BAZI DENETİMLİ ÖĞRENME ALGORİTMALARININ R PROGRAMLAMA DİLİ İLE KIYASLANMASI

COMPARISON OF SOME SUPERVISED LEARNING ALGORITHMS R PROGRAMMING LANGUAGE

НЕКОТОРЫЕ КОНТРОЛИРУЕМОГО ОБУЧЕНИЯ АЛГОРИТМОВ НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ R ПО СРАВНЕНИЮ С

Yusuf Murat KIZILKAYA[*]-Ayşe OĞUZLAR[**]

ÖZ

Yapay zekâ bilgisayarların insanların düşünce sistemlerini taklit ederek karmaşık problemlere çözüm üretebilme yeteneklerine verilen addır. Makine öğrenmesi ise yapay zekânın önemli bir alt dalıdır. Makine öğrenmesi, çeşitli görevlerin öğrenilmesi, mantıksal ve ikili çıkarımlar yoluyla otomatik hesaplama yöntemlerini kapsayan bir süreç olarak ele alınabilir. R yazılımı pek çok istatistiksel hesaplamanın yanı sıra makine öğrenmesi algoritmasında ki başarısıyla da ön plana çıkmaktadır. Bu çalışmada R yazılımının sınıflandırma amacıyla kullandığı çeşitli makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. Bu amaçla, UCI Makine Öğrenme Havuzundan, elde edilen gerçek verilere çeşitli makine öğrenme algoritmaları uygulanmış ve sınıflandırma algoritmaları birkaç kriter kullanılarak karşılaştırılmıştır. Hesaplanan kriterlerden olan; kesinlik, doğruluk, duyarlılık ve F ölçütü hareketle, sınıflandırma tekniklerinin kıyaslanması yapılmıştır. Yapılan karşılaştırmalar sonucunda, üç kriterde en iyi sınıflandırmayı yapan Lojistik Regrasyon algoritmasının diğer algoritmalara göre daha başarılı olduğu görülmüştür. Tüm ölçütlerden en iyi ikinci performansı gösteren algoritma Navie Bayes algoritması olmuştur.

Anahtar Kelimeler: Makine Öğrenmesi, Denetimli Öğrenme, R Programlama, Lojistik Regresyon, Navie Bayes.

ABSTRACT

Artificial intelligence is given to computers' ability to imitate people's thought systems and produce solutions for complex problems. Machine learning is an

[*]Öğr. Gör., Ardahan Üniversitesi SBMYO.
(muratkizilkaya@ardahan.edu.tr)
[**] Prof. Dr., Uludağ Üniversitesi İ.İ.F.B.
(ayseog@uludag.edu.tr)
DOI: 10.17498/kdeniz.405746

important subdivision of artificial intelligence. Machine learning can be viewed as a process involving the learning of various tasks and automatic calculation methods through logical and binary inferences. R programming comes to the forefront with its success in machine learning algorithm as well as many statistical calculations. In this study, the performances of various machine learning algorithms used by R programming for classification purposes are compared. For this purpose, various machine learning algorithms have been applied to real data obtained from UCI Machine Learning Pool and classification algorithms have been compared using several criteria. The calculated criteria are; precision, accuracy, sensitivity, and classification techniques based on the F-measure. As a result of these comparisons, it is seen that Logistic Regulation algorithm, which makes the best classification in the three criteria, is more successful than the other algorithms. The algorithm that has the second best performance of all criteria has been the Navie Bayes algorithm.

Key words: Machine Learning, Supervised Learning, R programming, Logistic Regression, Navie Bayes.

АННОТАЦИЯ

Искусственный интеллект дает способность компьютеров подражать системам мышления людей и выработать решения сложных проблем. Машиноведение является важным подразделением искусственного интеллекта. Машиноведение можно рассматривать как процесс, включающий изучение различных задач и автоматических методов расчета посредством логических и двоичных выводов. R-программирование выходит на первый план с его успехом в алгоритме машинного обучения, а также во многих статистических расчетах. В этом исследовании сравниваются характеристики различных алгоритмов машинного обучения, используемых программированием R для целей классификации. С этой целью различные реальные алгоритмы машинного обучения были применены к реальным данным, полученным из UCI пула машинного обучения, и алгоритмы классификации были сопоставлены по нескольким критериям. Вычисленные критерии: Сравнивались точность, точность, чувствительность и методы классификации на основе F-меры. В результате этих сравнений видно, что алгоритм логистической регуляции, который делает лучшую классификацию по трем критериям, более успешным, чем другие алгоритмы. Алгоритм, который имеет вторую лучшую производительность по всем критериям, был алгоритмом Наивного Байеса.

Ключевые слова: машинное обучение, контролируемого обучения, программирование на R, логистическая регрессия, Наивный Байес.

1.Giriş

Makine öğrenmesi, insanlar tarafından kolaylıkla anlaşılabilecek, basit sınıflandırıcı ifadeler üretmeyi amaçlar. Bunu yaparken de arka planda istatistiksel yöntemleri kullanır (Michie, Spiegelhalter ve Taylor, 1994:3). Makine öğrenmesi

sayesinde; insan tarafından yapılmaya kalkışıldığında çok uzun sürede yapılabilecek veya insanlar tarafından hesaplamanın imkânsız bir hal alabileceği bir durum, bilgisayarlar tarafından çok kısa sürede kolaylıkla yapılabilmektedir. Makine öğrenmesini günümüzde popüler hale getiren en önemli unsur şüphesiz ki bilgisayar teknolojilerinde meydana gelen ilerlemelerdir. Bu sayede artık ihtiyaç duyulan veriler rahatlıkla depolanabilmekte, rahatlıkla erişilebilmekte ve analiz için kullanılabilmektedir.

Makine öğrenmesi sayesinde, önceki tecrübelere veya örnek veri setlere dayanan bir işlemi optimize etmek için bilgisayarlar programlanabilir. İstenen sınıflandırmalar bilgisayarda kısa sürede ve etkili bir şekilde yapılabilir, bu süreçler sonunda bir model oluşturulur ve bu model geleceğe yönelik öngörülerde bulunabilir, denetim amacıyla kullanılabilir.

Makine öğrenmesi tüm üç farklı yöntemle çalışabilmektedir. Bunlar; denetimli (gözetimli) öğrenme, denetimsiz (gözetimsiz) öğrenme ve yarı denetimli öğrenmedir.

Denetimli öğrenme; sisteme eğitim veri seti ve test veri setinin yüklenmesi, veri setinde her bir veri için gerekli etiketlenmenin yapılması ve bu sayede girdi veri seti ile çıktı veri seti arasında ilişki kurulması mantığına dayanır. Temel amaç sonuçları bilinen veri setinden yapılan sınıflandırmadan hareketle sonuçları bilinmeyen veri setine dair etkili tahminler yapabilmektir (Aydın ve Özkul, 2015:38).

Denetimsiz öğrenme de ise kullanıcının sisteme herhangi bir müdahalesi yoktur. Sadece girdi verileri sisteme verilir ancak herhangi bir işaretleme yapılmaz. Sistem otomatik olarak keşifler yapar, ilişki ağını ortaya koymaya çalışır (Alpaydın, 2010:11).

Elde az sayıda etiketlenmiş veri buna karşın çok daha fazla sayıda etiketlenmemiş veri varsa denetimli öğrenme de denetimsiz öğrenme de yetersiz kalabilir. Bu durumda en ideal yöntem az sayıdaki etiketlenmiş veriden hareketle etiketlenmemiş veriler hakkında bilgi sahibi olmaya çalışmak, onları sınıflandırmaktır. Bu yönteme de yarı denetimli öğrenme denir. Denetimli öğrenme ile arasında ki en temel fark eldeki etiketlenmiş veri kümesidir. Denetimli öğrenmede etiketlenmiş veri sayısı fazla, tahmin edilmek istenen veri sayısı azken yarı denetimlide tam tersi bir durum söz konusudur.

2.Literatür

Makine öğrenmesinin çok geniş bir kullanım alanı olduğundan, literatür tarandığında hemen her konuda makine öğrenmesine dair çalışmalarla karşılaşmak mümkündür. Hangi alanda olursa olsun yapılan bu çalışmaların en büyük özelliği belli başlı sınıflandırma yöntemlerinin ön plana çıkmasıdır. Sağbaş ve Ballı (2016) yaptıkları çalışmada akıllı telefonların algılayıcıları üzerine 6 farklı makine öğrenmesi algoritmasını test etmiş, Random Forest algoritmasının diğerlerinden daha iyi sonuç verdiğini ortaya koymuştur.

Sevindi, yapmış olduğa çalışmada, Türkçe film yorumlarından hareketle duygu yönünü tespit etmeye çalışmış, makine öğrenmesi yöntemlerinden C4.5 karar

ağacı, Navie Bayes ve Destek Vektör Makineleri (Support vector Machines-SVM) algoritmalarını kullanmıştır. Yapılan çalışmada en iyi sonucu 0,8258 ile SVM sınıflandırıcısı vermiştir (Sevindi, 2013).

Makine öğrenmesi ulaşım türü tespitinde de sıklıkla kullanılan bir araç olmuştur. Lara ve diğerleri (2012) yaptıkları çalışmada Lojistik Regresyonun insan eylemelerini tanımlamada en iyi sınıflandırıcı olduğunu ortaya koymuştur.

Wu ve diğerleri (2009) yaptıkları çalışmada Karar Ağaçları C4.5 ile Destek Vektör Makinelerinin performanslarını kıyaslamış, Karar Ağaçlarının daha iyi sonuç verdiğini ortaya koymuştur.

Firmalarının başarısızlıkları üzerine yapılan çalışmalarda da makine öğrenmesinin kullanıldığı görülmektedir. Charitou ve diğerleri (2004) yaptıkları çalışmada lojistik regresyonun diğer sınıflandırıcılardan daha iyi sonuçlar verdiğini ortaya koymuştur.

3.Sınıflandırma Teknikleri

Literatürde denetimli makine öğrenmesinde kullanılan pek çok sınıflandırma tekniği mevcuttur. Ancak bu çalışma kapsamında söz konusu tekniklerden sadece üç tanesi kıyaslama amacıyla kullanılmıştır. Bu teknikler aşağıda ayrıntılı bir şekilde açıklanacaktır.

3.1. Navie Bayes

Bayes Teoremini temel alan olasılığa dayalı bir sınıflandırma yöntemidir. Navie Bayes Sınıflayıcıda test verisinden hareketle sistem öğrenmeyi gerçekleştirir ve en yüksek orana sahip olan örneği ilgili sınıfa dâhil eder.

C bir sınıfı gösterirken, c bilinen bir sınıfın etiketi, x=(x_1, x_2, x_3, … x_m) değerleri ise gözlemlenen nitelikler olmak üzere; x test verisinin ait olduğu sınıfı tahmin edebilmek için Bayes Teoremi ile olasılık hesaplanır;

$$p(C = cj|X = x) = \frac{p(C=cj)p(X=x|C=cj)}{p(X=x)}$$

En yüksek olasılık ile sınıfı tahmin edilir. X=x durumu; $X_1=x_1 \wedge X_2=x_2 \wedge X_3=\ldots..=X_n$ ifade eder. p(X=x) sınıflar arasında değişme göstermezse ihmal edilir ve (1) no'lu denklem aşağıdaki gibi değişir.

$$p(C = cj)|X = x) = p(C = cj)p(X = x|C = c)$$

(C=cj) ve p=(X=x|C=cj) öğrenme verilerinden tahmin edilir. x_i' ler birbirlerine koşullu olarak bağımsızdır. Bu durumda (2) no'lu denklemin alacağı şekil aşağıdaki gibi olur.

$$p(C = cj|X = x) = p(C = cj) \prod_i^m p(Xi = xi|C = cj)$$

(3)

Navie Bayes algoritması olarak adlandırılan (3) no'lu denklem kullanılarak test örneklerini hesaplamak ve eğitim verilerinden hareketle tahminlerde bulunmak daha kolaydır (Chandra vd; 2007).

3.2. Karar Ağaçları

Karar ağaçları genellikle eğitim ve test sürecinin hızlı olması ve elde edilen sonuçların kolay yorumlanabilmesi nedeniyle çok sık kullanılan bir sınıflandırma tekniğidir (Wu ve Banzhaf; 2010). Karar ağaçları ile sınıflandırma iki adımda gerçekleşir; önce ağaç oluşturulur daha sonra sınıflandırma kuralları elde edilir (Kaya ve Yıldız, 2014;93). Karar ağaçları kökler ve dallardan oluşan bir yapıdır. Çok farklı karar ağaçları algoritmaları vardır, kullanılan algoritmalar ağacın yapısını belirler. En yaygın algoritmalardan birisi de C4.5'tir ve bu çalışmada da C4.5 algoritması kullanılmıştır.

D veri tabanı ve C sınıf olmak üzere; $D=\{t_1,t_2,\ldots\ldots\ldots,t_n\}$ her bir kayıt t_i ile temsil edilmekte; $C\{C_1,C_2,\ldots,C_m\}$ ise m adet sınıftan oluşan sınıflar kümesini temsil etsin. Burada tüm C_j'ler ayrı bir sınıftır ve her bir sınıf kendisine ait kayıtları içerir. Yani, $C_j=\{t_i|=C_j,\ 1\leq i\leq n$ ve $t_i \in D\}$, dir. Veritabanındaki her bir kayıt için alanlar ise $\{A_1,A_2,\ldots A_n\}$ 'den oluşsun. Bu tanıma ilaveten her bir kayıt $C=\{C_1,C_2,\ldots C_m\}$ sınıflarından birine ait ise karar ağacı şöyle tanımlanabilir: Her bir düğüm A_i alanı ile isimlendirilir. Kök düğüm ile yaprak arasındaki düğümler birer sınıflandırma kuralıdır.

3.3. Lojistik Regresyon

Lojistik regresyon bağımsız değişkenleri veri iken kategorik bağımlı değişkenlerin olası sonuçlarını tahmin amacıyla kullanılır. m nitelikli n örnekleri için k sınıflarını oluğu durumda; son sınıf hariç j sınıfının olasılığı aşağıdaki bağıntıdan hesaplanabilir (Sağbaş ve Ballı, 2016;379)

$$Pj(Xi)=\frac{\exp(xibj)}{\sum_{j=1}^{k-1}\exp(xibj)+1} \quad (4)$$

Son sınıfın olasılığı;

$$1-\sum_{j=1}^{k-1}Pj(Xi)=\frac{1}{\sum_{j=1}^{k-1}\exp(xibj)+1} \quad (5)$$

ifadesi ile elde edilir.

4. Sınıflandırma Başarı Ölçütleri

Denetimli öğrenme tekniklerinin kullanıldığı sınıflandırma problemlerinde, model performansını değerlendirebilmek için en çok kullanılan yöntemlerden biri, hedef niteliğin sınıflarına ait gerçek ve tahmin değerlerinin bir arada gösterildikleri kontenjans tablosundan hareketle hesaplanan değerlerdir. Aşağıda bu tablo ve tablodan hareketle hesaplanan başarı ölçütleri gösterilmiştir (Balaban ve Kartal, 2015).

Tablo 1. Kontenjans Tablosu

DURUM	GERÇEK			
TAHMİNİ		POZİTİF	NEGATİF	TOPLAM
	POZİTİF	Doğru Pozitif (dp)	Yanlış Pozitif (yp)	tPoz
	NEGATİF	Yanlış Negatif (yn)	Doğru Negatif (dn)	tNeg
	TOPLAM	poz	neg	m

4.1. Doğruluk

Modelin başarılı olup olmadığını gösteren en basit ve en sık kullanılan ölçüt doğruluktur. Doğruluk, doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır. Doğruluk oranının 1'den çıkarılmasıyla da hata oranı hesaplanır. Hata oranı yanlış sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır.

$$Doğruluk\ (ACC) = \frac{dp+dn}{m} \quad (6)$$

Hata oranı (ERR) = 1 – ACC (7)

4.2. Duyarlılık

Doğru sınıflandırılmış pozitif örnek sayısının, toplam pozitif örnek sayısına oranıdır.

$$Duyarlılık\ (TPR) = \frac{dp}{poz} \quad (8)$$

4.3. Kesinlik

Doğru sınıflandırılan pozitif örneklerin toplam pozitif tahmin edilen örneklere oranıdır.

$$Kesinlik\ (PPV) = \frac{dp}{tPoz} \quad (9)$$

4.4. F-ölçütü

Kesinlik ve duyarlılık ölçütlerinin harmonik ortalamasıdır. Her iki ölçütü bir arada ele alma imkânı tanır.

$$F - ölçütü\ (F) = \frac{2*PPV*TPR}{PPV+TPR} \quad (10)$$

5. Uygulama

UCI Makine öğrenme havuzundan (makine öğrenme algoritmalarının ampirik analizi için kullandığı veri tabanları, alan teorileri ve veri setlerinin bir koleksiyonu

UCI 1987'de David Aha ve UC Irvine'deki diğer lisansüstü öğrenciler tarafından bir ftp arşivi olarak oluşturuldu. O zamandan beri, dünya çapında öğrenciler, eğitimciler ve araştırmacılar tarafından makine öğrenme veri setlerinin birincil kaynağı olarak yaygın bir şekilde kullanılmaktadır) elde edilen veriler, R programlamada denetimli makine öğrenmesi algoritmalarından, Lojistik Regresyon, Karar Ağaçları C4.5 ve Navie Bayes kullanılarak sınıflandırma yapılmıştır. Veri seti olarak, Cortez ve Silva (2008) tarafından derlenmiş olan, öğrenci performansı kullanılmıştır. Bu veri setinde Portekiz deki iki farklı liseye ait öğrencilerin demografik bilgileri, okulla ilgili bilgileri, aileleriyle ilgili bilgileri ve başarı durumları yer almaktadır. Verilerden hareketle öğrencilerin cinsiyetleri ve başarı durumları arasındaki ilişkiye dair makine öğrenmesi algoritmalarının başarıları kıyaslanmıştır.

R programlama dili açık kaynak kodlu olan bir yazılımdır. R da yer alan hazır paketler kullanılarak istenilen analizler yapmak mümkünken, programlama diline hakim olanlar tarafından istenilirse ihtiyaca yönelik paketler de oluşturulabilmektedir. Analiz kapsamında, Karar Ağacı için "partykit" paketi, Lojistik Regresyon için R ile varsayılan olarak yüklenen "stats paketinde glm()" fonksiyonu ve Navie Bayes için de "e1071" paketi kullanılmıştır. Veri seti üçte ikisi eğitim, üçte biri test olarak bölünmüş, elde edilen sonuçlar aşağıda özetlenmiştir.

Tablo 2. Başarı ölçütleri

Sınıflayıcı	Doğruluk	Duyarlılık	Kesinlik	F-ölçütü
Navie Bayes	0,74	0,88	0,71	0,79
Lojistik Regresyon	0,79	0,80	0,82	0,81
Karar Ağaçları	0,71	0,90	0,68	0,77

Doğruluk ölçütüne göre en iyi sınıflandırmayı Lojistik Regresyon algoritması yapmış, Lojistik Regresyonu Navie Bayes algoritması takip etmiştir.

Duyarlılık ölçütüne göre en iyi sınıflandırmayı Karar Ağaçları C4.5 algoritması yapmış, bunu Navie Bayes takip etmiştir.

Kesinlik ölçütüne göre en iyi sınıflandırmayı Lojistik Regresyon, daha sonra Navie Bayes algoritması yapmıştır.

Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olan ve her iki ölçütü aynı anda değerlendirmek için kullanılan F-ölçütü incelendiğinde; en iyi sınıflandırmayı Lojistik Regresyon algoritmasının yaptığı görülmüştür.

Hangi başarı ölçütü ele alınırsa alınsın genel olarak yüksek bir başarı ile algoritmaların sınıflandırma yaptığı görülmektedir. En iyi sınıflandırmayı dört ölçütten üçünde Lojistik Regresyon vermiştir. Dikkat çeken bir başka unsursa Navie

Bayes her ne kadar hiçbir ölçütte en iyi sınıflandırmayı yapamamış olsa da tüm başarı ölçütleri gösteriyor ki en iyi ikinci sınıflandırmayı yapan algoritma Navie Bayestir.

5.Sonuç

Bu çalışmada UC Irvine Makine Öğrenmesi veri deposundan alınmış olan, Polonya'daki öğrencilere ait reel veriler üzerinde çeşitli denetimli makine öğrenmesi algoritmalarının sınıflandırma başarıları test edilmiştir. Model başarım ölçütleri olarak; doğruluk, kesinlik, duyarlılık ve F-ölçütü kullanılmıştır.

En başarılı sınıflandırma algoritması olarak Lojistik Regresyon karşımıza çıkmaktadır. Lojistik Regresyonun en yüksek sonucu verdiği başarı ölçütü ise; 0,90 ile "duyarlılık" olmuştur.

Teşekkür

Bu çalışma Uludağ Üniversitesi BAP birimi tarafından desteklenen DDP(İ) – 2017/8 numaralı proje kapsamında yapılmıştır. Projemize verdikleri destekten dolayı Uludağ Üniversitesi BAP Birimine teşekkürü bir borç biliriz.

KAYNAKLAR

A. Charitou, E. Neophytou ve C. Charalambous,"Predicting corporate failure: empirical evidence for the UK", European Accounting Review, 13(3), 465-497, 2004.

B. İbrahim Sevindi, "Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması" Yüksek Lisans Tezi, 2013.

Chandra B, Gupta M ve Gupt MP. "Robust approach for estimating probabilities in Naive-Bayes classifier". Pattern Recognition and Machine Intelligence. Kolkata, India, 18-22 December 2007

Çetin KAYA, Oktay YILDIZ, "Makine Öğrenmesi Teknikleriyle Saldırı Tespiti: Karşılaştırmalı Analiz" Marmara Fen Bilimleri Dergisi 2014,3:89-104

D Michie, D J Spiegelhalter ve C C Taylor, *Machine Learning, Neural and Statistical Classification*, Overseas Press, 1994.s.3

Ensar Arif SAĞBAŞ, Serkan BALLI, "Akıllı telefon algılayıcıları ve makine öğrenmesi kullanılarak ulaşım türü tespiti", Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22(5), 376-383, 2016

Ethem Alpaydın, *Introduction to Machine Learning Second Edition*., The MIT Press, Cambridege, 2010, s.4

M. Erdal BALABAN, Elif KARTAL, Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları, Çağlayan Kitabevi, İstanbul, 2015

Lara OD, Pérez AJ, Labrador MA ve Posada JD. "Centinela: A human activity recognition system based on acceleration and vital sign data". Pervasive and Mobile Computing, 8(5), 717-729, 2012

P. Cortez, A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008

S. Y. Wu ve E. Yen, "Data mining-based intrusion detectors", Expert Systems with Applications, vol. 36, no. 3, pp. 5605-5612, 2009.

S. Wu ve W. Banzhaf, «The Use of Computational Intelligence in Intrusion Detection Systems: A Review,» Applied Soft Computing, cilt 10, no. 1, pp. 1-35, 2010.

Sinan Aydın ve Ali Ekrem Özkul, 'Veri Madenciliği Ve Anadolu Üniversitesi Açıköğretim Sisteminde Bir Uygulama', *Eğitim ve Öğretim Araştırmaları Dergisi*, Ankara, Cilt 4, Sayı 3, 2015, s. 38.

https://archive.ics.uci.edu/ml/index.php, Erişim tarihi, 15.01.2018