VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Kural Tabanlı Sınıflandırma Algoritmalarının Karşılaştırılması

Serpil *SEVİMLİ DENİZ*[*]
Van Yüzüncü Yıl Üniversitesi, Gevaş MYO, Bilgisayar Teknolojileri Bölümü, VAN

Makale Bilgisi

Başvuru:
09/02/2021
Kabul:
18/08/2021

Özet

Sınıflandırma, bir veri işleme yöntemi ve bir gruptaki öğeleri hedef sınıfa atayan bir veri madenciliği tekniğidir. Veri analizinde kullanılan yöntemleri uygulayarak verileri sınıflandırma, bir dizi girdi verisi için sınıflandırma modelleri oluşturmak için kullanılan bir prosedürdür. Sınıflandırma algoritmaları sinir ağı temelli algoritmalar, kural tabanlı algoritmalar ve istatistik tabanlı algoritmalar olmak üzere ayırılır. Bu çalışma kural tabanlı sınıflandırma algoritmalarından; bulanık sırasız kural algoritması (FURIA), kısmi karar algoritması (PART), karar ağaçları ve hata azaltma için tekrarlanan artımlı budama (JRIP) algoritmalarını karşılaştırmak ve analiz etmek üzere hazırlanmıştır. Bu algoritmalar kullanılarak bir veri madenciliği aracı olan WEKA platformunda Wine, Soybean, Labor, Sensör, Hipotiroid, Diabet, Credi Card, Messidor_Futures veri setleri incelenmiştir. Belirtilen veri setleri ve algoritmalar sınıflandırılan örneklerin sayısı, doğru sınıflandırma yüzdeleri, ortalama mutlak hata (MAE) ve ortalama hatanın karekökü (RMSE) özellikleri referans alınarak karşılaştırılmıştır. Bu karşılaştırmada en iyi algoritmanın bulanık mantık tabanlı algoritma olan FURIA algoritması olduğu görülmektedir. Veri sayısı büyüdükçe FURIA algoritmasının daha verimli olduğu tespit edilmiştir.

Anahtar Kelimeler: *Sınıflandırma, FURIA, Karar Ağaçları, PART, JRIP*

Comparison of Rule-Based Classification Algorithms

Abstract

Classification is a data processing method and a data mining technique that assigns items in a group to the target class. Classifying data by applying the methods used in data analysis is a procedure used to create classification models for a set of input data. Classification algorithms are divided into neural network based algorithms, rule based algorithms and statistics based algorithms. This study is one of the rule-based classification algorithms; It is designed to compare and analyze the fuzzy unordered rule algorithm (FURIA), partial decision algorithm (PART), decision trees and repetitive incremental pruning algorithms for error reduction (JRIP). WEKA is a data mining tool using algorithms. The data sets of Wine, Soybean, Labor, Sensor, Hypothyroid, Diabet, Credi Card, Messidor_Futures were examined on the WEKA platform. The specified data sets and algorithms were compared with reference to the number of samples classified, correct classification percentages, mean absolute error (MAE) and root mean square error (RMSE) characteristics. In this comparison, it is seen that the best algorithm is FURIA algorithm which is fuzzy logic based algorithm. It has been determined that the larger the number of data, the more efficient the FURIA algorithm.

Keywords: *Classification, FURIA, Decision Tree, PART, JRIP*

[*] İletişim e-posta: sdeniz@yyu.edu.tr

1 Giriş

Sınıflandırmanın amacı, verilerdeki her durum için hedef sınıfı doğru bir şekilde oluşturmaktır [1]. Sınıflandırma yöntemi, yeni bir kaydın veya nesnenin daha önce belirlenmiş kurallar ile nasıl sınıflandırması gerektiğini belirler. Bunun için sınıflandırma kuralını oluşturan üç adımın takip edilmesi gerekir [2]. Bu adımlar şöyledir:

1. adım: Veri seti üzerinden rastgele seçilen bir kısım veri, eğitim verisi olarak belirlenir.

2. adım: Sınıflandırma için eğitim veri seti sonucunda elde edilen model kullanılır. Modelin tahmin gücü bulunan modele dayanarak belirlenir. Test veri seti de yine tüm veri seti içerisinden rastgele seçilir.

3. adım: Eğer modelin doğruluğu test veri seti üzerinde gerçekleştirilen denemeler neticesinde yeterli bulunursa, model yeni sonuçları bulmak için kullanılır.

Sınıflandırma yöntemleri çalışan araştırmacılar, en yakın komşu yöntemleri, karar ağaçları, hata geri yayılımı, yeniden kuvvetlendirmeli öğrenme, tembel öğrenme, kurala dayalı öğrenme, istatistiksel öğrenme gibi birçok farklı sınıflandırma algoritması türü önermişlerdir. Bu sürekli artan yöntemlerde hangi çalışmada hangi yöntem kullanılmalı? sorusu önemli bir sorudur. Bu çalışmada da yukarıdaki soruya uygun bir yanıt bulmak amaçlanarak kural tabanlı sınıflandırma yöntemlerini incelenmiştir. Verinin sınıflandırılması işlemi, daha önce sınıf etiketleri belli olan veri setini eğitim ve test olmak üzere iki parçaya bölme işlemi ile başlamaktadır. Buradaki eğitim verisi ile oluşturulacak model tasarlanmakta ve test verisi ile de oluşturulan model test edilmektedir. Oluşturulan model yardımı ile yeni bir örnek ile karşılaşıldığında örneğin hangi sınıfa ait etiketle etiketleneceği belirlenmektedir. Sınıflandırmanın amacı, benzer özelliklerdeki verilerin önceden etiketlenmiş veri gruplarından hangisine ait olduğunun tahmin edilmesi işlemidir [3].

1.1 Yapay Sinir Ağı Tabanlı Algoritmalar

1960'ların ortalarında Nilsson, yapay sinir ağları (ANN) adı verilen sinir benzeri eşik birimlerine dayanan örüntü tanıma için yapay zekayı tanıttı. Yapay sinir ağları, çok katmanlı algılayıcılar (MLP), radyal temel işlev ağları, Self Organization Map ve geri yayılımlı ağlar (BP) gibi bazı algoritmaların geliştirilmesinden sonra bir yaklaşım haline geldi. MLP mimarisi, tümü ileri beslemeli ağırlıklarla birbirine bağlanan giriş, gizli ve çıkış katmanları olmak üzere üç nöron katmanından oluşur. Bir giriş modelinin de yapay sinir ağı, çıktı katmanındaki çıkışı tahmin etmek için sinyali ağ üzerinden geçirir daha sonra tahmin edilen hedef değeri gerçek hedefle karşılaştırır ve ağırlıkları değiştirmek için hatayı tahmin eder. Ağırlıklar skaler hata fonksiyonu ile, her girdiye doğru yanıt verene kadar öğrenme prosedürünü tekrarlayarak en aza indirilir [4]. BP, bir gradyan iniş yöntemi kullanarak hata fonksiyonunu en aza indirir. BP algoritmasının temel dezavantajı, diğer bazı popüler makine öğrenme tekniklerinden daha yavaş olması ve hata fonksiyonunun yerel minimumlarında sıkışıp kalma eğiliminde olmasıdır [5].

1.2 Kural Tabanlı Algoritmalar

Kural tabanlı öğrenme, özellikle karar ağaçları böl ve yönet yaklaşımı veya yukarıdan aşağıya bir indüksiyon yöntemidir. C4.5, ID3'ün [5] gelişmiş sürüm karar ağacı algoritmasıdır. ID3, üçüncü 'etkileşimli ikiye bölünme' prosedürleri serisi anlamına gelir. Yalnızca nominal veri kümelerini sınıflandırabilir. Gerçek değer öznitelikleri için, sırasız nominal değerler oluşturmak için önce aralığa gruplanır. Herhangi bir standart budama prosedürünü dikkate almaz. ANN gibi C4.5 de üç ana adımda çalışır. İlk olarak, ağacın üst düğümündeki kök düğüm tüm örnekleri dikkate alır ve "dal düğümü" adı verilen ikinci düğümdeki örnek bilgilerinden geçer. Dal düğümü, entropi ölçümüne dayalı olarak bir grup örnek için kurallar oluşturur. Bu aşamada C4.5, tüm öznitelik değerlerini göz önünde bulundurarak çok büyük bir ağaç kurar ve budama yaparak karar kuralını sonlandırır. Bölmelerin istatistiksel önemine dayalı olarak budama için sezgisel bir yaklaşım kullanır. En iyi kuralı düzelttikten sonra, dal düğümleri "yaprak düğüm" [6] adı verilen son düğümdeki nihai hedef değeri gönderir. OneR, çok basit, daha hızlı ve tek seviyeli bir karar ağacı algoritmasıdır. Bir veri kümesinden tek tek nitelikleri seçer ve eğitim setindeki hata oranına bağlı olarak farklı bir kurallar dizisi oluşturur. Son olarak, minimum hatayla kurallar sunan özelliği seçer ve nihai karar ağacını oluşturur [7]. PART, C4.5 ve RIPPER algoritmalarının geliştirilmiş versiyonu olan kısmi bir karar ağacı algoritmasıdır. PART algoritmasının temel özelliği, uygun kuralları üretmek için C4.5 ve RIPPER gibi global optimizasyonu gerçekleştirmesine gerek olmamasıdır. Ancak, ağacın daha büyük boyutta

olmasından dolayı karar ağaçları bazen daha sorunludur ve sınıflandırma problemleri için kötü performans gösterebilir [8].

1.3 İstatistiksel Öğrenme Algoritmaları

İstatistiksel öğrenme teorisi, 1990'ların ortasında Vapnik ve grubu tarafından Destek Vektör Çıkarımı (SVM) ile ilgi gördü. SVM, 1960'ların sonlarında Rusya'da geliştirilen Genelleştirilmiş Portre algoritmasının gelişmiş sürümüdür. SVM, ANN ve C4.5'e benzer şekilde çalışır. SVM için üç çalışma aşaması atayabiliriz, birincisi giriş aşaması veya dönüşüm aşaması, sonra öğrenme aşaması ve sonuncusu karar aşamasıdır. ANN ve C4.5 ilk aşamada önemli bir çalışma yapmaz. Ancak SVM en önemli işini, verileri kullanarak yüksek boyutlu bir özellik uzayına çekirdek eşlemesi dönüştürür. Çekirdek polinom fonksiyonu, Gauss veya diğerleri olabilir. Doğrusal ayrımın mümkün olduğu yüksek boyutlu uzay teorik olarak sonsuz olabilir. SVM, yüksek boyutlu özellik uzayındaki verileri öğrenme aşamasında, ayırma ile sınırlandırılan ağırlık vektörünün büyüklüğünü, çarpan parametresi yardımıyla, örneğin Lagrange çarpanı yardımıyla sınırsız bir probleme indirgeyerek öğrenmeye başlar. Bu aşamada, SVM yalnızca destek vektörlerini çıkarır. Destek vektörleri bilgilerine dayalı olarak SVM, karar aşamasında nihai çıktı fonksiyonunu üretir. ANN ve C4.5'ten farklı olarak, SVM, nihai karar fonksiyonunu oluşturmak için tüm örnekleri dikkate almaz. Dahası, SVM, yinelemeli yaklaşımlardan veya budamadan farklı olarak her zaman karar fonksiyonu için benzersiz bir çözüm elde eder. SVM'nin diğer bir özelliği, çoğu klasik öğrenme algoritması tarafından ele alınan ampirik riskten ziyade yapısal riski en aza indirmesidir [9]. Naive Bayes (NB), klasik istatistiksel teori "Bayes teoremi" ne dayanan basit bir sınıflandırıcıdır. "Naif" terimi, eğitim örneklerindeki özniteliklerin bağımsız olduğu ve tahmin prosedürlerinde gizli veya örtük özniteliklerin etkisi olmadığı varsayımına dayanarak maksimum posterior olasılığı hesaplamasıdır [10]. Çekirdek yoğunluğu (KD), parametrik olmayan doğrusal çekirdek tabanlı yoğunluk tahmin algoritmasıdır. Bu algoritma, tahmin için niteliklerin normal dağılımı gibi herhangi bir ön varsayıma ihtiyaç duymaz. Bu algoritmanın ayırt etme yeteneği, diğer bazı sınıflandırıcılardan nispeten daha hızlıdır. IBK, K-en yakın-komşu yöntemi gibi örnek tabanlı bir öğrenme yaklaşımıdır. Bu algoritmanın temel ilkesi, her görünmeyen örneğin bir mesafe ölçütü kullanılarak her zaman mevcut olanlarla karşılaştırılmasıdır; en yaygın olarak Öklid mesafesi ve en yakın mevcut örnek, test örneğine sınıf atamak için kullanılır. İstatistiksel öğrenme algoritması, SVM, iyi kurulmuş algoritmalar ANN ve karar ağacına göre bazı avantajlara sahiptir. Optimal hiper düzlemi oluşturmak için kümeleme veya enterpolasyon yerine özellik vektörlerinin iç çarpımını dikkate alır. ANN veya karar ağacı gibi modelleme sırasında önemli bilgileri kaybetme olasılığı daha düşüktür [11].

2 Sınıflandırma Değerlendirme Ölçütleri

Sınıflandırma çalışmalarında en önemli kriter, yüksek başarımlı bir sınıflandırıcı model oluşturabilmektir. Ancak başarımı etkileyen birçok neden bulunmaktadır. Kullanılan test yöntemlerinin yanı sıra veri setine ait özellikler de başarımı etkileyen unsurlardan biridir. Sınıflandırma algoritmaları ile oluşturulan modellerin değerlendirilmesi, hangi sınıflandırma modelinin daha doğru sonuçlar ürettiğinin belirlenmesinde bazı değerlendirme metrikleri kullanılmaktadır. Bunlar genel olarak karışıklık matrisi olarak isimlendirilen bir tabloya dayanmaktadır. Makine öğrenmesi ve istatistiksel sınıflandırma problemlerinde, bir sınıflandırıcının performansını görselleştirmek için geliştirilen bir tablo düzenidir [12].

2.1 Karışıklık Matrisi

Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek değerleri, kolonlar ise modelin tahminini ifade etmektedir.

Tablo 2.1. Confusion Matrisi [13]

Confusion Matrix		HEDEF			
		Pozitif	Negatif		
Model	Pozitif	a	b	Pozitif tahmin	a/(a+b)
	Negatif	c	d	Negatif tahmin	d/(c+d)
		Duyarlılık $\frac{a}{a+b}$	Özgünlük $\frac{d}{b+d}$	Accuracy $\frac{a+d}{a+b+c+d}$	

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negatif)

$$TPR(sensitivity) = \frac{TP}{TP + FN} \quad (1)$$

$$TNR(specificity) = \frac{TN}{TN + FP} \quad (2)$$

$$p = \frac{a}{a+c}, \quad r = \frac{a}{a+d} \quad (3)$$

$$F = \frac{2rp}{r+p} \quad (4)$$

Bu denklemlerde Duyarlılık: Pozitif olarak etiketlenmiş örneklerin (TP) gerçekten pozitif olan örneklerin (TP+FN) toplam sayısına oranıdır. F-Ölçütü: Kesinlik ve duyarlılık metrikleri kullanılarak hesaplanmaktadır. Sistemin, kesinlik veya duyarlılık yönüne doğru optimize edilmesinde kullanılmaktadır. F daha küçük hassasiyete daha yakın olma eğilimindedir. 10-katlı çapraz geçerlilik (cross validation) testine göre gerçekleştirilen çalışmalarda veri seti 10 parçaya bölünür, sırası ile bu 10 parçanın her biri test seti, diğerleri eğitim seti olarak kullanılarak sınıflandırma işlemi gerçekleştirilir. İşlem sonunda 10 sınıflandırma işleminin sonuçları genel başarı olarak alınır [13].

2.1.1 Ortalama Mutlak Hata (MAE)

Ortalama mutlak hata (MAE) modelin performans değerini gösteren bir indistir. Bu değer ne kadar küçük olursa performans değeri o kadar iyi olur. Ortalama mutlak hata iki sürekli değişken arasındaki farkın ölçüsüdür. MAE, her gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey mesafedir. MAE aynı zamanda her veri noktası ile en iyi uyan çizgi arasındaki ortalama yatay mesafedir. MAE, yönlerini dikkate almadan bir dizi tahmindeki hataların ortalama büyüklüğünü ölçen, tüm tekil hataların ortalamada eşit olarak ağırlıklandırıldığı doğrusal bir skordur. MAE değeri 0'dan ∞'a kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir. Ortalama mutlak hata tahminlerin nihai sonuçlar ile ne kadar yakın olduğunu ölçümlemek için kullanılır ve aşağıda belirtilen formüle göre hesaplanmaktadır [14].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|e_j\right| \quad (5)$$

Ortalama Hatanın Karekökü (RMSE)

Ortalama kare hatanın kökü (RMSE) öngörü başarısını ölçmek için kullanılır. Hatanın büyüklüğünü ölçen kuadratik bir metriktir. RMSE tahmin hatalarının (kalıntıların) standart sapmasıdır. Yani, kalıntılar, regresyon hattının veri noktalarından ne kadar uzakta olduğunun bir ölçüsüdür; RMSE ise bu kalıntıların ne kadar yayıldığının bir ölçüsüdür. Başka bir deyişle, verilere en iyi uyan çizgi etrafında o verilerin ne kadar yoğun olduğunu söyler. RMSE değeri 0'dan ∞'a kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir. RMSE değerinin sıfır olması modelin hiç hata yapmadığı anlamına gelir. RMSE, büyük hataları daha fazla cezalandırmanın avantajına sahiptir, bu yüzden bazı durumlara daha uygun olabilir. RMSE, birçok matematiksel hesaplamada istenmeyen mutlak değerin kullanılmasını engeller [15].

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}} \quad (6)$$

$$RMSE = \sqrt{MSE}$$

3 Çalışmada Kullanılan Algoritmalar

3.1 Karar Ağaçları

Karar ağaçları ağaç benzeri grafik veya modellerdir. Ters çevrilmiş bir ağaç gibidir, çünkü kökleri tepede bulunur ve aşağı doğru büyür. Tahmin, test noktaları ve dalları ile bir karar ağacı inşa ederek elde edilebilir. Bir karar ağacının her bir test noktası, belirli bir girdi değişkenini test etmeyi içerir ve her bir dal, yapılmakta olan kararı temsil eder. Daha fazla dal içermeyen bir düğüme yaprak düğüm denir. Düğümün derinliği, düğümden köke ulaşmak için gereken minimum adım sayısıdır. Bir karar ağacının girdi değerleri kategorik veya sürekli olabilir [16]. Karar ağaçlarında en önemli sorunlardan biri herhangi bir kökten itibaren bölünmenin başka bir ifade ile dallanmanın hangi kıstasa göre yapılacağıdır. Karar ağaçlarında sınıflandırma yöntemleri iki çeşittir. Bunlar, Entropiye Dayalı Algoritmalar ile Sınıflandırma ve Regresyon Ağaçları (CART)'lardır. ID3 Algoritması ve C4.5 Algoritması "Entropiye dayalı algoritmalar" arasındayken Twoing Algoritması ve Gini Algoritması "Sınıflandırma ve regresyon ağaçları" sınıfındadır[15]. Her öznitelik, eğitim örneklerinin sınıflandırmasına karar vermek için istatistiksel test kullanılarak değerlendirilir. Hangi özellik en büyük bilgi kazancını sağlarsa ağacın kökünde yer alacak

özellik olarak o seçilir. Bilgi kazancı için entropi adı verilen bir tanımlama kullanılır. Entropi, olayların olma olasılıklarıyla ilişkili olup belirsizliğin ölçülmesi için kullanılan bir ölçüttür. Entropi bilgi ile ilişkilidir ve belirsizlik arttıkça eldeki veriyi daha iyi tanımlamak için daha fazla bilgi gerekecektir. Entropi 0-1 arası değerler alır ve 1 değerine yaklaştıkça belirsizliğin arttığını gösterir. ID3, C4.5, CART algoritmaları en iyi ayırıcı özelliğe sahip değişkeni bulmak için entropiden faydalanır [17]. Karar ağaçlarında nitelik seçimi hesabı yapılırken çeşitli hesaplamalar kullanılır. Bunların bazıları bilgi kazancı (information gain), kazanç oranı (gain ratio), doğruluk (accuracy), least square, gini indeks fonksiyonlarıdır.

3.1.1 JRIP

JRip (RIPPER), temel ve en popüler algoritmalardan biridir. Sınıflar büyüyen boyutta incelenir ve artan azaltılmış hata JRip (RIPPER) kullanılarak sınıf için bir ilk kural seti oluşturulur. Eğitim verilerindeki belirli bir kararın tüm örneklerini bir sınıf olarak ele alarak ve bir dizi kural bularak devam eder ve o sınıfın tüm üyelerini kapsar. Daha sonra bir sonraki sınıfa geçer ve aynı şeyi yapar, tüm sınıflar işlenene kadar bunu tekrar eder. Varsayılan bir kuralla başlarlar ve bir eğitim veri kümesi kullanarak varsayılan istisnaları öngören kuralları öğrenmeye çalışırlar. Öğrenilen her kural, önerme değişmezlerinin bir birleşimidir. Her değişmez bilgi, tek bir özelliğin değerine dayalı olarak verilerin bölünmesine karşılık gelir. Karar ağaçlarına benzeyen bu algoritma ailesi, yorumlanması kolay olma avantajına sahiptir ve deneyler, JRip'in özellikle büyük veri kümelerinde verimli olduğunu göstermektedir. RIPPER ve IREP, doğrudan veri kümesinden çıkarılan sıralı bir kurallar kümesi oluşturmak için ayır ve yönet yöntemine dayalı bir strateji kullanır. Sınıflar, daha fazla unsura sahip olanlara öncelik verilerek tek tek incelenir. Bu algoritmalar, bir durdurma koşulu karşılanana kadar her sınıfa tekrar tekrar uygulanan dört temel adıma (büyütme, budama, optimize etme ve seçme) dayanmaktadır[18]. Bu adımlar şu şekilde özetlenebilir. Büyüme aşamasında, durdurma kriteri karşılanana kadar artan sayıda öngörücü dikkate alınarak kurallar oluşturulur. Budama aşamasında fazlalık ortadan kaldırılır ve uzun kurallar azaltılır. Optimizasyon aşamasında, önceki adımlarda oluşturulan kurallar, yeni öznitelikler eklenerek veya yeni kurallar eklenerek (mümkünse) geliştirilir. Son olarak, seçim aşamasında, en iyi kurallar seçilir ve diğerleri atılır [19].

3.1.2 PART

Algoritma, planlanan kurallar dizisi olan “karar listeleri” adı verilen kural kümeleri üretir. Sırayla listedeki her kuralla yeni bir veri karşılaştırılır ve öğeye ilk eşleşen kuralın sınıfı atanır. PART, her yinelemede kısmi bir C4.5 karar ağacı oluşturur ve "en iyi" yaprağını bir kural haline getirir. PART hem hesaplama performansı hem de sonuçlar açısından çok verimli bir algoritmadır. PART, tipik olarak karar ağacı öğreniminin böl ve yönet stratejisini, kural öğrenmeye özgü ayır ve yönet stratejisi ile birleştirir. Önce bir karar ağacı oluşturulur (C4.5 algoritması kullanılarak) ve en yüksek kapsama alanına sahip yaprak bir kurala dönüştürülür. Ardından, bu kural tarafından kapsanan örnekler grubu atılır ve süreç baştan başlar. Sonuç, önceki herhangi bir kuralı karşılamayan örnekler için geçerli olan varsayılan bir kural tarafından tamamlanan sıralı bir kurallar kümesidir [20].

3.1.3 FURIA

Bulanık kural tabanlı bir algoritmadır. Klasik mantıkta sıcak, soğuk, hızlı yavaş gibi kavramlar vardır. Fakat sıcak ile soğuk arasındaki ılık kavramı, hızlı ile yavaş arasındaki orta hızlı kavramı gibi kavramlar denilince bulanık mantık devreye girer. Bulanık mantıkta kümeye ait her bir eleman [0 1] arasında üyelik dereceleri alır. Geleneksel kurallardan daha geneldir ve birçok avantajı vardır. Örneğin, geleneksel (bulanık olmayan) kurallar “keskin” karar sınırları ve buna bağlı olarak farklı sınıflar arasında ani geçişler olan modeller üretir. Bu özellik şüphelidir ve pek sezgisel değildir. Bunun yerine, bir kural tarafından sağlanan bir sınıfa yönelik desteğin "tam" dan (kuralın özünün içinde) "sıfıra" (sınırın yakınında) ani bir şekilde değil, tedrici bir şekilde düşmesi beklenir. Bulanık kuralların temel özelliklerinden biri olan "yumuşak" sınırları vardır. Kuşkusuz, kesin bir sınıflandırma kararının verilmesi gerekiyorsa, yumuşak sınırların yeniden keskin sınırlara dönüştürülmesi gerekir [21]. FURIA, belirsiz kural tabanlı bir sınıflandırma algoritmasıdır. Bulanık Kuralı, özellikle kural esnetme yaklaşımı kullanır.

4 Veri Setleri

Tabo 4.1. Çalışmada kullanılan veri setleri

Veri Seti	Büyüklüğü	Sınıf sayısı	Veri Tipi
Messidor_İmage	1151 ×20	2	Tamsayı
Labor	57 × 16	2	Tamsayı
Diabets	7×768	2	Numeric
Wine	15×178	3	Numeric
Hipotiroid	3800×21	2	Tamsayı
Credit Card	211× 1000	9	Nümeric
Soy Bean	35×700	2	Tamsayı
Sensör	13×2212	3	Numeric

Tablo 4.1. de çalışmada kullanılan veri setlerine ait ad, büyüklük, sınıf sayısı ve veri tipi tanımlanmıştır. Bu veri setleri;

Messidor veri tabanı, diyabetik retinopatinin bilgisayar destekli tanıları ile ilgili çalışmaları kolaylaştırmak için oluşturulmuştur. Veri tabanında bulunan tüm görüntüler, gerçek klinik teşhisler yapmak için kullanılmıştır. Hasta mahremiyetinin en üst düzeyde korunmasını sağlamak için, bir hastayı tanımlamaya izin verebilecek bilgiler atılmıştır ve herhangi bir konuyu tanımlamak için görüntülerin tek başına veya başkalarıyla birlikte kullanılabileceğine dair gerçek bir bilgi yoktur [22].

Labor veri seti; Cinsiyete Göre Ayrıştırılmış İşgücü Veritabanı (GDLD), Dünya Bankası hane halkı anket toplamasına ve diğer kamu kaynaklarına dayanan küresel bir mikro işgücü veri tabanıdır. Bu veri tabanı, ekonomik faaliyetleri ve meslek kategorilerini yerel sınıflandırmadan uluslararası karşılaştırılabilir sınıflandırmalara kadar uyumlu hale getirmiştir. Eğitim, istihdam seviyeleri, ücretler, işgücü geliri ve genellikle mevcut olandan çok ayrıştırılmış ekonomik aktivite seviyesi ve meslek kategorisinde istihdam durumu hakkında ayrıntılı hesaplar sunarak küresel cinsiyet istatistiklerinde önemli bir bilgi boşluğunu doldurur [23]. Diabets veri seti; Diyabet hasta kayıtları iki kaynaktan elde edilmiştir. Bu veri seti aslen Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsünden alınmıştır. Veri setinin amacı, bir hastanın diyabetli olup olmadığını, veri setine dahil edilen belirli tanısal ölçümlere dayalı olarak tanısal olarak tahmin etmektir [24].

Wine veri seti; bu veriler, İtalya'da aynı bölgede yetiştirilen, ancak üç farklı çeşitten elde edilen şarapların kimyasal analizinin sonuçlarıdır. Analiz, üç şarap türünün her birinde bulunan 13 bileşenin miktarlarının sonuçlarıdır [25]. Hipotiroid veri seti; "machine learning repository" adlı internet sitesinin veri tabanı kullanılmıştır. Verileri dağılımı: "Hipertiroid=35 sonuç, Hypotiroid=30 sonuç ve sağlıklı sonuç sayısı =150" şeklindedir. Tanıyı koyabilmek için kullanılan giriş verileri 5 adet (T3RU, T4, T3, TSH and MAD-TSH), çıkış ise bir adettir (0=sağlıklı,1=hipertiroid, 2=hipotiroid)[26].

Credit Card verileri; Kredi kartı şirketlerinin hileli kredi kartı işlemlerini fark edebilmeleri önemlidir, böylece müşteriler satın almadıkları ürünler için ücretlendirilmezler.
Bu veri kümesi, iki gün içinde gerçekleşen ve 284.807 işlemden 492'sini sahtekarlık yapılan işlemleri içerir[27].

SoyBean verileri; bu, eğitim ve test veritabanının tek bir dosyada birleştirildiği UCI deposundaki büyük soya veritabanıdır. Sadece ilk 15'i önceki çalışmalarda kullanılmış olan 19 sınıf vardır. Bazıları nominal ve bazıları sıralı 35 kategorik nitelik vardır. Özniteliklerin değerleri sayısal olarak kodlanır, ilk değer "0", ikincisi "1" vb. Olarak kodlanır[28]. Sensör Verileri; Sensör verileri, fiziksel ortamdan bir tür girdiyi algılayan ve bunlara yanıt veren bir aygıtın çıktısıdır. Çıktı, başka bir sisteme bilgi veya girdi sağlamak veya bir süreci yönlendirmek için kullanılabilir[29].

Veri madenciliği için kullanılacak olan verilerin az ve sınırlı olması durumunda elde edilen sonuçların değerlendirilmesi için bekletme yaklaşımına seçenek olarak kullanabileceğiniz değerlendirme yöntemi k-katlı çapraz geçerleme yaklaşımıdır. Toplam N örnekten oluşan veri seti k adet (5 veya 10 gibi) eşit parçaya ayrılır (Eğer toplam örnek sayısı k sayısına tam olarak bölünemiyorsa ayrılan son parçadaki örnek sayısının diğer k-1 parçadaki örnek sayılarından daha az olacağı unutulmamalıdır). Bu durumda k adet analiz ardışık olarak gerçekleştirilmektedir. Çapraz geçerleme sürecinde sırasıyla k parçanın her biri test veri seti olarak kullanılırken diğer k-1 parçalar eğitim veri seti olarak kullanılmaktadır [30].

Çalışmada kullanılan tüm veri setleri 10-kat çaprazlama yöntemi kullanılarak dört algoritmada test edilmiştir.

Araştırmada kullanılan bilgisayarların donanım özellikleri (CPU: P4 1.8 Mhz, R.A.M: 256 Mb, Ekran Kartı: 64 MB paylaşımsız, Sabit Disk: 40 Gb, Cd_Rom: 52x) standart bir yazılımı yeterince hızlı çalıştırabilecek düzeydedir.

5 Deneysel Sonuçlar

Ek-1'de gösterildiği gibi dört sınıflandırma algoritması Karar Ağaçları, JRIP, PART ve FURIA

kullanılarak analiz edilmiştir. Dört algoritma açık erişimli, veri madenciliği doğrulama yöntemleri için kullanılan sekiz gerçek veri seti kullanılarak uygulanmıştır. Bu veri setleri farklı büyüklüklerde olup ortak yanları kategorik ve sayısal verilerden oluşmalarıdır. Bu algoritmalar arasında FURIA algoritmasının doğru sınıflandırma yüzdesinin yüksek, ortalama mutlak hatasının düşük olduğu görülmektedir. Bu araştırmada, FURIA'nın yüksek düzeyde doğruluğa ulaştığı saptanmıştır.

Sırasıyla veri sayısı en yüksek olan Hipotiroid ve Sensör veri setlerinde doğru sınıflandırma oranlarının en yüksek olduğu görülmektedir. İyi bir sınıflandırma algoritmasının büyük veri setleri için de doğru sınıflandırma yapabilmesi beklenir. Elde edilen bulgulara göre veri seti büyük iken de FURIA algoritmasının kullanılması önerilebilir.

6 Sonuçlar

Sınıflandırma, istatistikçiler ve makine öğrenimi araştırmacıları tarafından geleneksel problemleri çözmek için kullanılan bir yöntemdir. Bir modelin işleyişi, sınıf üyeliği, yanlış kategorize etme, eğitimin boyutu ve test setleri gibi eğitim algoritmasının yanı sıra diğer özelliklere de bağlı olabilir. Her eğitim örneği, bir hipotezin doğru olma olasılığına bağlı olarak aşamalı olarak artabilir veya azalabilir. Gruplama ve tahmin, veri analizi için kullanılan iki yöntemdir.

Bu çalışmada kural tabanlı sınıflandırma algoritmalarından PART, JRIP, FURIA ve Karar Ağaçları algoritmaları kullanılmıştır. Bu algoritmalar WEKA platformunda birçok alandan Wine, Soybean, Labor, Sensör, Hipotiroid, Diabet, Kredi card, Messidor_futures veri setleri sınıflandırılan örneklerin sayısı, yüzdesi, ortalama mutlak hata (MAE), kök ortalama kare hata (RMSE) özellikleri referans alınarak karşılaştırılmıştır. Bu karşılaştırmada doğru sınıflandırma açısından en iyi algoritmanın bulanık mantık tabanlı algoritma olan FURIA algoritması olduğu görülmektedir. Diğer algoritmalara göre en düşük performans gösteren algoritma ile ilgili bir yorum yapılamaz. Büyük veri analiz yöntemlerinde FURIA algoritmasının sınıflandırma algoritması olarak kullanılması önerilmektedir.

Kaynaklar

[1] Andreeva P., Dimitrova M., ve Radeva P. "Data Mining Learning Models And Algorithms For Medical Application", Proceedings Of The 18-Th Conference On Saer, Pp 11-18, 2004.

[2] Küçüksille, E. "Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İmkb Hisse Senetleri Piyasasında Bir Uygulama", Doktora Tezi, Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü. Isparta, 2009.

[3] Fawcett T. An introduction to ROC analysis, Pattern recognition letter; 27 (8): 861–874, 2006.

[4] Duda R.O., Hart P.E., Stork D.G. Pattern Classification, second ed., Wiley, New York, 2001.

[5] Craft J.L. Statistics and Data Analysis for Social Workers, second ed., F.E. Peacock Publishers, USA, 1990.

[6] Cohen, W.W. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9–12; pp. 115–123, 1995.

[7] Holte R.C. Very simple classification rules perform well on most commonly used dataset, Mach. Learn. 11 – (63–91), 1993.

[8] Frank E. ve Witten I.H. Generating accurate rule sets without global optimisation, in: J. Shavlik (Ed.), Machine Learning, Proceedings of the Fifteenth International Conference, Morgan Kaufmann, San Francisco, CA, 1998.

[9] Hühn J. ve Hüllermeier E. FURIA: an algorithm for unordered fuzzy rule induction. Data Min Knowl Disc 19(3):293–319, 2009.

[10] John G.H, Langley P.Estimating continuous distributions in Bayesian classifiers, in: Proceeding of the Eleventh Confer- ence on Uncertainty in Artificial Intelligence, Morgan Kauf- mann, San Mateo, CA, pp. 338–345, 1995.

[11] Lodhi H., Shawe-Taylor J., Christianini N., Watkins C. Text classification using string kernels, in: T. Leen, T. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, vol. 13, MIT Press, 2, 2001.

[12] Japkowicz N. Performance evaluation for learning algorithms, Cambridge University Press, Cambridge, 2011.

[13] Flach, P.. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 1st, Cambridge University Press Glasgow, UK, ISBN: 978-1-107-09639-4, 2012.

[14] Mean Absolute Error, http://en.wikipedia.org/wiki/Mean_absolute_error, Erişim Tarihi: 08.12.2019.

[15] http://en.wikipedia.org/wiki/Mean_absolute_error Erişim Tarihi: 08.12.2019.

[16] Dietrich, D., Heller, B., Yang, B. Data Science & Big Data Analytics. John Wiley & Sons, U.S.A., 409, 2015.

[17] Haciefendioğlu, Ş. Makine Öğrenmesi Yöntemleri ile Glokom Hastalığının Teşhisi. Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2012.

[18] Rajput, A., Aharwal, R.P., Dubey, M., Saxena, S., Raghuvanshi, M. J48 and JRIP rules for e-governance data. *IJCSS*, *5*, 201, 2011.

[19] Witten, I.H., Frank, E.; Hall, M.A. Introduction to Weka. In *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Witten, I.H., Frank, E., Hall, M.A., Eds.; The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann: Boston, MA, USA; pp. 403–406, 2011.

[20] Qidwai U., Chaudhry J., Jabbar S., Zeeshan HMA., Janjua N., Khalid S. Using casual reasoning for anomaly detection among ECG live data streams in ubiquitous healthcare monitoring systems. J Ambient Intell Humaniz Comput 10(10):4085–4097, 2019.

[21] Salem O., Serhrouchni A., Mehaoua A., Boutaba R. Event detec- tion in wireless body area networks using Kalman filter and power divergence. IEEE Trans Netw Serv Manag 15(3):1018–1034, 2018.

[22] https://www.adcis.net/en/third-party/messidor/ Erişim Tarihi: 08.12.2019.

[23] https://datacatalog.worldbank.org/dataset/gender-disaggregated-labor-database Erişim Tarihi: 08.12. 2019.

[24] https://www.kaggle.com/uciml/pima-indians-diabetes-database Erişim Tarihi: 08.12.2019.

[25] https://archive.ics.uci.edu/ml/datasets/wine Erişim Tarihi: 08.12.2019.

[26] https://archive.ics.uci.edu/ml/datasets/thyroid+disease. Erişim Tarihi: 08.12.2019.

[27] https://www.kaggle.com/mlg-ulb/creditcardfraud. Erişim Tarihi: 08.12.2019.

[28] https://datahub.io/machine-learning/soybean. Erişim Tarihi: 08.12.2019.

[29] https://data.world/datasets/sensors Erişim Tarihi: 08.12.2019.

[30] Bramer, M. Principles of Data Mining (2nd ed.). London: Springer-Verlag, 2013.

Ek-1 Deneysel Sonuçlar

Veri Seti	Algoritma Adı	Doğru sınıflandırılan örneklerin Sayısı-Yüzdesi	Ortalama Mutlak Hata	Hata Kareler ortalamasının karekökü
Messidor_Features	Karar Ağaçları	718- %62.38	0.43	0.47
	JRIP	724-%62.90	0.45	0.48
	PART	744-%64.63	0.39	0.48
	FURIA	**753-%65.41**	**0.34**	**0.54**
Credit Card	Karar Ağaçları	710-%71	0.36	0.43
	JRIP	717-%71.7	0.37	0.44
	PART	702- %70.2	0.32	0.49
	FURIA	**725-%72.5**	**0.27**	**0.49**
Diabet	Karar Ağaçları	547- %71.22	0.34	0.42
	JRIP	577-%75.13	0.34	0.42
	PART	578- %75.26	0.31	0.41
	FURIA	**581-%75.65**	**0.25**	**0.47**
Hipotiroid	Karar Ağaçları	3747-%99.33	0.02	0.07
	JRIP	3747-%99.33	0.004	0.05
	PART	3750-%99.41	0.003	0.05
	FURIA	**3752-%99.46**	**0.003**	**0.04**
Sensör	Karar Ağaçları	2051-%92.72	0.11	0.20
	JRIP	2157-%97.51	0.02	0.12
	PART	2168-%98.01	0.01	0.11
	FURIA	**2182-%98.64**	**0.011**	**0.08**
Wine	Karar Ağaçları	146- %85.88	0.16	0.26
	JRIP	156-%91.7	0.06	0.22
	PART	156- %91.76	0.08	0.229
	FURIA	**159- %93.5**	**0.05**	**0.20**
Soybean	Karar Ağaçları	576-%84.33	0.06	0.15
	JRIP	630- %92.24	0.011	0.08
	PART	628-%91.94	0.013	0.084
	FURIA	**637- %93.265**	**0.008**	**0.07**
Labor	Karar Ağaçları	43-%75.43	0.29	0.40
	JRIP	44-%77.19	0.22	0.45
	PART	45-%78.94	0.28	0.43
	FURIA	**46-%80.70**	**0.18**	**0.39**