

Makine Öğrenmesi Algoritmalarının Farklı Veri Setleri Üzerinde Karşılaştırılması

Comparison of Machine Learning Algorithms on Different Datasets

Yazarlar Gizlenmiştir

Özetçe —Makine öğrenmesi algoritmaları verileri sınıflandırmak için kullanılan yöntemlerdir. Bu çalışmanın amacı, farklı veri setleri üzerinde makine öğrenmesi algoritmalarının karşılaştırılmasıdır. Bu çalışma için 9 farklı makine öğrenmesi algoritması WEKA yazılımında 10 katlı çapraz doğrulama tekniği ile 3 farklı veri seti üzerinde sınıflandırılmıştır. Sınıflandırma sonucuna göre, yüksek doğruluk değerine sahip makine öğrenmesi algoritması 3 veri seti için farklı çıkmıştır. Car Evaluation veri seti için Multilayer Perceptron algoritması, Image Segmentation veri seti için Random Forest algoritması ve User Knowledge Modeling veri seti için Simple Logistic algoritması elde edilmiştir.

Anahtar Kelimeler—veri seti, makine öğrenmesi, sınıflandırma, WEKA.

Abstract—Machine learning algorithms are methods used to classify data. Aim of this study is comparison of machine learning algorithms on different dataset. For this study, 9 different machine learning algorithms with 10 fold cross validation method in WEKA is classified on 3 different datasets. As a result of classification, machine learning algorithm which has high accuracy rate is different for 3 datasets. Multilayer Perceptron algorithm for Car Evaluation dataset, Random Forest algorithm for Image Segmentation dataset and Simple Logistic algorithm for User Knowledge Modeling dataset were obtained.

Keywords—dataset, machine learning, classification, WEKA.

I. GİRİŞ

Günümüzde verilerin işlenmesi ve sınıflandırılması makine öğrenmesi yöntemleri ile kolaylık kazanmıştır. Bu çalışmada, makine öğrenmesi algoritmalarının farklı veri setleri üzerinde etkisi incelenmiştir. Makine öğrenmesi algoritmaları olarak Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve Logistic Model Trees(LMT) kullanılmıştır. Ayrıca bu makine öğrenmesi algoritmaları WEKA yazılımında 10 katlı çapraz doğrulama tekniğine göre sınıflandırılmıştır. Veri seti olarak ise, UCI Machine Learning Repository¹ websitesinden alınan

Car Evaluation, Image Segmentation ve User Knowledge Modeling veri setleri kullanılmıştır.

Makine öğrenmesi algoritmalarının farklı veri setleri üzerinde sınıflandırılmasıyla elde edilen sonuçlara göre; Car Evaluation veri setinde doğruluk değeri %99.537 ile Multilayer Perceptron algoritması, Image Segmentation veri setinde %98.0519 ile Random Forest algoritması ve User Knowledge Modeling veri setinde %95.7816 ile Simple Logistic algoritması elde edilmiştir.

Bildirinin devamında, bu konuya benzer çalışmaların yer aldığı İlgili Çalışma bölümü; WEKA yazılımı, bildiride kullanılan veri setleri, algoritmalar ve metriklerin tanıtıldığı Yöntemler bölümü; çalışmada elde edilen sonuçların yer aldığı DeneySEL Sonuçlar bölümü ve son olarak da Sonuç bölümü bulunmaktadır.

II. İLGİLİ ÇALIŞMA

Bilgin [1] makine öğrenmesi yöntemlerini 10 parça çapraz doğrulama tekniği ile 6 farklı biyomedikal veri seti üzerindeki performans analizini incelemiştir. Bilgin [1] makine öğrenmesi algoritması olarak Naive Bayes, Random Forest, Sequential Minimal Optimization(SMO), IBk(k-nearest neighbor), Decision Table, J48 kullanmıştır ve SMO algoritması ile daha yüksek doğruluk değeri elde etmiştir.

Akçetin vd. [2] ADTree, BFTree, Fonksiyonel Ağaçlar, J48, J48 Aşı, LADTree, LMT, NBTree, Rastgele Orman Karar Ağacı, Rastgele Ağaç, REPTree Sınıflandırıcısı, SimpleCART algoritmaları 10 katlı çapraz doğrulama ile UCI elektronik posta veri seti üzerinde sınıflandırmıştır ve Rastgele Orman karar ağacı %94.68 ile en iyi sınıflandırma başarısı elde etmiştir.

Caruana vd. [3] SVMs, neural nets, logistic regression, naivebayes, memory based learning, random forests, decision trees, bagged trees, boosted trees ve boosted stumps algoritmalarını 11 tane veri seti üzerinde sınıflandırmışlardır. Ayrıca Platt Scaling ve Isotonic Regression yöntemlerini de kullanarak bir karşılaştırma yapmışlardır. Bu iki yöntemin kullanılmasıyla en iyi algoritma boosted trees algoritması elde edilmiştir,

¹<https://archive.ics.uci.edu/ml/datasets.html>

normal sınıflandırmanın yapılmasıyla en iyi algoritma bagged trees algoritması elde edilmiştir.

Kaynar vd. [4] Naive Bayes, Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları ve Destek Vektör Makineleri yöntemlerini film yorumları veri seti üzerinde kullanarak sonuçları karşılaştırmışlardır. Eğitim ve test verileri için Yapay Sinir Ağları ve Destek Vektör makineleri yöntemlerinde daha iyi sonuç elde etmişlerdir.

III. YÖNTEMLER

A. WEKA

WEKA [9], veri madenciliği çalışmaları için kullanılan, makine öğrenmesi algoritmalarını içeren açık kaynak kodlu bir yazılımdır.² Bu çalışmada, veri setleri WEKA yazılımında farklı makine öğrenmesi algoritmalarıyla sınıflandırılmıştır.

B. Veri Seti

Çalışma için üç farklı veri seti kullanıldı. Bunlar Tablo I’de de görüldüğü gibi Car Evaluation [7], Image Segmentation [7] ve User Knowledge Modeling [8] veri setleridir.

veri seti	özellik sayısı	örnek sayısı
Car Evaluation	7	1728
Image Segmentation	20	2310
User Knowledge Modeling	6	403

TABLO I: Veri seti detayları

Car Evaluation veri seti³: Bu veri seti araba ile ilgili özellikleri kapsar. Örneğin, arabanın kapı sayısı, kişi kapaitesi, güvenilirliği gibi özellikleri içerir. Bu veri setinde 1728 örnek vardır.

Image Segmentation veri seti⁴: Bu veri seti 7 dış mekan resimlerinden rastgele çekilen örneklerden oluşur. Bu veri setinde 20 özellik ve 2310 örnek vardır.

User Knowledge Modeling veri seti⁵: Yazarların sezgisel bilgi sınıflandırıcı, k-en yakın komşu algoritmasını kullanarak kullanıcıların bilgi sınıfını sınıflandırmasıyla ilgili veri setlerini kapsar. Bu veri seti 6 özellik ve 403 örnekten oluşur.

C. Algoritmalar ve Metrikler

Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve LMT algoritmaları WEKA yazılımında varsayılan özellikler kullanılarak sınıflandırıldı.

- **Multilayer Perceptron:** Öğrenme oranı(learning rate) 0.3 ve gizli katmanda 5 düğüm kullanıldı.
- **Naive Bayes:** Denetimli ayrıştırma(supervised discretization) *false* kullanıldı.
- **Logit Boost:** Sınıflandırıcı olarak *Decision Stump* algoritması kullanıldı.

- **J48:** Güven faktörü(confidence factor) 0.25 kullanıldı.
- **Random Forest:** Bag size percent değeri olarak 100 kullanıldı.
- **Bayes Net:** Tahmin(estimator) algoritması olarak *Simple Estimator* algoritması ve arama(search) algoritması olarak *K2* algoritması kullanıldı.
- **Simple Logistic:** Boosting iteration sayısı 0 olarak kullanıldı.
- **Bagging:** Sınıflandırıcı olarak *REPTree* algoritması kullanıldı.
- **LMT:** Boosting iteration sayısı -1 olarak kullanıldı.

Bu çalışmada kullanılan metrikler aşağıda verilmiştir.

- **Doğruluk(accuracy) [4]:** Doğru olarak sınıflandırılan örneklerin toplam örnek sayısına oranıdır.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Kesinlik(precision) [4]:** Pozitif olarak etiketlenen örneklerin sayısının pozitif olarak sınıflandırılan toplam örneklere oranıdır.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

- **Duyarlılık(recall) [4]:** Pozitif olarak etiketlenen örneklerin gerçekten pozitif olan örneklerin toplam sayısına eşittir.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

- **F-ölçütü(f-measure) [4]:** Kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır.

$$f - measure = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

- **Ortalama mutlak hata(mean absolute error) [6]:** Mutlak hataların ortalamasıdır. Hata, tahmin edilen değer ile gerçek değer arasındaki farktır.

Tablo II’ye göre [5], TP(true positive) ve FN(false negative) değerleri doğru ve yanlış örneklerin doğru şekilde sınıflandırıldığını gösterir; FP(false positive) ve TN(true negative) değerleri ise doğru ve yanlış örneklerin doğru şekilde sınıflandırılmadığını gösterir.

	Beklenen Pozitif	Beklenen Negatif
Gerçek Doğru	TP	FN
Gerçek Yanlış	FP	TN

TABLO II: Karışıklık Matrisi [5]

²<https://www.cs.waikato.ac.nz/ml/weka/>

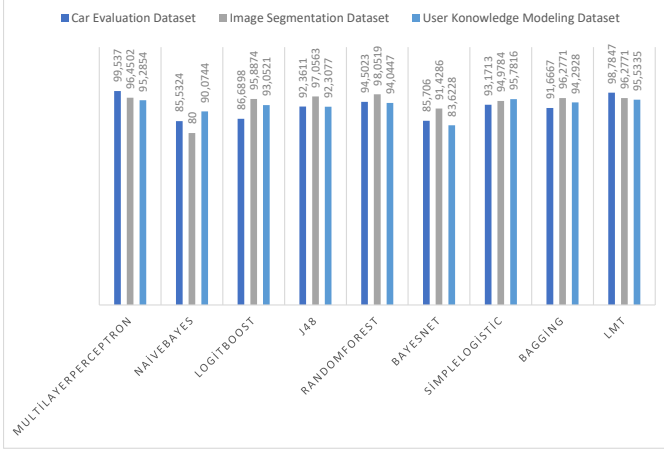
³<https://archive.ics.uci.edu/ml/datasets/Car+Evaluation>

⁴<https://archive.ics.uci.edu/ml/datasets/Image+Segmentation>

⁵<https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling>

IV. DENEYSEL SONUÇLAR

Bu çalışmada, Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve LMT algoritmalarının farklı veri setleri üzerinde etkisi incelenmiştir. Bu çalışmaya göre elde edilen doğruluk değerleri Şekil 1’de gösterilmiştir.



Şekil 1: Doğruluk değerleri

Car Evaluation veri seti için en yüksek doğruluk değeri **%99.537** ile Multilayer Perceptron algoritması, en düşük doğruluk değeri **%85.5324** ile Naive Bayes algoritmasıdır. Image Segmentation veri seti için en yüksek doğruluk değeri **%98.0519** ile Random Forest algoritması, en düşük doğruluk değeri **%80** ile Naive Bayes algoritmasıdır. User Knowledge Modeling veri seti için en yüksek doğruluk değeri **%95.7816** ile Simple Logistic algoritması, en düşük doğruluk değeri **%83.6228** ile Bayes Net algoritmasıdır. Sonuç olarak, doğruluk değerlerinin yüksek olduğu algoritmalar üç veri seti için de farklıdır; fakat, doğruluk değerlerinin düşük olduğu algoritmalar Car Evaluation veri seti ve Image Segmentation veri seti için aynıdır.

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	0.995	0.995	0.995
NaiveBayes	0.852	0.855	0.847
LogitBoost	0.869	0.867	0.861
J48	0.924	0.924	0.924
RandomForest	0.946	0.945	0.945
BayesNet	0.854	0.857	0.849
SimpleLogistic	0.933	0.932	0.932
Bagging	0.919	0.917	0.916
LMT	0.988	0.988	0.988

TABLO III: *Car Evaluation* veri seti için sonuçlar

Tablo III’de Car Evaluation veri seti için elde edilen kesinlik, duyarlılık ve f-ölçütü değerleri verilmiştir. Bu tabloya göre, Multilayer Perceptron algoritması **0.995** değeri ile en yüksek kesinlik, duyarlılık ve f-ölçütü değerine sahiptir. Naive Bayes algoritması ise **0.852**, **0.855** ve **0.847** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	0.965	0.965	0.964
NaiveBayes	0.817	0.800	0.779
LogitBoost	0.959	0.959	0.959
J48	0.971	0.971	0.971
RandomForest	0.981	0.981	0.980
BayesNet	0.914	0.914	0.914
SimpleLogistic	0.951	0.950	0.950
Bagging	0.963	0.963	0.963
LMT	0.963	0.963	0.963

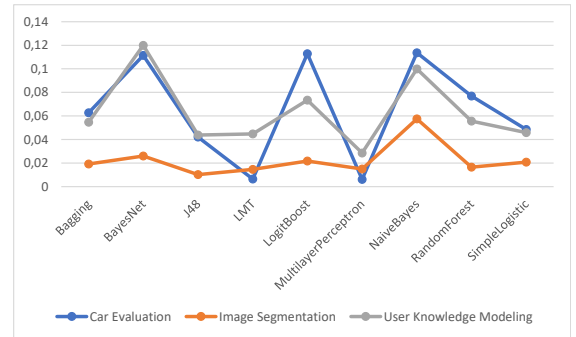
TABLO IV: *Image Segmentation* veri seti için sonuçlar

Tablo IV’te Image Segmentation veri seti için elde edilen kesinlik, duyarlılık ve f-ölçütü değerleri verilmiştir. Bu tabloya göre, Random Forest algoritması **0.981**, **0.981** ve **0.980** değerleri ile sırasıyla en yüksek kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir. Naive Bayes algoritması ise **0.817**, **0.800** ve **0.779** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	0.953	0.953	0.953
NaiveBayes	0.908	0.901	0.901
LogitBoost	0.931	0.931	0.931
J48	0.923	0.923	0.923
RandomForest	0.942	0.940	0.940
BayesNet	0.842	0.836	0.837
SimpleLogistic	0.958	0.958	0.958
Bagging	0.945	0.943	0.943
LMT	0.956	0.955	0.955

TABLO V: *User Knowledge Modeling* veri seti için sonuçlar

Tablo V’te User Knowledge Modeling veri seti için elde edilen kesinlik, duyarlılık ve f-ölçütü değerleri verilmiştir. Elde edilen bu sonuçlara göre, Simple Logistic algoritması **0.958** değeri ile en yüksek kesinlik, duyarlılık ve f-ölçütü değerine sahiptir. Bayes Net algoritması ise, **0.842**, **0.836** ve **0.837** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.



Şekil 2: Ortalama mutlak hata değerleri

Şekil 2’de veri setlerine göre ortalama mutlak hata değerleri verilmiştir. Buna göre, Car Evaluation veri setinde Multilayer Perceptron algoritması en düşük değere sahip ve değeri **0.0062**, Image Segmentation veri setinde J48 algoritması en düşük değere sahip ve değeri **0.0102**, User Knowledge Modeling veri setinde Multilayer Perceptron algoritması en düşük değere sahip ve değeri **0.0286**’dır.

V. SONUÇ

Makine öğrenmesi algoritmalarının farklı veri setleri üzerinde karşılaştırılmasıyla farklı sonuçlar elde edilmiştir. Bu çalışma için kullanılan algoritmalar; Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve Logistic Model Trees(LMT) algoritmalarıdır. Kullanılan veri setleri ise UCI Machine Learning Data Repository websitesinden alınan Car Evaluation, Image Segmentation ve User Knowledge Modeling veri setleridir.

Sınıflandırma sonucuna göre, Multilayer Perceptron algoritması Car Evaluation veri setinde %99.537 doğruluk değeri ile, Random Forest algoritması Image Segmentation veri setinde %98.0519 doğruluk değeri ile, Simple Logistic algoritması User Knowledge Modeling veri setinde %95.7816 doğruluk değeri ile en yüksek değere sahip algoritmalarıdır.

KAYNAKLAR

- [1] M. BİLGİN, "Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi," Breast, vol. 2, no. 9, pp. 683–688.
- [2] E. Akçetin and U. Çelik, "The Performance Benchmark of Decision Tree Algorithms for Spam e-mail Detection," J. Internet Appl. Manag., vol. 5, no. 2, pp. 43–56, 2014.
- [3] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," Proc. 23rd Int. Conf. Mach. Learn., vol. C, no. 1, pp. 161–168, 2006.
- [4] O. Kaynar, Y. Görmez, M. Yıldız, and A. Albayrak, "Makine Öğrenmesi Yöntemleri ile Duygu Analizi Sentiment Analysis with Machine Learning Techniques," no. September, 2016.
- [5] <https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance>
- [6] <https://docs.microsoft.com/tr-tr/azure/machine-learning/studio/create-experiment>
- [7] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- [8] H. T. Kahraman, Sagioglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, pp. 283-295, 2013.
- [9] 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.