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
Elektrik ve Bilgisayar Mühendisliği
KTO Karatay Üniversitesi - Konya, Türkiye
emrehan.ekentok@ogrenci.karatay.edu.tr

***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; Classification Analysis;

***Abstract*— In this study, the results obtained by applying the Attribute 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 Attribute Selection techniques has been mentioned.**

Keywords — Machine Learning; Attribute Selection; Sınıflandırma Analizi;

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 yoğunluğun 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

performasına negatif olarak etki edecek gürültülü bilgi içerebilir, bu özniteliklerin ayrıştırılması algoritma başarı oranı artırmakla beraber, algoritmada kullanılacak veri boyutunun indirgenmesinin avantajlarınıda beraberinde getirmektedir.

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

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

B. Veri Setleri

Bu çalışmada UCI veritabanından, *Waveform Database Generator (version 1) v*eri 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 öznitelik ve 5000 adet örnek bulunmaktadır. *Vote* veri setinde ise 16 öznitelik ve 465 örnek bulunmaktadır.

C. Sınıflandırma

Bildiriye konu olan bu çalışmada her iki veri setindede 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 öznitelikler birbirinden bağımsız olarak değerlendirilir. Bir Bayes yaklaşımı olarak, n boyutlu uzayda tanımlı olan X vektörü (x_1,..,x_2), m adet sınıf bulunan C_k (C_1,.,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 (k10) 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 özniteliğe ve 1 adet sınıflandırma niteliğine sahiptir (class) sahiptir. 40 öznitelik ile Naive Bayes sınıflandırıcı ile 10k çapraz doğrulama test tekniği kullanılarak gerçekleştirilen çalışmanın sonucunda, % 80 ‘lik bir başarı oranı elde edilmiştir. Sonuçlar Tablo 1’de Gösterilmiştir.

Doğru Sınıflandırılmış Örnekler (Correctly Classified Instances)	4.000 (% 80)
Kappa İstatistiği (Kappa Statistic)	0.7005
Ortalama Mutlak Hata (Mean Absolute Error)	0.1357
Ortalama Karesel Hatanın Karekökü (Root Mean Squared Error)	0.3369
Göreceli Mutlak Hata (Relative Absolute Error)	% 30.5282
Göreceli Karesel Hatanın Karekökü (Root Relative Squared Error)	% 71.4731
Toplam Örnekleme Sayısı	5.000

Karmaşa Matrisi			
A	**B**	**C**	
890	**394**	**408**	**a = 0**
16	**1547**	**90**	**b = 1**
2	**90**	**1563**	**c = 2**

Tablo 1; Özgün Öznitelikler ile Sınıflandırma Sonuçları

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

Doğru Sınıflandırılmış Örnekler (Correctly Classified Instances)	4.006 (% 80.12)
Kappa İstatistiği (Kappa Statistic)	0.7023
Ortalama Mutlak Hata (Mean Absolute Error)	0.1349
Ortalama Karesel Hatanın Karekökü (Root Mean Squared Error)	333
Göreceli Mutlak Hata (Relative Absolute Error)	% 30.3476
Göreceli Karesel Hatanın Karekökü (Root Relative Squared Error)	% 70.6464
Toplam Örnekleme Sayısı	5.000

Karmaşa Matrisi			
A	**B**	**C**	
890	**394**	**404**	**a = 0**
15	**1546**	**92**	**b = 1**
4	**85**	**1566**	**c = 2**

Tablo 2; Azaltılmış Öznitelikler ile Sınıflandırma Sonuçları

Weka 3.8.1 versiyonlu yazılımın kütüphanesinde bulunan *Vote (oy)* veri seti özgün olarak 16 adet öznitelik ve 1 adet sınıflandırma (class) niteliğine ve 465 adet örneğe sahiptir. 16 öznitelik 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 Tablo 3' de gösterilmiştir.

Doğru Sınıflandırılmış Örnekler (Correctly Classified Instances)	392 (% 90.1149)
Kappa İstatistiği (Kappa Statistic)	0.7949
Ortalama Mutlak Hata (Mean Absolute Error)	0.0995
Ortalama Karesel Hatanın Karekökü (Root Mean Squared Error)	0.2977
Göreceli Mutlak Hata (Relative Absolute Error)	% 20.9815
Göreceli Karesel Hatanın Karekökü (Root Relative Squared Error)	% 61.1406
Toplam Örnekleme Sayısı	435

Karmaşa Matrisi		
A	**B**	
238	**29**	**a = demokrat**
14	**154**	**b = cumhuriyetçi**

Tablo 3; Özgün Öznitelikler ile Sınıflandırma Sonuçları

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

Doğru Sınıflandırılmış Örnekler (Correctly Classified Instances)	418 (% 96.092)
Kappa İstatistiği (Kappa Statistic)	0.9177
Ortalama Mutlak Hata (Mean Absolute Error)	0.0575
Ortalama Karesel Hatanın Karekökü (Root Mean Squared Error)	0.1768
Göreceli Mutlak Hata (Relative Absolute Error)	% 12.1285
Göreceli Karesel Hatanın Karekökü (Root Relative Squared Error)	% 36.3023
Toplam Örnekleme Sayısı	435

Karmaşa Matrisi		
A	**B**	
258	**9**	**a = demokrat**
8	**160**	**b = cumhuriyetçi**

Tablo 4; Azaltılmış Öznitelikler ile Sınıflandırma Sonuçları

Öznitelik Seçim Tekniğinin Veri Setlerine Uygulanması ve Özgün Öznitelikler ile Sınıflandırma Başarı Sonuçların Karşılaştırılması					
Veri Seti	**Özgün Öznitelik**	**Azaltılmış Öznitelik**	**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 5; Sınıflandırma Sonuçlarının Karşılaştırılması

IV. SONUÇLAR VE DEĞERLENDİRME

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

Her iki veri setindede bildiriye konu olan öznitelik azaltma işleminin önemi ve başarısı açıkça görülmüştür. Daha az öznitelik ile elde edilen başarı oranı analiz edilerek tespit edilmiştir. Ayrıca daha çok özniteliğe sahip olan veri setlerinde ise Öznitelik seçim tekniklerinin başarı oranlarında azaltılmamış özgün öznitelik 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 öznitelikler sayesinde başarı oranlarının artırılması, işlem sürelerinin azaltılması, en uygun öznitelikler ile algoritmaların çalıştırılması, daha az özniteliğe sahip olunması gibi yararları gözlemlenmiştir.

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

KAYNAKLAR

1. Hall M, “Correlation-Based Feature Selection For Machine Learning” Phd Thesis, Department Of Computer Science, Waikato University, New Zealand, 1999
2. A. Gümüşçü, İ. B. Aydilek, R. Taşaltın , “Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması” HU Muh. Der. 01 (2016) p.1-7. [1]
3. H. Nizam, S. S. Akın, 2014, “Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması” XIX. Türkiye'de İnternet Konferansı, İzmir. [2]
4. T. PARLAR, Esra SARAÇ, S.A. Özel, 2017, “Türkçe Twitter Verilerinde Duygu Analizi için Nitelik Seçim Yöntemlerinin Karşılaştırılması” 25. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı
5. B. Yazıcı, F. Yaslı, H. Yıldız Gürleyik, U.O. Turgut M. S. Aktaş, Oya Kalıpsız “Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama” UYMS 2015
6. M. Bilgin, 2017, “Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi”
7. B. Bektaş, S. Babur, “Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi” Tıp Teknolojileri Kongresi 2016.
8. Liu H, Setiono R, “Chi2: Feature Selection And Discretization Of Numeric Attributes” In: Proceedings Of The IEEE 7th International Conference On Tools With Artificial Intelligence 338- 391, 1995.
9. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
10. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.