Afyon Kocatepe Üniversitesi Uluslararası Mühendislik Teknolojileri ve Uygulamalı Bilimler Dergisi

Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences

AKÜ IJETASCilt**3(2)** (2020) Aralık (52-59 s) AKU J.Eng.App.Sci. **Vol3(2)** (2020) Decemeber (52-59pp)

Araştırma Makalesi / Research Article

e-ISSN 2667-4165 (https://dergipark.org.tr/akuumubd)

Makine Öğrenmesi Yöntemleri ile Web'den Bilgi Çıkarımı Sürecinin İyileştirilmesi

Erkan Özhan[1]

[1] *Tekirdağ Namık Kemal Üniversitesi, Çorlu Mühendislik Fakültesi, BilgisayarMühendisliği Bölümü, Çorlu, Tekirdağ.*

e-posta: erkanozhan@gmail.com, ORCID ID: http://orcid.org/0000-0002-3971-2676

Geliş Tarihi:22.08.2020 ; Kabul Tarihi:15.09.2020

Anahtar kelimeler
Web bilgi çıkarımı; Makine öğrenmesi; Sınıflandırma; Veri madenciliği

Öz

Web ortamı bilginin doğduğu, yayıldığı ve yaşadığı bir formata sahiptir. Gün geçtikte bilgi morfolojik olarak değişim geçirmekte ve bu değişimle birlikte avantajlar yanında istenilen anlamlı bilgiye ulaşmada zorluklar artmaktadır. Zaman, depolama, iletişim ve veri işleme maliyetleri açısından istenilen bilgiye en verimli şekilde ulaşmak kritik bir görevdir. Bunun yanında verinin yaşam süreci boyunca kullanılabilirliğini de artırabilir. Web sayfalarının "layout" adı verilen bölümlerinin sınıflandırılması bu sorunların çözümüne önemli katkılar sağlayabilir. Özellikle bu bölümlerdeki gereksiz içeriğin bilinmesi faydalı ve anlamlı bilgiye ulaşmayı kolaylaştırıcı ve maliyetleri düşürücü etki sağlayabilir. Bu çalışma makine öğrenmesi yöntemleri ile web sayfası bölümlerinin sınıflandırılması sürecini iyileştirmek amacıyla farklı algoritmalara odaklanmış ve bu algoritmaların iyileştirici sonuçlarını ortaya koymuştur. Elde edilen sonuçlara göre Random Forest ve KStar algoritmalarının süreci iyileştirici modeller olduğu görülmüştür. Random Forest algoritması %98.46 doğru sınıflandırma oranı sunarken, KStar hız faktörüyle öne çıkmıştır. Çalışmada karar ağacı ve entropi tabanlı algoritmaların başarımları da karşılaştırılmış ve bulgular hesaplama zamanlarıyla birlikte sunulmuştur.

Improvingthe Information ExtractionProcessfromthe Web with Machine Learning Methods

Keywords
Web informationextraction; Machine learning; Classification; Data mining

Summary

The web environment has a format in which information is born, propagated and lived. Information changes morphologically day by day, and with this change, difficulties in reaching the desired meaningful information increase as well as advantages. It is a critical task to reach the desired information in the most efficient way in terms of time, storage, communication and data processing costs. In addition, it can increase the availability of data throughout its life cycle. Classification of the parts of web pages called "layout" can make important contributions to the solution of these problems. In particular, knowing the unnecessary content in these sections can facilitate access to useful and meaningful information and provide a cost-reducing effect. This study focuses on different algorithms in order to improve the process of classifying web page sections with machine learning methods and reveals the improvement results of these algorithms. According to the results, it has been seen that Random Forest and KStar algorithms have process improvement solutions. While the Random Forest algorithm offers 98.46% correct classification rate, KStar stands out with its speed factor. In the study, especially the performance of tree and entropy-based algorithms were compared and the findings were presented together with the computation times.

1. Giriş

Bilgisayarlar veya diğer cihazların iletişim ortamına dahil edilmesinin birincil nedeni bilgiyi paylaşmaktır. Bu iletişimin yönü tek yönlü olabildiği gibi çift yönlü de olabilir. Bir web sayfasındaki metin bloğunu okumak tek yönlü iken, bu yazıya yorum yapmak bir anda çift yönlü bilgi paylaşımına neden olur. Dünyada iletişim için kullanılan bu araçların çeşitliliği, yetenekleri, kapasiteleri ve iletişim-işlem hızları arttıkça iletişim sonucu ortaya çıkan bilgi yoğunluğu da dramatik bir şekilde artmaktadır. Bu yoğunluk ulaşılabilecek bilgi miktarı ve çeşitliliğinin artması anlamına gelir. Bu sonuç ilk bakışta olumlu görünse de istenilen bilgiye ulaşmak için çok sayıda ayıklama yapmanız gerekeceği anlamına gelir. Bu sorun yeni değildir ve en ilgili sonuçları ortaya çıkarmak için arama motorları geliştirilmiştir. Ancak arama motorları da devasa boyutlara ulaşmış, web kaynaklarını elde etmede ve onları değerlendirmede sorunlar yaşamaktadır. Web sayfaları yalnızca gerçek içerikten değil, aynı zamanda afişler (banner), gezinme öğeleri, reklamlar, telif hakkı vb. gibi diğer unsurlardan da oluşur (Wu, Liu and Fan, 2015). Web içerik çıkarımı (web content extraction) istenilen bilgiye veya ona en yakın olana ulaşmanın yollarını arar. Arama sonuçlarını en ilgili ve en hızlı bir şekilde ortaya çıkarmak için performans artırıcı çok sayıda teknik kullanırlar. Veri indirgenimi, haritalama, yüksek başarımlı hesaplama, veri madenciliği ve makine öğrenmesi (machine learning-ML) gibi çok sayıda fark yaratabilecek disiplinden faydalanmaktadırlar. Kısacası verinin elde edilmesi, depolanması, işlenmesi ve sunumunda iyileştirme çalıştırmaları sürekli yapılmaktadır.

İnternet ortamında bulunan veriler metin, resim, video, ses gibi birkaç farklı formatta olabilir. Bu verilerin sunumu için ise direkt veya dolaylı olarak dosyalar kullanılır. Veriler .html, .php, .asp vb. gibi dosyalar içerisinde gömülü olarak bulunabilir. Bunun yanında veri tabanlarından, sensörlerden vb. elde edildikten sonra sayfaya gömülerek yine web ortamında sunulabilir. Çoğu web sayfası; gazete, alışveriş kataloğu, başvuru formu gibi uzayıp giden bir listenin elektronik versiyonu gibidir (Duckett, 2011).

Web sayfaları layout adı verilen bölümlerden oluşur. Web öğelerinin konumlandığı bölgelere layout denir. Bu bölümlerin çeşitliliği ve sayısı oldukça değişkendir. Günümüzde web sayfası oluşturmak için çok sayıda biçimlendirme ve web programlama dili geliştirilmiştir. Web sayfaları için temel biçimlendirme dili HTML (Hypertext Markup Language)'dir. HTML aynı zamanda web sayfaları oluşturmak ve yönetmek için temel standarttır. Web sayfası (HTML belgesi), düğümlerin HTML öğeleri olduğu bir ağaç olarak temsil edilebilir (Štěpánek and Šimková, 2013). HTML için özgün ortam dışına gönderim formatı denebilir (Raggett, 1994). Bu formata göre metinler, resimler, tablolar, formalar vb. için geçerli olan biçimlendirme gereksinimleri "tag" adı verilen tanımlayıcı etiketlerle temsil edilebilir. Dahası bu etiketler yardımıyla özgün ortam dışına aktarıldığında biçimi bozulmadan tekrar bir araya getirilerek görüntülenebilir. Temelinde ise etiketler arasında kurulmuş hiyerarşik bir yapı vardır. Her etiket hiyerarşik yapıya uymak zorundadır.

Verilerin bir aygıttan diğerine aktarıldığında yeni konumunda nasıl görüntüleneceği ve depolanacağı çeşitlilik gösterse de anlaşılırlık ve tutarlılık temel gereksinimdir. Veriler TCP-IP (Transmission Control Protocol and Internet Protocol) protokollerine göre paketlere ayrılır, paketler telefon hattı üzerinden gönderilir ve alıcı bilgisayar tarafından kendi internet yazılımı kullanılarak etiketlere göre tekrar bir araya getirilir (Berners-Lee and Fischetti, 2000).

Bilgisayarlar için çok sayıda veri kaynağını HTML yapısının esaslarına göre gömülü veya doğrudan yazılarak barındıran sunucular (server) bilginin düğüm noktaları olarak düşünülebilir. Yoğun bir şekilde veri veya hizmet barındıran bu düğümler içerisinden istenilen bilgiyi tam olarak alabilmek için sıkı bir ayıklama yapmak kaçınılmazdır. Bu süreç sunucu için bant genişliği, işlemci ve bellek gibi kaynakların tüketiminde, istemci (client) içinse kaynak ve zaman maliyetlerini doğrudan etkiler. Bu nedenle istenilen bilginin az kaynak ile en doğru ve kısa sürede edilmesi amaçlanmaktadır. Bunun yanında web sayfalarından "yararlı ve ilgili" içeriğin çıkarılması, cep telefonu ve PDA taraması, görme engelliler için konuşma oluşturma ve metin

özetleme gibi birçok uygulamaya sahiptir (Gupta, Kaiser, Neistadt and Grimm, 2003). Web sayfalarının bölümlere ayrılması ve gürültünün (bilgilendirici olmayan bölüm) kaldırılması, duyarlılık analizi, metin özetleme ve bilgi erişimi gibi çeşitli uygulamalarda önemli ön işleme adımlarındandır (Pappas, Katsimpras and Stamatatos, 2012).

Sonuçta günümüzde web belgeleri üzerine yerleşmiş olan ve çok büyük miktarda kozmikleşmiş veri barındıran web ekosistemi ortaya çıkmıştır. Bu kozmik veriden anlamlı, işe yarar (knowledge) sonuçların çıkarılması için veri madenciliği (data mining) ve yapay zeka (artificial intelligence) tekniklerinin kullanılması kaçınılmazdır.

Veri madenciliği, büyük ve karmaşık veriler içerisinden anlamlı ve işe yarar bilgiyi ortaya çıkarmanın yöntemlerini inceleyen bir disiplindir. 1990'lardan bu yana, veri madenciliği kavramı, akademik alandan iş dünyasına veya tıbbi faaliyetlere kadar pek çok ortamda ortaya çıkmıştır (Gorunescu, 2011). Anlamlı bilginin keşfinde bir diğer ilgili disiplin ise yapay zekadır. Yapay zeka (Artificial Intelligence-AI), insanın öğrenme sürecine benzer olarak bilgisayarların veriden öğrenmesinin yöntemlerini inceleyen bilim dalıdır. Yapay zeka disiplininin altında yer alan makine öğrenmesi alt alanında denetimli (Supervised), denetimsiz (Unsupervised) ve Yarı-denetimli (semi-supervised) öğrenme olmak üzere üç tür öğrenme bulunmaktadır. Veri madenciliği ve yapay zeka veri içerisinden daha önce ortaya çıkarılmamış öngörülemeyen bilgiler gibi işe yarar bilgiyi ortaya çıkarmak için kullanırlar. Yapay zeka sistemleri, boyut ve karmaşıklık açısından giderek daha yetenekli olma eğilimindedir (Shuldiner, 2019). Sınıflandırma, kümeleme (clustering), birliktelik ilişkileri kurma gibi veriler üzerinde çok farklı görevleri yerine getirebilirler. Özellikle gelecekteki veriler üzerinde öngörü sunabilmesi ve faydalı örüntüler keşfedebilmesi onları cazip kılar.

Web içeriği çıkarma (web content extraction) teknikleri iki kategoride gruplanabilir: el yapımı kurallar ve otomatik ayıklama (Uzun, Serdar Güner, Kılıçaslan, Yerlikaya and Agun, 2014). Ele alınan verinin karmaşık, çok boyutlu ve büyük olması gibi anatomik özelliklerinden dolayı akıllı ve otomatik bir sistem geliştirmek oldukça faydalı ve başarılı olabilir.

Bu çalışmanın birinci amacı yapay zeka teknikleri ile layout-bölüm sınıflandırma işleminin başarımını artırmaktır. İkinci amacı ise algoritmaların başarım süreleri yanında işlem hızlarını da elde ederek analiz etmektir. Çalışmanın ikinci bölümünde önceki çalışmalar özetlenmiştir. Üçüncü bölümünde ise makine öğrenmesi teknikleri ve değerlendirme metrikleri hakkında bilgi verilmiştir. Dördüncü bölümde ise bulgular sunulmuş, son bölümde sonuçlar verilerek tartışılmıştır.

2. Önceki Çalışmalar

Web sayfalarındaki gerçek içeriği ayıklamak için çok sayıda akademik çalışma yapılmıştır. Wu ve ark. (Wu et al., 2015) yaptıkları araştırmada web sayfalarının DOM (Document Object Model-Belge Nesne Modeli) ağaç düğümü özelliklerini kullanarak birden çok özellik elde etmişler ve bu özellikleri makine öğrenimi yöntemini ile modellemeye çalışmışlardır. Araştırmacılar gerçek içeriğin uzamsal ve sürekli bir blokta yer aldığını gözlemlemişlerdir. Gupta ve ark. (Gupta et al., 2003) ise yine DOM ağacı ile orijinal verileri özetlemek yerine tanımlayarak ve koruyarak İçerik çıkarmaya çalışmışlardır. Weninger ve ark. farklı bir teknikle HTML belgesinin etiket oranlarını kullanarak çeşitli web sayfalarından içerik metni çıkarmak için Etiket Oranları (Content Extraction via Tag Ratios-CETR) adlı bir yöntem önermişlerdir. Uzun ve ark. (Uzun et al., 2014) ise yaptıkları araştırmada yedi farklı blok üzerinden web içeriğini otomatik olarak elde eden iCrawler adlı bir akıllı tarayıcı geliştirmişlerdir. Araştırmacılar daha sonra topladıkları içeriği makine öğrenmesi algoritmalarından DecisionTable (Karar tablosu) algoritması ile yüksek doğruluk oranı ile modellemeyi başarmışlardır. Yang ve Song(Yang and Song, 2010) ise heterojen yapıdaki web sayfaları ile başa çıkmada daha fazla uyarlanabilirliğe sahip gürültü ve karakteristiği gidermek üzere kullanılmayan aday düğümleri düzeltmeye dayalı bir yöntem önermişlerdir. Pappas ve ark. (Pappas et al., 2012), web sayfasının

görsel ve görsel olmayan özelliklerini hesaba katan ve kullanıcı tarafından oluşturulan içeriği (Haberler, Bloglar, Tartışmalar) içeren üç ana sayfa kategorisinden gürültülü bölümleri kaldırabilen bir algoritma önermişlerdir. Diğer yandan Bu ve ark. (Bu, Zhang, Xia and Wang, 2014) ana metin içeriğini web sayfalarından çıkarmak için bulanık ilişki kuralları (fuzzy association rules-FAR) kavramını kayan pencere (sliding window-SW) kavramıyla bütünleştiren istatistik tabanlı bir yaklaşım önermişlerdir. Uzun ve ark. (Uzun, Agun and Yerlikaya, 2013) gürültülü içeriği ortadan kaldırmak ve istenilen bilgiye ulaşmak için hibrit bir yöntem önermişlerdir. Lin ve ark. (Lin, Sheng, Vo and Tata, 2020)FreeDOM adını verdikleri araştırmada her site için örnek gerektiren ve web sitelerinin görsel yapısı üzerine inşa edilen sezgisel içerik çıkarım yöntemlerin sınırlılıkları olduğunu belirtmişlerdir. Araştırmacılar bu sorunu çözmek için FreeDOM'un web sayfalarının metin ve biçim bilgilerini birleştirerek sayfadaki her DOM düğümünün temsilini (Word embedding) öğrendiğini ve bu bilgiyi bir sinir ağı ile semantik ilişkiler elde etmek için kullandığını göstermişlerdir. Uçar ve ark. (Uçar, Uzun and Tüfekci, 2016) birbirini tamamlayan iki aşamalı bir algoritma önererek yüksek doğruluk elde etmişlerdir.

Bu çalışmada bu veri seti için daha önce kullanılmamış olan algoritmalar ve algoritma optimizasyon araçları kullanılmış ve başarım oranı aynı veriyi kullanan önceki çalışmalara göre artırılmıştır. Bunun yanında model hesaplama zamanları da çıkarılarak karşılaştırılıştır.

3. Materyal ve Metot

Çalışmada veri setinin analizi için farklı makine öğrenmesi algoritmaları eğitilmiştir. Daha sonra, testlerden elde edilen bulgular kaydedilerek değerlendirme metriklerine göre değerlendirilmiştir.

3.1 Makine Öğrenmesi Algoritma Testleri

Çalışmanın makine öğrenmesi algoritmaları testi Weka (Frank et al., 2009) (Waikato Environment for Knowledge Analysis) adlı Waikato üniversitesi tarafından geliştirilmiş açık kaynak kodlu, Java tabanlı yazılım aracılığı ile yapılmıştır. Weka bünyesinde çok sayıda yapay zeka algoritmasını barındırmaktadır. Package Manager sayesinde yeni algoritma ve veri işleme araçları yüklemek mümkündür. Veriler üzerinde çok sayıda algoritma denenmiş ve en başarılı ilk 5 algoritma (Random Forest, Random Tree, JRIP, Bagging, KStar) tespit edilmiştir.

İlk algoritma Random Forest'tır. Random Forest algoritması, hem sınıflandırma hem de regresyon görevlerini gerçekleştirebilen çok yönlü ve akıllı bir makine öğrenmesi yöntemi olarak tanımlanabilir (Sullivan, 2018). Random Forest algoritması, ağaç indüksiyon algoritmasının rastgele bir varyantından türetilen ve bir karar ağaçları topluluğu (veya ormanı) oluşturmayı içerir (Louppe, 2014). Bu karar ağaçları kullanılarak özellikle büyük veriler üzerinde etkili çözümler elde edilebilir.

Ağaçlar, döngüsüz bağlı grafikler olarak tanımlanır ve özellikleri grafik (graph) teorisinin temelleridir (Drmota, 2009). Düğümler ve düğümlere komşu olan diğer düğümlerden meydana gelirler. Bu düğümlerden bir tanesi (root-r) kök düğüm olarak adlandırılır. Ancak köksüz de olabilirler. Ağaç tabanlı algoritmalarda ağacı ters çevrilmiş olarak düşünürsek bir verinin sınıfı kök düğümden başlanarak aşağı doğru her bir düğümdeki kritere göre yönlendirilerek bulunur. Random Tree sınıflandırıcı birkaç karar ağacını, önyüklemeye paralel olarak eğiten ve ardından bagging (torbalama) adı verilen işlemle bir araya getiren yöntemler topluluğudur (Misra, Li and He, 2019).

JRIP algoritması Cohen(Cohen, 1995) tarafından geliştirilmiş eğitim örnekleriyle hızlı ölçeklenebilen ve yüzbinlerce örnek içeren gürültülü veri kümelerini verimli bir şekilde işleyebilen kural tabanlı bir algoritmadır. Bu tür algoritmalar basit deterministik mantıksal kurallar üretir ve tüm örneklerin mükemmel bir doğrulukla sınıflandırılmasına izin verir (Nosofsky, Gluck, Palmeri, Mckinley and Glauthier, 1994). Kural tabanlı bir sınıflandırıcının kuralları bir karar ağacından çıkarılabilir (Yucalar, Ozcift, Borandag and Kilinc, 2020).

Bagging ensemble (birlikte) öğrenme temeline dayalı bir algoritmadır. Birlikte öğrenme yöntemleri, tek bir karar ağacı sınıflandırıcısından daha iyi tahmin performansı üretmek için birkaç karar ağacı sınıflandırıcısını birleştirir (Sarkar, 2019). Birliktelik modelinin arkasındaki ana ilke, bir grup zayıf öğrenicinin güçlü bir öğrenici oluşturmak için bir araya gelmesi ve böylece modelin doğruluğunun arttırılmasıdır. Bagging algoritması, önce bir sınıflandırıcılar komisyonu kurar ve ardından bunların sonuçlarını çoğunluk oylamasıyla toplar (Hsu and Srivastava, 2012).

KStar algoritması, olası tüm dönüşümler arasından rasgele seçim yaparak entropik ölçüm kullanır (Madhusudana, Kumar and Narendranath, 2016). Uzaklık ölçümü olarak entropi kullanımı sembolik niteliklerin, gerçek değerli niteliklerin ve eksik değerlerin ele alınmasında tutarlı bir yaklaşım sağlar (Cleary and Trigg, 1995).

3.2 Değerlendirme Metrikleri

Makine öğrenmesi algoritmalarının performanslarını değerlendirmek için standart bazı metrikler kullanılır. Bunların sayısı çok olmakla birlikte kullanılacak olan alan ve beklentiye göre değerlendirme metrikleri özel olarak seçilebilir. Örneğin pozitif örneklerin önemi büyük ise ve yanlış pozitifliğin maliyeti düşük ise eşik değeri düşük tutularak tüm pozitifiler yakalanabilir. Çalışmada kullanılan algoritmalar için değerlendirme metriği olarak F-Measure, Kappa, RMSE ve Correctly Classified Instances seçilmiştir.

F-Measure, Precision (kesinlik) ve Recall değerlerinin harmonik ortalamasıdır. Recall model tarafından öğrenilen veya deneyimlenen bir şeyi hatırlama eylemi veya yeteneğini ifade eder. Precision ise modelin tahminlerindeki kesinliği ifade eder. Modelin çıktıları 4 durumdan biri olabilir bunlar:

- TP: Gerçekte pozitif iken model tarafından da pozitif olarak sınıflandırılanlar
- FP: Gerçekte negatif iken model tarafından pozitif olarak sınıflandırılanlar.
- TN: Gerçekte negatif iken model tarafından da negatif olarak sınıflandırılanlar.
- FN: Gerçekte pozitif iken model tarafından negatif olarak sınıflandırılanlardır.

Precision denklem 1'de gösterildiği gibi modelin sınıflandırdığı pozitif örnek sayısının toplam pozitif girdi sayısına oranıdır.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{1}$$

Recall ise yine denklem 1'de gösterildiği gibi modelin pozitif sınıflandırdığı örnekler içerisinden kaç tanesinin gerçekten pozitif olduğunun ölçüsüdür. F-Measure veya literatürde F-Ölçüsü, F-Score olarak adlandırılan metrik ise denklem 2'deki gibi hesaplanır.

$$F - Measure = \frac{2 \times P \times R}{P + R} \tag{2}$$

Correctly Classified Intances değeri ise modelin doğru sınıflandırdığı örnek sayısının yüzdelik ifadesi olarak temsil edilen metriktir. Kappa ise gözlenen ve beklenen değerler arasındaki uyuşmayı gösterir. Yani modelin çıktısı ile beklenen çıktı (gerçek) arasındaki uyuşmayı temsil eder. Kappa -1 ile +1 arasında değer alabilir. Model için 1'e yakın bir kappa değeri istenir.

RMSE (Root mean squared error) ise denklem 3'te gösterildiği gibi modelin tahminleri ile gerçek değer arasındaki hataların karasel ortalaması alındıktan sonra toplanması ve sonucun karekökün alınması ile elde edilir. Bu değerlendirme hassasiyeti sağlar.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\text{Gerçek değer-Model Tahmini})^2}{n}} \tag{3}$$

İki algoritmanın performansı birbiriyle neredeyse aynı ise, o zaman RMSE'ye bakarak hangisinin daha iyi olduğu ayırt edebilir (Pradham, Younan and King, 2008). RMSE'nin rakip algoritmaya göre düşük değerli olması model sonuçlarının daha doğru olduğunu gösterir(Aydin, Yucel and Sadikoglu, 2018).

3.3 Veri Seti

Araştırmada Uzun ve ark. (Uzun et al., 2014) tarafından elde edilmiş olan 49 girdi özniteliği ve 7 çıktısı (sınıfı) olan veri seti kullanılmıştır. Veri setinin çıktıları web sayfalarında bulunan ve "main, menü,

links, summary, empty, headline, others" olarak etiketlenmiş bölüm adlarıdır. Araştırmacılar veri setini 2011 yılına kadar Goggle News'den rastgele seçilen 110 farklı Web alanından olmak üzere toplam 2414 web sayfasından elde etmişlerdir. Veri seti (Uzun, 2014) 14742 satırdan oluşmaktadır. Veri setinin sınıf dağılımları ise Tablo 1'de gösterilmiştir.

Tablo 1'e bakıldığında en yüksek değerin menü, links ve empty'e ait olduğu görülür. En düşük değer ise summary sınıfına aittir.

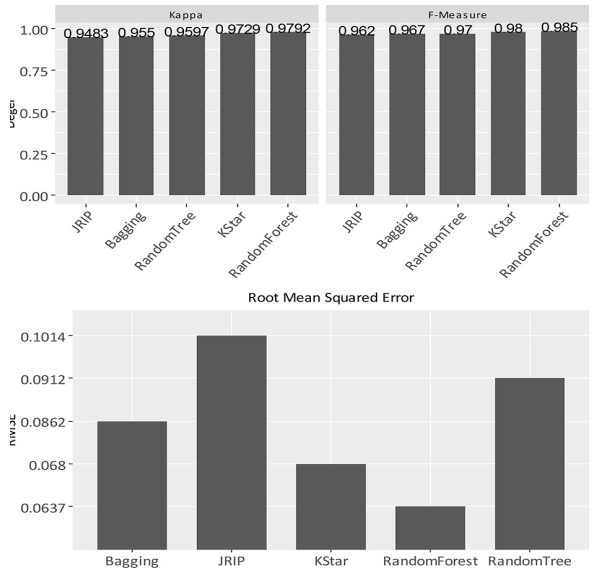
Tablo 1. Veri setinin sınıf dağılımı

Sınıf Adı	Sayısı
Main	549
Headline	553
Summary	73
Others	1889
Menu	5643
Links	4054
Empty	1981

Veri setinin sınıf dağılımı incelendiğinde dengeli bir dağılım olmamasına karşın sonraki bölümde paylaşılacak olan sonuçlara göre analiz gerekliliklerini sağladığı yani yanlılık ile karşılaşılmadığı görülmektedir.

4. Bulgular

Yapılan makine öğrenmesi algoritma testlerine göre elde edilen performans metrikleri ve değerleri Şekil 1'de sunulmuştur.

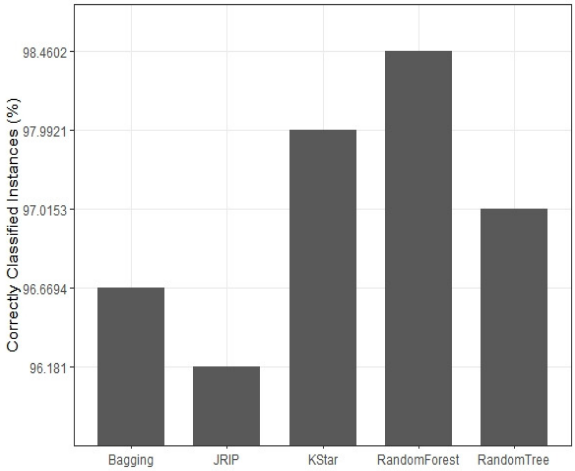


Şekil 1. Algoritmaların Kappa ve F-Measure değerleri

Şekil 1'in üst bölümü Kappa ve F-measure değerlerini göstermektedir. Random Forest ve KStar algoritmalarının Kappa ve F-measure değerlerine göre en iyi sınıflandırıcı algoritmalar olduğu görülmektedir.

Bunun yanında RMSE değerlerine bakıldığında en az hata oranına sahip olan algoritmaların sırasıyla Random Forest ve KStar olduğu görülmektedir.

Uzun ve ark. (Uzun et al., 2014)'nın Decision Table algoritmasıyla elde ettikleri doğru sınıflandırma (Accuracy) değeri %96.87'dir. Bu çalışmada elde edilen bulgular Şekil 2'de gösterilmiştir. Şekil 2'de gösterildiği gibi Random Forest algoritması ile elde edilen doğru sınıflandırma oranı %98.46'dir. Bunun yanında KStar (%97.99), Random Tree (%97.01) algoritmalarının da önceki çalışmaya göre daha iyi sonuçlar ürettiği görülmüştür.

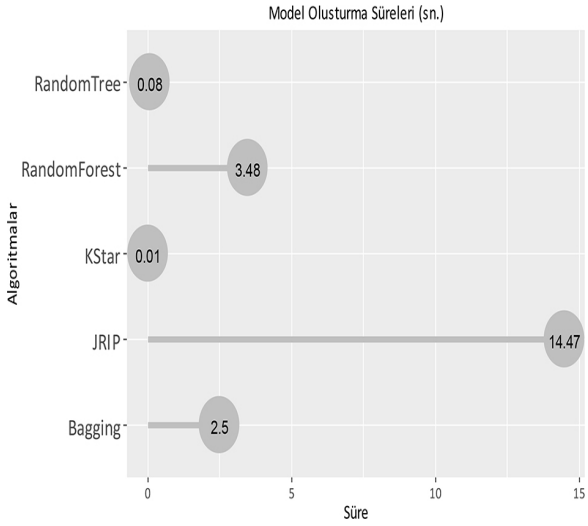


Şekil 2. Algoritmaların Correctly Classified Instances Değerleri

Başarımı diğerlerine göre daha düşük olan algoritmalar ise JRIP ve Bagging algoritmalarıdır.

Özellikle entropi temelli uzaklık ölçümü kullanan KStar ve kural tabanlı ağaç algoritmalarının veri setinin sınıflandırılmasında yüksek başarı gösterdikleri söylenebilir. Veriler tür olarak sayısal verilerdir. Sadece çıktılar nominal türdedir.

Şekil 2'deki Correctly Classified Instances ve Şekil 3'teki F-Measure, Kappa ve RMSE değerleri düşünüldüğünde hız/performans açısından KStar algoritmasının Random Forest'ın alternatifi olarak kullanılabileceği söylenebilir.



Şekil 3. Algoritmaların model oluşturma süreleri

Şekil 3'te görüldüğü gibi KStar algoritması en hızlı model kurma zamanına sahip algoritmadır.

5. Sonuçlar ve Tartışma

Web sayfalarından içeriğin ayıklanması ve çıkarılması özellikle erişim ve işlem hızı, depolama, bant genişliği ve istenilen bilgiye en doğru şekilde ulaşmak açısından son derece önemli bir görevdir. Web sayfalarının layout'larının tespit edilmesi bilgi çıkarımı için iyi bir başlangıç olabilir. Bu çalışmada çeşitli web sayfalarından toplanarak oluşturulan veri seti üzerinde bölüm sınıflandırması işlemi makine öğrenmesi algoritmalarıyla iyileştirilmeye çalışılmıştır. Sınıflandırma başarım oranı Random Forest algoritması ile %1.59 oranında iyileştirilmiştir.

Çalışmada ayrıca sınıflandırıcı makine öğrenmesi algoritmalarının model oluşturma süreleri de analiz edilmiş bu analizler sonucunda hem başarım hem de hesaplama süresi bakımından alternatif olarak KStar algoritmasının kullanılabileceği görülmüştür.

Çalışmada elde edilen sonuçlar makine öğrenmesi algoritmalarının yapısal açıdan değerlendirildiğinde ağaç ve entropi tabanlı algoritmaların bu veri seti üzerinde daha başarılı sonuçlar verdiğini göstermiştir.

Gelecekte araştırmacılar bu veri seti üzerinde performans artırıcı çalışmalar yürütebilirler. Bunun yanında veri setinin ve barındırdığı özniteliklerin geliştirilmesi ve iyileştirilmesi de düşünülebilir.

Teşekkür

Bu araştırmada kullanılan verileri sağlayan ve açık erişim şekilde yayınlayarak paylaşan Tekirdağ Namık Kemal Üniversitesi, Çorlu Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Doç. Dr. Erdinç Uzun'a teşekkürlerimi sunarım.

6. Kaynaklar

Aydin, E. S., Yucel, O. and Sadikoglu, H. (2018). Chapter 2.6 - Numerical Investigation of Fixed-Bed Downdraft Woody Biomass Gasification. I. Dincer, C. O. Colpan and O. B. T.-E. Kizilkan Energetic and Environmental Dimensions (Eds.), (pp. 323–339). Academic Press. doi:https://doi.org/10.1016/B978-0-12-813734-5.00018-4

Berners-Lee, T. and Fischetti, M. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (1st ed.). New York, NY, USA: Harper Business.

Bu, Z., Zhang, C., Xia, Z. and Wang, J. (2014). An FAR-SW based approach for webpage information extraction. *Information Systems Frontiers*, *16*(5), 771–785. doi:10.1007/s10796-013-9412-2

Cleary, J. G. and Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. *Machine Learninginternational Workshop Then Conference*, *5*, 1–14. doi:10.1.1.51.4098

Cohen, W. W. (1995). Fast Effective Rule Induction. A. Prieditis and S. B. T.-M. L. P. 1995 Russell (Eds.), (pp. 115–123). San Francisco (CA): Morgan Kaufmann. doi:https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Drmota, M. (2009). *Random trees: An interplay between combinatorics and probability*. *Random Trees: An Interplay Between Combinatorics and Probability*. doi:10.1007/978-3-211-75357-6

Duckett, J. (2011). *HTML and CSS: Design and Build Websites*. (C. Long, Ed.). Indianapolis, Indiana: John Wiley & Sons.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H. and Trigg, L. (2009). Weka-A Machine Learning Workbench for Data Mining. *Data Mining and Knowledge Discovery Handbook* in . doi:10.1007/978-0-387-09823-4_66

Gorunescu, F. (2011). *Data Mining*. Intelligent Systems Reference Library (Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-19721-5

Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P. (2003). DOM-Based Content Extraction of HTML Documents. *Proceedings of the 12th International Conference on World Wide Web* in , WWW '03 (pp. 207–214). New York, NY, USA: Association for Computing Machinery.

doi:10.1145/775152.775182

Hsu, K.-W. and Srivastava, J. (2012). Improving Bagging Performance through Multi-algorithm Ensembles. L. J. Huang J.Z., Bailey J., Koh Y.S. (Ed.), *New Frontiers in Applied Data Mining* in (pp. 471–482). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-28320-8_40

Lin, Y., Sheng, Y., Vo, N. H. and Tata, S. (2020). FreeDOM: A Novel Transferable Neural Architecture for Structured Data Extraction over Web Documents. *KDD 2020* in .

Louppe, G. (2014, July). *Understanding Random Forests from Theory to Practice*. https://arxiv.org/pdf/1407.7502.pdf from retrieved.

Madhusudana, C. K., Kumar, H. and Narendranath, S. (2016). Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal. *Engineering Science and Technology, an International Journal*, *19*(3), 1543–1551. doi:https://doi.org/10.1016/j.jestch.2016.05.009

Misra, S., Li, H. and He, J. (2019). *Machine Learning for Subsurface Characterization*. Elsevier Science. https://books.google.com.tr/books?id=WdO1DwAAQBAJ from retrieved.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C. and Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369. doi:10.3758/BF03200862

Pappas, N., Katsimpras, G. and Stamatatos, E. (2012). Extracting Informative Textual Parts from Web Pages Containing User-Generated Content. *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* in , i-KNOW '12. New York, NY, USA: Association for Computing Machinery. doi:10.1145/2362456.2362462

Pradham, P., Younan, N. H. and King, R. L. (2008). 16 - Concepts of image fusion in remote sensing applications. T. B. T.-I. F. Stathaki (Ed.), (pp. 393–428). Oxford: Academic Press. doi:https://doi.org/10.1016/B978-0-12-372529-5.00019-6

Raggett, D. (1994). A review of the HTML + document format. *Computer Networks and ISDN Systems*, *27*(2), 135–145. doi:https://doi.org/10.1016/0169-7552(94)90127-9

Sarkar, P. (2019, 14 October). Bagging and Random Forest in Machine Learning: How do they work? 18 August 2020 tarihinde https://www.knowledgehut.com/blog/data-science/bagging-and-random-forest-in-machine-learning from retrieved.

Shuldiner, A. (2019). Chapter 8 - Raising Them Right: AI and the Internet of Big Things. W. Lawless, R. Mittu, D. Sofge, I. S. Moskowitz and S. B. T.-A. I. for the I. of E. Russell (Eds.), *Artificial Intelligence for the Internet of Everything* in (pp. 139–143). Academic Press. doi:https://doi.org/10.1016/B978-0-12-817636-8.00008-9

Štěpánek, J. and Šimková, M. (2013). Comparing Web Pages in Terms of Inner Structure. *Procedia - Social and Behavioral Sciences*, *83*, 458–462. doi:10.1016/j.sbspro.2013.06.090

Sullivan, W. (2018). *Decision Tree and Random Forest - Machine Learning and Algorithms: The Future Is Here!* CreateSpace Independent Publishing Platform. https://books.google.com.tr/books?id=x-u_tAEACAAJ from retrieved.

Uçar, E., Uzun, E. and Tüfekci, P. (2016). A novel algorithm for extracting the user reviews from web pages. *Journal of Information Science*, *43*(5), 696–712. doi:10.1177/0165551516666446

Uzun, E. (2014). iCrawler/Dataset at master · erdincuzun/iCrawler. 18 August 2020 tarihinde https://github.com/erdincuzun/iCrawler/tree/master/Dataset from retrieved.

Uzun, E., Agun, H. V. and Yerlikaya, T. (2013). A hybrid approach for extracting informative content from web pages. *Information Processing & Management*, *49*(4), 928–944. doi:https://doi.org/10.1016/j.ipm.2013.02.005

Uzun, E., Serdar Güner, E., Kılıçaslan, Y., Yerlikaya, T. and Agun, H. V. (2014). An effective and efficient Web content extractor for optimizing the crawling process. *Software: Practice and Experience*, *44*(10), 1181–1199. doi:10.1002/spe.2195

Wu, S., Liu, J. and Fan, J. (2015). Automatic Web Content Extraction by Combination of Learning and Grouping. *Proceedings of the 24th International Conference on World Wide Web* in , WWW '15 (pp. 1264–1274). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741659

Yang, D. and Song, J. (2010). Web Content Information Extraction Approach Based on Removing Noise and Content-Features. *2010 International Conference on Web Information Systems and Mining* in (Vol. 1, pp. 246–249). doi:10.1109/WISM.2010.82

Yucalar, F., Ozcift, A., Borandag, E. and Kilinc, D. (2020). Multiple-classifiers in software quality engineering: Combining predictors to improve software fault prediction ability. *Engineering Science and Technology, an International Journal*, *23*(4), 938–950. doi:https://doi.org/10.1016/j.jestch.2019.10.005