Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması

Cengiz COŞKUN[1], Yrd. Doç. Dr. Abdullah BAYKAL[2]

[1] Dicle Üniversitesi, Bilgi İşlem Daire Başkanlığı, Diyarbakır

[2] Dicle Üniversitesi, Fen Fakültesi Matematik Bölümü, Diyarbakır

ccoskun@dicle.edu.tr, baykal@dicle.edu.tr

Özet: Veri madenciliği yöntemleri ile eldeki veriler sınıflandırılarak, gruplandırılarak ya da veriler arasında ilişkiler, bağıntılar, istatistiksel sonuçlar oluşturularak modeller oluşturulur. Oluşturulan model, oluşturulduğu veri kümesinde olmayan yeni bir kayıt geldiğinde, yeni gelen kayıt hakkında tahminleme yapma imkanı verir. Yapılan tahminlerin doğruluk derecesi oluşturulmuş olan modelin veri üzerindeki başarımını ortaya koyar. Dolayısı ile bir veri madenciliği uygulamasında hangi algoritma ile daha iyi sonuçlar üretildiği uygulamanın başarımı açısından önemlidir. Ayrıca sürekli geliştirilmekte olan yeni algoritmaların başarım derecesinin var olan algoritma sonuçları ile karşılaştırılması yeni geliştirilen algoritmanın kabul edilebilirliğini ortaya koyması açısından önemlidir. Bu çalışmada WEKA (Waikato Environment for Knowledge Analysis) programı ile SEER (Surveillance Epidemiology and End Results) veri kaynağı üzerinde çeşitli sınıflandırma metotlarına ait algoritmalar kullanılarak modeller oluşturulmuş, elde edilen modellerin tahminleme başarımları karşılaştırılarak kullanılan veri kaynağında hangi algoritmanın daha başarılı modeller oluşturduğu irdelenmiştir.

Anahtar Sözcükler: Veri Madenciliği Uygulaması, Veri Madenciliği Sınıflandırma Algoritmalarının karşılaştırılması, Veri madenciliğinde Model Başarım Ölçütleri (doğruluk, kesinlik, duyarlılık ve f-ölçütü)

An Application for Comparison of Data Mining Classification Algorithms

Abstract: Technologic advances in the computer world and the rapidly spreading usage of computers in different fields has led us to deal with a great amount of data, and in turn, made it difficult to analyze and produce beneficial and meaningful information out of those bulk data. As a prediction, it is supposed that total amount of data in the world doubles every 20 months. Within recent decades, data mining methodologies has emerged based on the fundamentals of statistics, AI and machine learning which focuses on analyzing, interpreting and producing information from data. A large variety of algorithms exist to construct models. In a data mining application it is important to obtain the model that gives more accurate results. Although comparison of data mining algorithms faces some critics in the academic literature, it is beneficial and inevitable. Detecting which algorithm produces the best explanatory model is an important issue while working on a data mining application for a given data. Also comparing algorithms plays an important role for determining if a newly developed algorithm is acceptable. In this study, different models were constructed for several classification algorithms using WEKA (Waikato Environment for Knowledge Analysis) program over SEER (Surveillance Epidemiology and End Results) data, and the algorithms were compared using the accuracy, precision, recall and f-measure values that their models resulted in.

Keywords: Data Mining Application, Comparison of several Data Mining Classification Algorithms, Metrics for Data Mining Model Comparison, accuracy, precision, recall, f-measure

1. Giriş

Gelişen bilgisayar teknolojileri ile birlikte her geçen gün kullanılmakta olan veri miktarı da hızla büyümekte; hızla artan veri miktarı da bu verilerin analiz edilmesini zorlaştırmaktadır. Bir tahmine göre her 20 ayda bir dünyadaki veri miktarının ikiye katlandığı ifade edilmektedir[2]. Son on yıllarda verileri analiz ederek yararlı ve anlamlı bilgilere ulaşmak için kaynağını istatistik, yapay zeka ve makine öğrenmeden alan veri madenciliği disiplini oluşturulmuş, endüstri, ekonomi, ve iş çevrelerinden gelen yoğun ilgi ile beraber akademik alanda da ilgi odağı olmuştur. Sağlık, temel bilimler, bankacılık, finans, pazar araştırmaları gibi pekçok alanda kullanılan veri madenciliği yöntemleri çok sayıda algoritma barındırır.

Veri madenciliğinde amaç, analiz edilerek bilgi çıkarılması zor olan büyük veri yığınlarını analiz ederek anlamlı, gizli ve faydalı olabilecek bilgi çıkarmak; bu bilgileri içerisinde barındıran bir model oluşturarak yeni gelecek bir veri nesnesi hakkında yorum yapmayı ve bu veri hakkında tahminde bulunmayı sağlamaktır. Üzerinde çalışılan veri kümesinden çıkarılan bilgi bir doğruluk derecesine sahip olup deterministik bir bilgi değildir. Oluşturulan modellerin başarım derecelerini belirleyen doğruluk, kesinlik, duyarlılık ve f-ölçütü gibi kriteler kullanılarak kullanılan algoritmaların başarıları değerlendirilir.

Veri madenciliği algoritmalarının empirik yollarla karşılaştırılarak değerlendirilmesi tüm bilimsel çalışmalarda yapılagelmiş olsa da akademik literatürde böyle çalışmaların objektif ve kesin sonuçlar üretmeyeceğine dair eleştiriler mevcuttur. Bu eleştiriler model oluşturuma basamaklarında uygulanan veri önişleme, parametre seçimi, test ve öğrenme kümelerinin oluşturulması gibi adımların uygulamayı yapan kullanıcının insiyatifinde olmasından dolayı oluşturulan model başarımlarının uygulamayı yapan kullanıcıya bağlı olmasından kaynaklanmaktadır. Bir başka eleştiri de yeni bir algoritmanın var olan eski algortmalarla kıyaslandığı akademik çalışmalarda yapılan uygulamalarda karşılaştırmaların geliştiricinin yanlı duruşu nedeni ile objektif bir sonuç vermeyeceği yönündedir[3]. Yapılan bir başka eleştiride ise akademik literatürde yapılmış olan karşılaştırma çalışmalarının çoğunda gerçek veriler kullanılmadığı, bu nedenle yapılan değerlendirmelerin doğru sonuç üretmemiş olduğu yönündedir[4]. Tüm bu eleştirilere rağmen algoritmaların karşılaştırılması gerekliliği ortak bir görüş olarak kabul edilmiş, gerek uygulama gerekse geliştirme anlamında yapılan akademik çalışmalarda ve güncel uygulamalarda yer edinmiştir.

Literatürde veri madenciliği algoritmalarının karşılaştırmasına yönelik farklı çalışmalar mevcuttur. Bunlardan bazıları var olan ya da yeni geliştirilen algoritmaların mevcut algoritmlarla karşılaştırılarak, geliştirilen algoritmanın kabul edilebilirliğine yönelik yapılan çalışmalar; bir kısmı da farklı veriler üzerinde farklı algoritmaların karşılaştırılarak değerlendirildiği çalışmalardır.

European Stat Logs projesi kapsamında farklı veri madenciliği sınıflandırma algoritmalarının karşılaştırıldığı ve sonradan "Machine Learning, Neural and Statistical Classification" adında kitap haline getirilen çalışmada endüstri ihtiyaçlarının hangi algoritma tarafından daha iyi karşılandığına dönük kapsamlı testler yapılmıştır. Çalışmada istatiksel, makine öğrenme ve yapay sinir ağları metotları farklı veri kümeleri üzerinde karşılaştırılmıştır. Çalışma sonucunda farklı algoritmaların farklı veri kümelerinde daha iyi modeller ürettiği vurgulanmıştır[5].

Diğer benzer karşılaştırma çalışmalarında farklı sonuçlara ulaşılmıştır.[6,7,8,9,10,11]

2. Sınıflandırma Algoritmalarının Karşılaştırılmasında Önemli Hususlar

Veri önişleme, parametre seçimi ve test kümesi seçimi veri madenciliği uygulamasında ortaya çıkacak olan modelin başarımını

etkiler. Dolayısı ile yapılan karşılaştırma sonuçları büyük ölçüde uygulamacıya bağlıdır.

2.1. Veri Önişleme

Algoritmaların karşılaştırılarak hangi algoritmanın daha iyi olduğunu bulmaya yönelik çalışmalara yapılan eleştirilerden biri uygulama sırasında yapılan veri önişleme basamağıdır. Bu adımda veri temizleme, veri birleştirme, veri dönüşümü, veri azaltma yöntemleri kullanılarak, veri analize hazır hale getirilir. Bu işlemler oluşacak modelin başarımını etkileyebilir. Yapılan işlemler uygulamacının bakış açısına bağlıdır. Veri kümesi üzerinde yapılan bazı farklı müdahaleler farklı algortmalarda farklı neticelere sebep olabilir. Yapılacak çalışmanın iyi sonuçlar üretmesi uygulamacının uygulama yapılan alan hakkında bilgili olmasını ya da bu alan uzmanlarıyla birlikte çalışmasını gerektirir.

2.2. Parametre Seçimi

Veri madenciliğinde kullanılan farklı algoritmaların farklı parametreleri olabilir. Örneğin yapay sinir ağlarında gizli nöron sayısı, karar ağaçlarındaki budama işleminin parametreleri, algoritmaların kullanacağı parametrik değerleri belirler. Bu parametreler algoritmadan algoritmaya değişebilir, ya da kullanılan veri madenciliği araç programlarında farklı olabilir. Bunların seçimi oluşacak olan modelin başarımını etkileyecektir.

2.3. Test Kümesinin Seçimi

Model oluşturulurken kullanılan öğrenme ve test kümelerinin belirlenmesinin de modelin başarımı üzerinde etkisi vardır. Eldeki verinin öğrenme kümesi ve test kümesi olarak ayrılmasında farklı metotlar kullanılabilir. Kullanılan veri madenciliği programında bu işlem için farklı seçenekler bulunabilir. Öğrenme kümesi ve test kümesi farklı dosyalardan programa verilebileceği gibi, programın bir veri dosyasını belirtilen bir oranda test kümesi olarak kullanması ya da n-fold metodu ile programın veri kümesini n sayıdaki parçalara ayırarak sırayla her parçayı test kümesi olarak kullanması sağlanabilir.

2.4. Model Başarım Ölçütleri

Model başarımını değerlendirirken kullanılan temel kavramlar hata oranı, kesinlik, duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle alakalıdır.

Test sonucunda ulaşılan sonuçların başarım bilgileri karışıklık matrisi ile ifade edilebilir. Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin tahminlemesini ifade eder.

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a	b
	Sınıf=0	c	d

a: TP (True Pozitif) c: FP (False Pozitif)
b: FN (False Negatif) d: TN (True Negatif)

2.4.1. Doğruluk – Hata oranı

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP+FN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır.

$$Doğruluk = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Hata\,Oranı = \frac{FP + FN}{TP + FP + FN + TN}$$

2.4.2. Kesinlik

Kesinlik, sınıfı 1 olarak tahminlenmiş True Pozitif örnek sayısının, sınıfı 1 olarak tahminlenmiş tüm örnek sayısına oranıdır

$$Kesinlik = \frac{TP}{TP + FP}$$

2.4.3. Duyarlılık

Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır.

$$Duyarlılık = \frac{TP}{TP + FN}$$

2.4.4. F-Ölçütü

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$F - Ölçütü = \frac{2 \times Duyarlilik \times Kesinlik}{Duyarlılık + Kesinlik}$$

3. Uygulama

Uygulamada göğüs kanseri hastalarının kayıtları incelenmiş, hastaların hayatta olup olmadıkları, hayatta değil iseler ne kadar süre hayatta kaldıkları ve ölüm sebepleri gözönünde tutularak herhangi bir hastanın hastalığı yenip yenemeyeceği sınıflandırılarak ileriye dönük tahminlerde bulunabilme amacı ile farklı algoritmalarla oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

Uygulamada bir karar ağacı algoritması olan ve temeli ID3 ve C4.5 algoritmalarına dayanan J48, istatistiksel bir algoritma olan Bayes sınıflandırma algoritmalarından Naive-Bayes, regresyon tabanlı algoritmalardan lojistik regresyon ve örnek tabanlı sınıflandırma algoritmalarından Kstar algoritmaları kullanılarak modeller oluşturulmuş ve oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

3.1. Kullanılan Veri Kümesi

Algoritmaların karşılaştırılması SEER (Surveillance Epidemiology and End Results) veri kaynağı kullanılarak yapılmıştır.

SEER veri tabanı farklı kanser gruplarını içeren ve bilimsel araştırmalarda son derece önemli bir yer tutan, güvenilir, dokümante edilmiş, eşine az rastlanır bir veri kümesidir. National Cancer Institute (NCI)'in sağladığı Amerika Birleşik Devletleri'nin belli başlı coğrafi bölgelerini kapsayan, nüfusunun %26'sını ilgilendiren ve bu kanser vakaları hakkında istatistiksel önem taşıyan bilgiler içerir. Yıllık olarak güncellenen bu veritabanı bilimsel çalışma yapanlara, sağlık sektöründe çalışanlara, halk sağlığı konusunda görevli kurumlara açık bir veri kaynağı olup, binlerce bilimsel çalışmada kaynak olarak kullanılmıştır. Veri kaynağı, kurumun web sitesinden veri kullanma talep formu doldurularak imzalandıktan sonra elektronik olarak indirilebilir.

1973 yılı itibarı ile başlanmış olan kanser verileri farklı yılları kapsayan, farklı tümör tiplerine göre gruplar altında metin formatında, 118 nitelikten oluşan, oldukça büyük veri kaynağıdır. Bazı nitelikler daha önceki yıl verilerinde yokken sonraki verilerde eklenmiş, bazı niteliklerin sonraki yıllarda değerleri alınmamış, bazı nitelikler farklı bir tümör tipinde değer taşırken bazı tiplerde bir anlam ifade etmediği için değer kullanımdan kaldırılmıştır. Her ne kadar bu veri kaynağı oldukça düzenli ve dokümante edilmiş olsa da yaptığım çalışma için bir önişlemden geçirilmesi gerekmiştir.

Bu çalışmada yıllık olarak güncellenen SEER veri kaynağının 2008 yılına ait olan versiyonu kullanılmıştır.

3.2. WEKA

Bu çalışmada Waikato Üniversitesinde java programlama diliyle geliştirilmiş olan Weka (Waikato Environment for Knowledge Analysis) programı kullanılmıştır. Weka, kullanımı ücretsiz, açık kaynak kodlu, içerisinde pek çok sınıflandırma, regresyon, demetleme, bağıntı kuralları, yapay sinir ağları algoritmaları ve önişleme metotları barındıran, yaygın kullanımlı bir veri madenciliği aracıdır.

WEKA, ham verinin işlenmesi, öğrenme metotlarının veri üzerinde istatistiksel olarak değerlendirilmesi, ham verinin ve ham veriden öğrenilerek çıkarılan modelin görsel olarak izlenmesi gibi veri madenciliğinin tüm basamaklarını destekler. Geniş bir öğrenme algoritmaları yelpazesine sahip olduğu gibi pek çok veri önişleme filtreleri içerir. Explorer, Experimenter, Knowledge Flow ve Simple CLI adı verilen 4 temel uygulamayı barındırır.

Desteklediği temel veri kaynakları, metin tabanlı arff, csv, c45, libsvm, svmlight, Xarff formatlarıdır; jdbc sürücüsü bulunan veritabanlarına direk bağlantı yapabilir ve internet üzerinden http protokolünü kullanarak bu formatlara uygun dosyaları okuyabilme yeteneğine sahiptir.

3.3. Veri üzerinde Uygulanan Ön İşlemler

SEER veri kaynağı, hasta kayıtlarını, niteliklerin değerleri arasında herhangi bir ayraç kullanılmadan, her bir kayıt bir satırda olacak şekilde farklı kanser tipleri için farklı metin dosyalarında tutmaktadır. Her satır 264 karakterden oluşmaktadır.

Çalışmada veri madenciliği algoritmalarının karşılaştırmasını yaparken 2000-2006 yılları arası Greater California, Kentucky, Louisiana, and New Jersey bölgelerine ait göğüs kanseri verileri kullanılmıştır.

3.3.1. Veri Formatının Ayarlanması

Veri madenciliği analizi için kullanılan Weka Programı csv, arff, c4.5 libsvm, xarff gibi formatları desteklemektedir. Bunun için C programlama dili ile veri dosyasını arff formatına dönüştürecek bir program yazılarak weka programının bu dosyayı okuması sağlanmıştır. Bu transfer yapılırken aynı zamanda veri önişleme -nitelik azaltma- işlemi uygulanmış, analizde kullanılmayacak nitelikler göz ardı edilerek yeni dosyada bu değerlere yer verilmemiştir. Bu alanlar belirlenirken konu ile ilgili uzmanlardan yaralanılmıştır.

3.3.2. Sınıfların Belirlenmesi

SEER veri kaynağında hastaların hastalığı yenip yenemediğini ifade eden bir alan olmamakla birlikte, bu bilgiyi çıkarabileceğimiz Survival Time Recode – STR-, Cause of Death Recode to Seer Site Recode -COD- ve Vital Status Recode – VTR-, nitelikleri mevcuttur. Delen ve ark.'ın kullandığı STR niteliğine[12] Bellaachia ve arkadaşlarının ek olarak kullandığı VTR ve COD niteliklerini[13] göz önüne alarak sınıflandırma yapılmıştır. Özetle, sınıfı belirlemek için;

```
if (VSR ?= '1') /* hayatta mı? - Evet */
  {if ( STR? > 60)
     Class= '1'; /* 60 aydan fazlaysa kurtuldu */
   }
else     /* hayatta değil */
  {   /* Ölüm Sebebi Kanser mi? */
        if (COD ?='Kanser')
        Class='2';
  }
```

algoritması kullanılmıştır. Bu kriterlerin dışında kalanlar ise göz ardı edilmiştir.

Veri kümesi arff formatına çevrildikten sonra Weka programı ile veri üzerinde katar tipli değişkenler nominale dönüştürülmüş, geniş aralıklı numerik değerler ise kısa aralıklara dönüştürülerek analizde daha anlamlı

sonuçlar elde edilmesi için düzenelemeler yapılmıştır.

3.4. Parametre Seçimi

Karşılaştırılacak algoritmaların parametreleri WEKA programının kendi varsayılan değerleri olacak şekilde bırakılmış, model başarımına olumlu ya da olumsuz etki edebilecek değişikliklerden kaçınılmıştır.

3.5. Test Kümesi Seçimi

Algoritmaları çalıştırırken test yöntemi olarak WEKA programının sağladığı "10-kat çapraz doğrulama" metodu kullanılmıştır. Bu yöntemle veri kaynağı 10 bölüme ayrılır ve her bölüm bir kez test kümesi, kalan diğer 9 bölüm öğrenme kümesi olarak kullanılır.

3.6. Oluşturulan Modellerin Başarım Ölçütleri

J48 Algoritması

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%86.36	%85.57	%96.53	%90.72

NaiveBayes Algoritması

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%85.21	%85.80	%94.15	%89.78

Lojistik Regresyon Algoritması

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%85.36	%86.30	%93.66	%89.83

KStar Algoritması

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%85.44	%83.70	%98.00	%90.28

4. Sonuç ve Öneriler

Çalışma sonuçları incelendiğinde J48 algoritmasının model testine ait %86.36 doğruluk derecesiyle en iyi sonucu ürettiği söylenebilir. Doğruluk ölçütü oldukça basit ve önemli bir kriterdir. Bu ölçüte göre J48 algoritmasını sırasıyla KStar, Lojistik Regresyon ve NaiveBayes algoritmaları izlemektedir.

Kesinlik ölçütü bakımından lojistik regresyon en iyi sonucu oluşturmuş olup, diğer algoritmalar bu ölçüte göre NaiveBayes, J48 ve KStar şeklinde sıralanabilir. Ancak kesinlik ölçütü tek başına yorumlanırsa değerlendirme yanlış sonuçlara götürebilir. Bu ölçütü duyarlılık ölçütüyle beraber ele almak gerekir. Tablodan görüleceği üzere algoritmalar, duyarlılık ölçütüne göre KStar, J48, NaiveBayes ve Lojistik Regresyon olarak sıralanabilir. Görüleceği üzere, kesinlik ölçütü ve duyarlılık ölçütü birbiriyle zıt bir sıralama ortaya koymuştur.

Kesinlik ve duyarlılık ölçütlerini beraber değerlendirmek için, her iki değerin harmonik ortalaması olan F-Ölçütüne baktığımızda sıralamanın J48, KStar, Lojistik Regresyon ve NaiveBayes şeklinde olduğunu görüyoruz. Bu sıralamanın yine doğruluk kriter sıralamasıyla birebir aynı olduğu görülmektedir.

Veri madenciliğinde bilgiye erişmede farklı metotlar kullanılmaktadır. Bu metotlara ait pek çok algoritma vardır. Bu algoritmalardan hangisinin daha üstün olduğu üzerine pek çok çalışma yapılmış, yapılan bu çalışmalarda farklı sonuçlar elde edilmiştir. Bunun en önemli sebebi, işlem başarımının, kullanılan veri kaynağına, veri üzerinde yapılan önişleme, algoritma parametrelerinin seçimine bağlı olmasıdır. Farklı kişiler tarafından, farklı veri kaynakları üzerinde, farklı parametrelerle yapılan çalışmalarda farklı sonuçlar oluşması doğaldır. Ancak, yaptığımız çalışma, "benzer veri kümelerinde belli yöntemlerin daha başarılı olduğu" şeklindeki çıkarıma[5] uygun olarak, diğer çalışmalarla[12,13] benzer sonuç vermiştir. Göğüs kanseri vakalarının farklı yıllarını içeren SEER veri kaynağının kullanıldığı her iki çalışmada da, bir karar ağacı algoritması olan C4.5 algoritmasının diğer

algoritmalardan daha iyi sonuç ürettiği sonucuna ulaşıldığı belirtilmiştir[12,13]. Bu çalışmada da, yapılan karşılaştırma sonucunda, C4.5 algoritmasının Weka implementasyonu olan J48 karar ağacı algoritması, benzer şekilde diğer algoritmalara göre daha başarılı bulunmuştur.

Bu çalışmada, modellerin oluşturulması için ücretsiz bir yazılım olan Weka aracı kullanılmıştır. Var olan diğer veri madenciliği araçları üzerinde aynı algoritmalar çalıştırılarak farklı araçların benzer sonuçlar üretip üretmediği kontrol edilebilir.

Algoritmaların veri kaynağı üzerinde çalıştırılması sırasında algoritma parametreleri olarak her algoritmanın o parametre için varsayılan değeri kullanılmıştır. Her algoritma ve her veri kaynağı için başarım derecesini maksimize edecek parametre değerleri tespit ederek bu parametrelerle algoritma sonuçlarını karşılaştırmak farklı sonuçlara götürebilecektir. Ancak, böyle bir karşılaştırmada yanlılık oluşabilecektir.

Bu çalışmada, algoritmaların ürettiği modellerin başarım sonuçları karşılaştırılmıştır. Benzer şekilde, algoritmaların hızı ve hafıza kullanımı da karşılaştırılarak algoritmaların performans karşılaştırması yapılabilir.

5. Kaynaklar

1. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Limited-Use Data (1973-2006), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2009, based on the November 2008 submission.

2. Witten, I. H. ; Frank, E. ; Data Mining, Practical Machine Learning Tools and Techniques ; Morgan Kaufmann , USA; ***2005****.*

3. Hand, D. J. ; Classifier Technology and the Illusion of Progress; Statistical Science, Vol. 21; Institute of Mathematical Statistics, ***2006****; 1-15.*

4. Salzberg, L. ; Methodological Note On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach ; Data Mining and Knowledge Discovery 1 ; Kluwer Academic Publishers, Boston ; ***1997****; 317–328.*

5. Michie, D. ; Spiegelhalter, D. J. ; Machine Learning, Neural and Statistical Classification; Taylor, C. C. ; Prentice Hall, ***1994****.*

6. Wilson, R. L. ; Sharda, R. ; *Bankruptcy prediction using neural networks*; Decision Support Systems, Vol. 11, Issue 5 ; Elsevier Science Publishers B. V. Amsterdam, The Netherlands; **1994**; 545-557.

7. Lin, M. ; Huang, S. ; Chang, Y. ; *Kernel-based discriminant technique for educational placement*; Journal of Educational and Behavioral Statistics, Vol 29; **2004**; 219-240.

8. King, R. D. ; Feng, C. ; Sutherland, A. ; *StatLog: Comparison of Classification Algorithms on Large Real-World Problems*; Applied Artificial Intelligence, Vol 9, Issue 3 ; **1995**; 289-333.

9. Sabzevari, H. ; Soleymani, M. ; Noorbakhsh, E. ; *A comparison between statistical and Data Mining methods for credit scoring in case of limited available data*; Eleventh Annual APRIA Conference; **2007**.

10. Rajavarman, V.N. ; Rajagopalan, S.P. ; *Comparison between Traditional data mining Techniques and Entropy-based Adaptive Genetic Algorithm for Learning Classification Rules*; International Journal of Soft Computing Vol 2 Issue 4; **2007**; 555-561.

11. Zurada, J. ; Lonial, S. ; Comparison Of The Performance Of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry; The Journal of Applied Business Research Vol 21; ***2005****; 37-53.*

12. Delen, D. ; Walker, G. ; Kadam, A. ; Predicting breast cancer survivability:a comparison of three data mining methods; Artificial Intelligence in Medicine, Vol 34, issue 2 ; ***2004****; 113-127.*

13. Bellaachia, A. ; Guven, E. ; Predicting breast cancer survivability: a comparison of three data mining method ; Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006) ; ***2006****.*

14. http://seer.cancer.gov/

15. http://en.wikipedia.org/wiki/Weka_(machine_learning*)*

16. http://www.cs.waikato.ac.nz/ml/weka/