Zamanla Değişen ve Gelişen Arabaların Konsept Yapılarına Göre Sınıflandırılması
Classification of changing and developing cars according to concept structures

Sezerin Küçük
Elektrik-Elektronik Mühendisliği Bölümü
Kto Karatay Üniversitesi
Konya, Türkiye
Sezerin.kucuk@gmail.com

***Özetçe*—Teknolojinin ve zamanın ilerlemesiyle her gün daha da gelişen arabalar günümüzde ihtiyaçlarımız listesinde önemli bir rol oynamaktadır. Bu alandaki çalışmalar, farklı öznitelik çıkarma ve sınıflandırma yöntemlerinin kullanılmasıyla her geçen gün artmaktadır. Bu çalışmada Marko Bahonec'in 1997'de sunmuş olduğu Car Evaluation veri kümesi kullanılmıştır. Araba modelleri değişen yapı konseptlerine göre incelenmiştir. Weka'da İki bölüme veri kümesi ayrılmış olup, bir bölüm eğitim veri kümesi tahmin modeli üretmek için kullanılır ve diğer bölüm test veri modeli doğruluğunu test etmek için set olarak kullanılır. Veri kümesinden gelen tüm veriler eğitim veri kümesine ve test veri kümesine seçilebilir anlamına gelir ve bunun için çapraz doğrulama yöntemi (Cross-validation) kullanıldı. En yüksek sınıflandırma doğruluğu, belirtilen dönüşüm yöntemlerinin birlikte kullanılması ve sınıflandırma yönteminin Rastgele Orman olduğu durumda %94,6335 olarak elde edilmiştir ve bu da tercih edilen sınıflandırma yöntemi ve dönüşüm yöntemlerinin literatürdeki diğer sınıflandırıcılara bakıldığında etkili olduğunu göstermektedir.**

***Abstract*— With the advancement of technology and time, more and more developing cars play an important role in the list of our needs. The studies in this area increase every day with the use of different feature extractions and classification methods. In this study, the car Evaluation data set presented by Marko Bahonec in 1997 was used. We separated the data set into two parts, one part is used as training data set to produce the prediction model, and the other part is used as test data set to test the accuracy of our model.(Cross-validation).The highest classification accuracy was obtained as 94,6335% when the specified conversion methods were used together and the classification method was Random Forest, indicating that the preferred classification method and conversion methods are effective when compared to other classifiers in the literature.**

I. GİRİŞ

İnsanlık tarihinin en önemli buluşlarından biri olan arabanın icat edilmesinden bugüne kadar arabalar pek çok değişim geçirmiştir. Bu veri setimizde hangi araba özelliklerinin daha çok değişime uğradığı, hangi özelliklerin daha çok tercih edildiğini kullanacağımız yöntemlerle bunları değerlendireceğiz. Weka'da kullandığımız yöntemlerden Random Forest, J48, Random Tree, REP tree, Decision Stump yöntemlerinin arasından en çok anlık doğruluk yüzdesini Random Forest (Rastgele Orman) algoritmasi ile %93,64 oranında bir değer bulmuş olmaktayız. Bu çalışmada veri kümesi 1528 örneği oluşur ve her kayıt yedi özellikleri içerir.Bunlar; fiyat satın alma, bakım fiyatı, kapı sayısı, taşımak için kişi açısından kapasite, bagaj çizme boyutu, arabanın tahmini emniyet ve araç kabul edilebilirliği. Araç kabul edilebilirliğinin niteliği, müşterilerin kabul ettiği aracın derecesini sınıflandırmak için kullanılan bir sınıf etiketidir, diğer nitelikler ise tahmin değişken olarak görülmektedir .

Attribute name	Description	Domain
b_price	Buy Price	v-high / high / med / low
m_price	Repair price	v-high / high / med / low
door	The number of the door	2 / 3 / 4 / 5-more
person	The number of passenger	2 / 4 / more
size	Suitcase capacity	Small / med / big
safety	Safety evalution	Low / med / high
class	level of customer acceptance	Unacc / acc / good / vgood

Tablo 1.Veri seti detayları

Bu altı nitelikleri, kötü kategori olsa bile, aynı zamanda bazı müşteriler kabul edebilir, örneğin bagaj kapasitesi (boyutu) küçük olsa bile, hala "iyi" sınıflandırmak olabilir. Ama oldukça özel olan iki özellik vardır, bunlar güvenlik ve kişi sayısıdır."Kişi " değeri 2'dir, sınıfın tümü unacc'dir. Güvenlik özniteliği değeri düşük, sınıfın tümü unacc'dir. Bu nedenle araba seçtiğimiz zaman bu özellikler müşteri için çok önemlidir. İki bölüme veri kümesi ayrılmış olup, bir bölüm eğitim veri kümesi tahmin modeli üretmek için kullanılır ve diğer bölüm test veri modeli doğruluğunu test etmek için set olarak kullanılır. Veri kümesinden gelen tüm veriler eğitim veri kümesine ve test veri kümesine seçilebilir anlamına gelir çapraz doğrulama yöntemi (Cross Validation) kullanıldı.

II. MATERYAL VE YÖNTEM

A. Veriseti Tanıtımı

Yapılan çalışmada kullanılan veri kümesi Car Evaluation veri seti kullanılmıştır. Modeli oluşturmak için kullandığımız veri madenciliği yöntemi sınıflandırmadır. Bu sınıflandırma için Weka progamını seçtik. Sınıflandırıcılar içersinden kulandığımız algoritmalardan (Trees) sınıflandırma yöntemi, daha iyi sonuç gösterdi. V 10-kat çapraz doğrulama yöntemi kullanıldı. Tablo 1'de kullandığımız veri setinin özelliklerine göre nasıl ayrıştığını göstermektedir. Veri setimizde 1528 örnek yer almaktadır. Örneğin, araba için en önemli olabilecek güvenlik özelliğinin yüksek, düşük ,orta seçimleri bulunmaktadır.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1446               94.6335 %
Kappa statistic                          0.8747
Mean absolute error                      0.0728
Root mean squared error                  0.1568
Relative absolute error                 34.1628 %
Root relative squared error             48.0927 %
Total Number of Instances             1528
```

Tablo.2 Anlık doğruluk yüzdesi

Yukarıdaki tabloda Rastgele Orman (Random Forest) algoritması ile kullanılmış anlık doğrulama sınıflandırma yüzdesi verilmiştir. Yüzdemiz %94,6335 olarak elde edilmiştir ve bu da tercih edilen sınıflandırma yöntemi ve dönüşüm yöntemlerinin literatürdeki diğer sınıflandırıcılara bakıldığında etkili olduğunu göstermektedir.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure
               0,972     0,030     0,988       0,972    0,980
               0,950     0,047     0,861       0,950    0,904
               0,387     0,003     0,706       0,387    0,500
               0,622     0,006     0,719       0,622    0,667
Weighted Avg.  0,946     0,033     0,946       0,946    0,945
```

Tablo 3. RO sınıfına göre ayrıntılı doğruluk

Yukarıdaki tabloda Rastgele Orman sınıfına göre belirlenmiş doğruluk değerleri verilmiştir.

B. Sınıflandırma

Çalışmada Rastgele Orman (Random Tree) algoritması ile en iyi başarı elde edilmiştir. Bu sınıflandırıcının yanında Bayes Net, Random Tree, REP Tree yani Karar Ağacı algoritma sınıflandırıcıları ile de performans incelenmesi yapılmıştır. 10 kat çapraz (K Fold Cross Validation) onaylama ile kullanılan özniteliklerin başarısı incelenmiştir. Çapraz onaylamada öncelikle eğitim verileri rastgele 10 eşit parçaya bölünmüş ve bu 10 parçadan bir tanesi onaylama verileri olarak kullanılmış geri kalan veriler ise eğitim verileri olarak kullanılmıştır. Her bir parça onaylama kümesi olacak şekilde parçalar taranmış ve her seferinde en yüksek sınıflandırma doğruluğunu verecek parametre elde edilmeye çalışılmıştır. Bu adımlar 100 kere tekrarlanarak optimum parametre elde edilmiş ve test verilerinin sınıflandırılmasında kullanılmıştır.

C. K Fold Cross Validation

Veri madenciliği çalışmalarında, uygulanan yöntemin başarınının sınanması için, veri kümesini eğitim ve test kümeleri olarak ayırılmaktadır. Bu ayırma işlemi çeşitli şekillerde yapılabilir. Örneğin veri kümesinin %66'lık bir kısımını eğitim %33'lük bir kısmını ise test için ayırmak ve eğitim kümesi ile sistem eğitildikten sonra test kümesi ile başarısının sınanması kullanılabilecek yöntemlerden birisidir. Bu eğitim ve test kümelerinin rastgele olarak atanması da farklı bir yöntemdir. K katlamalı çarpraz doğrulama için k-kere yöntem çalıştırılır. Her adımda veri kümesinin 1/k kadar, daha önce test için kullanılmamış parçası, test için kullanılırken, geri kalan kısmı eğitim için kullanılır.Buna göre her aslında yapılan işlem aşağıdaki şekilde daha resmi bir gösterimle yazılabilir:

$$t_i \in VK \text{ olmak üzere }, Sonuç = \frac{\sum_{i=0}^{k} SF(t_i, VK - t_i)}{k}$$

Buradaki S F(test, eğitim), sınıflandırma fonksiyonu , VK, veri kümesi, k , kaç parça katlama kullanıldığı ve t ise veri kümesi üzerinden seçilen her bir test kümesi olarak verilmiştir. Yukarıda formülize edildiği üzere, sonuç bütün sınıflandırma fonksiyonlarının performanslarının toplamının, k sayısına bölünerek ortalaması alınmasıdır

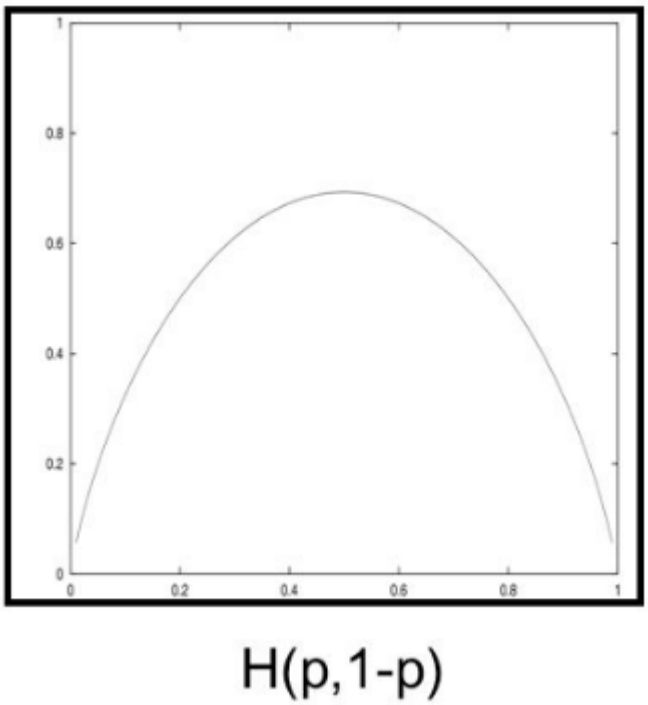
C. ID3 Karar Ağacı Algoritması

ID3 bir Karar Ağacı algoritmasıdır. ID3 algoritması sınıflandırmak için en ayırıcı özelliğe sahip olan entropi kavramından yararlanır. Entropi kavramı beklentisizliğin maksimumlaştırılmasıdır. Bir veri kümesindeki rastgeleliği, belirsizliği ölçmek için kullanılır. Eğer elimizdeki veriler tek bir sınıfa ait olsaydı entropi 0 olurdu. Örneğin herkes aynı siyasi partiye oy vermiş olursa elimizde tek bir sınıf olduğu için entropi değeri sıfır olacaktı. Entorpi 0 – 1 aralığında bir değer alır. Bütün olasılıklar eşit olduğunda ise entropi değeri maksimum değere ulaşır. Entropi matematiksel olarak şöyle ifade edilebilir:
$< P_1, P_2 \dots P_n >$ olasılıkları ifade ederse, tüm bu olasılıkların toplamı 1 olmak zorundadır. Bu durumda entropi aşağıdaki gibi olacaktır:

$$\sum_{i=1}^{n} P_i = 1$$

$$H(P_1, P_2 \dots P_n) = \Sigma(P_i \log(1/P_i)) \quad (1)$$

Bir veri setinin tamamının entropisi hesaplanır; fakat bu veri seti çeşitli alt bölümlere ayrılıyorsa her alt bölümün de entropisi hesaplanır. Buradaki H(), veri setindeki herhangi bir durumdaki halini temsil eder.



Şekil 4. Veri setindeki herhangi bir durum

ID3 algoritması veri setini bölmeden önce doğru sınıflandırma yapmak için gelen bilgiyle, veri seti bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanarak, öncelikle düğüme ve dallanmalara karar verir. Aradaki bu farka kazanım denir. Veri seti bölündükçe ve dallanmalar oluştukça doğru sınıflandırma için gerekli bilgi sayısı da azalacaktır. Kazanım şu şekilde hesaplanır.

$$Kazanım(D;S) = H(D) - \sum_{i=1}^{n} P(D_i)H(D_i) \quad (2)$$

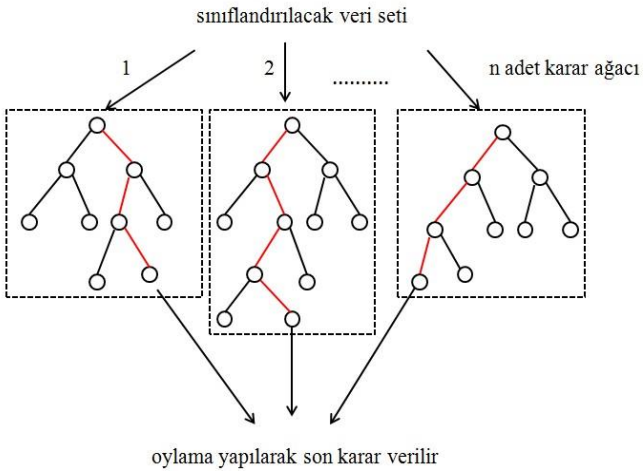
1) Rastgele Orman (Random Tree)

Rastgele Orman (RO) sınıflandırıcısı Breiman [1] tarafından ortaya atılan, karar ağaçları ve düğümlerden oluşan bir sınıflandırıcıdır. Bu sınıflandırıcı, düğümleri dallarına ayırırken her düğümde rastgele olarak seçilmiş değişkenlerden en iyi olanını kullanır.

Oluşturulan karar ağaçlarının (Out of Bag, OOB) iç hatalarına göre her bir ağaca belirli bir ağırlık verilir. En düşük hataya sahip karar ağacı en yüksek, en yüksek hataya sahip karar ağacı ise en düşük ağırlığa sahip olur. Bu ağırlıklara göre yapılan sınıf tahmininde oy verme işlemi gerçekleştirilir. Sonrasında oylar toplanarak en son karar verilir.
Rastgele orman sınıflandırıcısının adımları aşağıda verilmiştir:

- Veri setinin öznitelikleri kullanılarak oluşturulacak karar ağacı adedi (n) belirlenir.
- Karar ağaçları içersindeki her düğümde rastgele m adet değişken seçilir ve en iyi dal belirlenir. (gini indeksi ile hesaplama yapılarak
- Belirlenen en iyi dal tekrar iki alt dala ayrılır ve gini indeksi sıfıra ulaşıncaya kadar yani her bir yaprak düğümde bir sınıf kaldığında ağaç dallanma işlemi sonlanır [2].
- n karar ağacının ayrı ayrı yaptığı tahminler arasında en çok oyu alan sınıf son karar tahmini olarak seçilir.

Şekil 2'de RO sınıflandırıcısının genel yapısı verilmiştir.



Şekil 5. RO Sınıflandırıcısının Genel Yapısı

Bu sınıflandırıcıda en önemli seçim, her düğümde kullanılacak değişken sayısı (m) ve geliştirilecek ağaç sayısı (n) parametrelerinin seçimidir. Breiman [3]'e göre, m parametresi değişken sayısının karekköküne eşit alındığında en kararlı sonuç elde edilmektedir.

2) Navie Bayes:

Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes'den alan bir sınıflandırma/ kategorilendirme algoritmasıdır. Naïve Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar. Naïve Bayes sınıflandırma yönteminin birçok kullanım alanı bulunabilir fakat, burada neyin sınıflandırıldığından çok nasıl sınıflandırıldığı önemli. Yani öğretilecek veriler binary veya text veriler olabilir, burada veri tipinden ve ne olduğundan ziyade, bu veriler arasında nasıl bir oransal ilişki kurduğumuz önem kazanıyor.

3) Doğrusal Analiz Ayırıcı: Bu yöntemde ise sınıflar arasındaki dağılım maksimum olurken sınıf içi varyansı minimum yapacak doğru bulunmaya çalışılır. Verilerin normal dağılıma sahip olduğu varsayılmaktadır.

III. SONUÇLAR VE TARTIŞMA

Çalışma sonuçları incelendiğinde Rastgele Orman (Random Forest) algoritmasının model testine ait %94,6365 doğruluk derecesiyle en iyi sonucu ürettiği ve Tablo 2'de de gösterilerek en iyi başarı bu algoritma ile elde edilmiştir. Yukarıda da belirtildiği gibi, "güvenlik" özelliği çok önemlidir. Böylece, "güvenlik" değeri düşükse, sonuç doğrudan kabul edilemez (unacc) oluyor. Ve güvenlik değeri ne olursa olsun, eğer "kişinin değeri 2 ise, giriş de kabul edilemez doğrudan oda (unacc) yani kabul edilemez olacak.

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.

KAYNAKLAR

[1] 12- Breiman, L. "Random forests," *Machine learning*, 45(1):5–32, 2001

[2] 13- Pal M., "Random forest classifier for remote sensing classification", *International Journal Of Remote Sensing*, 26(1): 217-222, 2005.

[3] 15- Breiman, Leo. "Manual on setting up, using, and understanding random forests v3. 1." *Statistics Department University of California Berkeley*, CA, USA, 2002.

[4] 14- Watts J. D.,Powell S. L., Lawrence R. L., Hilker T., "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery", *Remote Sensing of Environment*, 115(1): 66–75, 2011.

[5] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[6] 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

[7] Kaufmann, Fourth Edition, 2016.