

Makine Öğrenmesi Algoritmalarında Veri Setlerine Öznitelik Seçim Tekniklerinin Uygulanması ve Karşılaştırmalı Analizi

Application and Comparative Analysis of Attribute Selection Techniques to Data Sets in Machine Learning Algorithms.

Emrehan EKENTOK

Bilgisayar ve Elektrik Elektronik Mühendisliği
KTO Karatay Üniversitesi - Konya, Türkiye
e.ekentok@gmail.com

Özetçe— Bu çalışmada, Makine öğrenmesi algoritmalarına tabi tutulan veri setlerinde öznitelik seçim tekniğinin uygulanması ile elde edilen sonuçların, tüm özniteliklerin kullanıldığı veri setiyle elde edilen sonuçlar (özgün veri setinin öznitelikleri) ile karşılaştırılması, deneysel gözlemlerin sonuçları ve öznitelik seçim tekniklerinin kullanımının önemi hakkında çıkarımlarda bulunulmuştur.

Anahtar Kelimeler — Makine Öğrenmesi; Öznitelik Seçimi; Naive Bayes;

Abstract— In this study, the results obtained by applying the feature selection technique to the data sets subjected to the machine learning algorithms were compared with the results obtained with the data set of all the attributes (original data sets attributes). the results of the experimental observations and the importance of using the feature selection techniques has been mentioned.

Keywords — Machine Learning; Attribute Selection; Naive Bayes

I. GİRİŞ

Günümüzde Makine Öğrenmesi algoritmaları Üretim, Pazarlama-Satış, Sağlık Bilimleri, Finans ve Mali Hizmetler, Enerji, Seyahat veya Sosyal Platformlar gibi birçok alanda sıklıkla kullanılmaktadır. Makine Öğrenmesi Algoritmalarının kullanılmakta olduğu bazı alan ve uygulamalarda kullanılan veri setlerinin sahip olduğu öznitelik sayıları oldukça yüksek sayılardan oluşabilmektedir. Hal böyle olunca araştırmacılar öznitelik seçim yöntemlerine önceki çalışmalara nazaran daha çok ihtiyaç duymaktadırlar. Öznitelik seçimi uygulanmış veri setinde yapılacak işlem sayısı azalmakla beraber gürültülü veya veri setiyle ilgisiz olan öznitelikler veri setinden çıkarılarak,

sınıflandırma başarı oranı yükselmekte, sınıflandırma prosesi (eğitim zamanı, daha az bellek kullanımı) kolaylaşmakta, daha az öznitelik ile veri setinin tüm özniteliklerinin kullanılmasına kıyasla ufak hata oranları, aynı başarı oranları veya daha iyi sonuçlar elde edilmektedir.

Özellikle, Örnekleme sayısının az fakat öznitelik sayısının fazla olduğu veri setlerine sahip uygulamalar ele alındığında, öznitelik seçimi yaparak sınıflandırmaya daha uygun özniteliklerin belirlenebilmesi ve/veya işlem yoğunluğunun sadeleştirilmesi konusunda öznitelik seçim tekniğinin önemi ortaya çıkmaktadır.

Bu bildiride iki farklı veri seti üzerinde öznitelik seçimi tekniği uygulanarak elde edilen sonuçların özgün veri setinden elde edilen başarı sonuçları ile karşılaştırılması analiz edilmiş ve öznitelik seçim tekniklerinin öneminin vurgulanması hedeflenmiştir.

II.

MATERYAL VE METOT

A. Öznitelik Seçimi

Öznitelik seçim işlemi veri setindeki bağımlı değişken (Supervised Variable) ile hiç ilgisi bulunmayan, gürültülü özniteliklerin elenmesi veya bağımlı değişkeni açıklama unsuru daha yüksek olanların kümelerin belirlenmesi işlemidir.

Genel kanı, veri setinde tüm özelliklerin kullanılması ile daha başarılı sonuçlar elde edileceği yönündedir. Ancak bu yaklaşım çok sayıda öznitelik içeren veri setlerinde her zaman doğru olmayabilir. Veri setindeki her öznitelik algoritmaya uygun açıklayıcı veya doğru bilgiler taşımayabilir. Diğer bir deyişle veri setlerinde bazı öznitelikler algoritmanın işleyiş performansına negatif olarak etki edecek gürültülü bilgi

içerebilir, bu öz niteliklerin ayrıştırılması algoritma başarı oranı artırmakla beraber, algoritmada kullanılacak veri boyutunun indirgenmesinin avantajlarında beraberinde getirmektedir.

Veri setleri içerisindeki etkin öz nitelikleri seçme işlemi WEKA programı ile gerçekleştirilmiştir. Öz nitelik seçim yöntemi olarak Correlation-based Feature Subset Selection (CfsSubsetEval), arama metodu olarak bestFirst tercih edilmiştir.

CFS, 1999 yılında Hall tarafından geliştirilmiştir [1]. CFS, alt öz nitelik kümelerini korelasyon-bazlı değerlendirerek en iyi alt öz nitelik kümesini bulmayı hedefleyen filtre modeli öz nitelik seçme algoritmalarından biridir. Temel prensip olarak kendi aralarında korelasyonu az, sınıf etiketleri ile korelasyonu fazla öz nitelik alt kümesini seçmeye çalışan bir algoritmadır.

B. Veri Setleri

Bu çalışmada UCI veritabanından, *Waveform Database Generator (version 1)* veri seti ve Waikato üniversitesinde açık kaynak kodlu olarak JAVA dili üzerinde geliştirilmiş Weka yazılımının (versiyon 3.8.1) kütüphanesinde bulunan *Vote* veri seti kullanılmıştır.

Waveform Database Generator veri setinde 3 adet sınıf, 40 öz nitelik ve 5000 adet örnek bulunmaktadır. *Vote* veri setinde ise 16 öz nitelik ve 465 örnek bulunmaktadır.

C. Sınıflandırma

Bildiriye konu olan bu çalışmada her iki veri setinde de makine öğrenmesi algoritmalarından Naive Bayes algoritması kullanılmıştır.

Naive Bayes Sınıflandırma algoritması Bayes teoremine dayanan temel bir olasılıksal sınıflandırma yöntemidir. Hali hazırda sınıflandırılmış durumdaki örnek verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine ait olma olasılığını hesaplayan bir yaklaşımdır. Bu sınıflandırma algoritmasında öz nitelikler birbirinden bağımsız olarak değerlendirilir. Bir Bayes yaklaşımı olarak, n boyutlu uzayda tanımlı olan X vektörü (x_1, \dots, x_n) , m adet sınıf bulunan C_k (C_1, \dots, C_n) veri kümesinde son olasılığı maksimize eden bir sınıf etiketi C arar.

III. DENEYSEL SONUÇLAR

Deniz dalga formlarının sınıflandırılması için elde edilen *Waveform Database Generator* veri seti ve *Vote* veri seti Naive Bayes algoritması, çapraz doğrulama ($k=10$) tekniği ile çalıştırılmıştır. Çapraz doğrulama test yönteminde veri seti 10 eşit kümeye bölünmekte, 9 küme eğitim için 1 küme test için kullanılmaktadır. Test sonunda 10 adet performans metriği elde edilmekte ve elde edilen her bir metriğin aritmetik ortalaması alınmaktadır.

Waveform Database Generator veri seti özgün olarak 40 adet öz niteliğe ve 1 adet sınıflandırma niteliğine sahiptir (class) sahiptir. 40 öz nitelik ile Naive Bayes sınıflandırıcı ile 10k çapraz doğrulama test tekniği kullanılarak gerçekleştirilen çalışmanın sonucunda, % 80 'lik bir başarı oranı elde edilmiştir.

Correctly Classified Instances	4000	80	%
Kappa statistic	0.7005		
Mean absolute error	0.1357		
Root mean squared error	0.3369		
Relative absolute error	30.5282 %		
Root relative squared error	71.4731 %		
Total Number of Instances	5000		
=== Confusion Matrix ===			
a	b	c	<-- classified as
890	394	408	a = 0
16	1547	90	b = 1
2	90	1563	c = 2

Resim 1; Özgün Öz nitelikler ile Sınıflandırma Sonucu

Özgün veri setine uygulanan Öz nitelik Seçim tekniği *Correlation-based Feature Subset Selection (CfsSubsetEval)* olarak seçilmiş, arama metodu ise *bestFirst* tercih edilmiştir. Bu işlem sonucunda veri setinde 15 adet öz nitelik kalmış 25 adet öz nitelik çıkartılarak veri setinde kullanılmamıştır. 15 öz nitelik ile aynı şekilde Naive Bayes ile 10k çapraz doğrulama test tekniği kullanılarak sınıflandırma algoritması çalıştırılmıştır. Başarı oranı % 80.12 'ye yükselerek sadece 15 adet öz nitelik ile bu sonuç elde edilmiştir. Sonuçlara ilişkin diğer Detaylar ise aşağıdaki Resim 2 görselinde verilmiştir.

Correctly Classified Instances	4006	80.12	%
Kappa statistic	0.7023		
Mean absolute error	0.1349		
Root mean squared error	0.333		
Relative absolute error	30.3476 %		
Root relative squared error	70.6464 %		
Total Number of Instances	5000		
=== Confusion Matrix ===			
a	b	c	<-- classified as
894	394	404	a = 0
15	1546	92	b = 1
4	85	1566	c = 2

Resim 2; Azaltılmış Öz nitelik ile Sınıflandırma Sonucu

Weka 3.8.1 versiyonlu yazılımın kütüphanesinde bulunan *Vote (oy)* veri seti özgün olarak 16 adet öz nitelik ve 1 adet sınıflandırma (class) niteliğine ve 465 adet örneğe sahiptir. 16 öz nitelik özgün olarak Naive Bayes sınıflandırıcı algoritması ile 10k çapraz doğrulama test tekniği kullanılarak çalıştırıldığında % 90.1149'luk başarı oranı elde edilmiştir. Sonuçlar ile ilgili detaylar Resim 3' de gösterilmiştir.

Correctly Classified Instances	392	90.1149	%
Kappa statistic	0.7949		
Mean absolute error	0.0995		
Root mean squared error	0.2977		
Relative absolute error	20.9815 %		
Root relative squared error	61.1406 %		
Total Number of Instances	435		
=== Confusion Matrix ===			
a	b	<-- classified as	
238	29	a = democrat	
14	154	b = republican	

Resim 3; Weka Sınıflandırma Sonucu

VOTE veri setine *CfsSubsetEval* — *bestFirst* öz nitelik seçim tekniği uygulandığında ise veri setinde 4 adet öz nitelik seçilmiş 12 adet öz nitelik veri setinden çıkartılarak kullanılmamıştır. 4 öz nitelik aynı şekilde Naive Bayes ile 10k çapraz doğrulama test tekniği kullanılarak sınıflandırma algoritması çalıştırılmıştır. Başarı oranı % 96.092 'ye yükselmiştir.

Correctly Classified Instances	418	96.092 %
Kappa statistic	0.9177	
Mean absolute error	0.0575	
Root mean squared error	0.1768	
Relative absolute error	12.1285 %	
Root relative squared error	36.3023 %	
Total Number of Instances	435	

=== Confusion Matrix ===

a	b	<-- classified as
258	9	a = democrat
8	160	b = republican

Resim 4; Azaltılmış Öz nitelik ile Sınıflandırma Sonucu

Öznitelik Seçim Tekniğinin Veri Setlerine Uygulanması ile Elde Edilen Sonuçların Karşılaştırılması					
Veri Seti	Özgün Öz nitelik	Azaltılmış Öz nitelik	Algoritma	Başarı Oranı	Az.Öz.Başarı Oranı
WaveForm	40	15	Naive Bayes	%80	% 80.12
Vote	16	4	Naive Bayes	% 90.1149	%96.092

Tablo 1; Başarı Sonuçlarının Karşılaştırılması

IV. SONUÇLAR VE DEĞERLENDİRME

Çalışmaya Konu olan Makine Öğrenme Algoritmalarında kullanılan veri setlerinde filtre olarak da değerlendirilebilen Öz nitelik seçim tekniği iki adet veri setine uygulanmıştır. Bu veri setlerinin her ikisinde de öz nitelik sayıları CFS Tekniği ile azaltılmış ve Naive Bayes algoritması 10 katlı Çapraz Doğrulama test tekniği ile çalıştırılmıştır. Elde edilen sonuçlar karşılaştırılmıştır.

Her iki veri setinde de bildiriye konu olan öz nitelik azaltma işleminin önemi ve başarısı açıkça görülmüştür. Daha az öz nitelik ile elde edilen başarı oranı analiz edilerek tespit edilmiştir. Ayrıca daha çok öz niteliğe sahip olan veri setlerinde ise Öz nitelik seçim tekniklerinin başarı oranlarında azaltılmamış özgün öz nitelik sayılı veri seti ile yapılan çalışmalardan daha başarılı olabileceğine dair sonuçlar gözlemlenmiştir.

Sonuç olarak günümüzde birçok alan ve uygulamada kendine yer bulan makine öğrenmesi algoritmalarında ayrıştırılmış ve/veya azaltılmış öz nitelikler sayesinde başarı oranlarının artırılması, işlem sürelerinin azaltılması, en uygun öz nitelikler ile algoritmaların çalıştırılması, daha az öz niteliğe sahip olunması gibi yararları gözlemlenmiştir.

Bu çalışmada yeni bir yöntem veya metot önerimi yapılmamış makine öğrenmesi algoritmalarında öz nitelik seçim tekniklerinin kullanılmasının önemi ve sonuçları tecrübe edilerek aktarılmıştır. ileride bu konu ile ilgili yapılacak çalışmalara bir temel olması amaçlanmıştır.

KAYNAKLAR

1.