilerin akıllı sistemler ile sınıflandırma başarımlarının analizi. Yayınlanmamış doktora tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü.

[17] Patterson, D., Liu, F., Turner, D., Concepcion, A., Lynch, R. 2008. Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security, Mart, Orlando, FL.

[18] 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.

[19] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. 1996. Circulation, 93, 1043.

[20] Wiklund, U., Akay, M., Niklasson, U. 1997. Short-Term Analysis of Heart-Rate Variability by Adapted Wavelet Transforms. IEEE Engineering and Medicine in Biology, 113–118.

[21] Witten I.H., Frank E., Hall MA. 2011. Data mining: practical machine learning tools and techniques. Elsevier, London.