Web Sitelerinde Kimlik Avı
Phishing on Websites

Recep Ali AY
Elektrik Bilgisayar Mühendisliği
KTO Karatay Üniversitesi
Konya, Türkiye
recepaliay@gmail.com

Özetçe — Günümüzde web uygulamaları kullanıcılar tarafından yoğun olarak kullanılmaktadır. Kullanımın artmasına eş olarak web uygulamaların sayısında artış yaşanmaktadır. Bu artışlara birlikte kötü amaçlı kişiler tarafından kötücül web siteleri hazırlanmakta ve saldırılara maruz kalan son kullanıcıya ciddi zararlar vermektedir. Kişisel ve hassas bilgilerin çalınmasına yönelik bu saldırılardan biri Kimlik Avı saldırısıdır. Yayımlan an güvenlik raporlarında son yıllarda milyonlarca yeni kimlik avı sahteciliği yapan web sayfası tespit edildiği ifade edilmektedir. Böylesi kritik bir durumda bu web sayfalarının tespiti büyük önem arz etmektedir. Bu çalışmada, bir veri kümesi ile birlikte literatürde bulunan makine öğrenmesi sınıflandırma algoritmaları kullanılarak karşılaştırmalı analiz yapılmıştır. Analiz sonuçları, Kimlik Avı Sahteciliği çalışmalarında kullanılan sınıflandırma 4 algoritmanın verilerin analizindeki performansları kıyaslanacaktır.

Anahtar Kelimeler — Sınıflandırma algoritması; makine öğrenmesi; Multilayer perceptron; Random Forest; IBk; Naive Bayes

Abstract — Today, web applications are used extensively by users. As the usage increases, the number of web applications increases. Along with these increases, bad websites are being prepared by malicious people and serious damage is done to the end user exposed to the attack. One of these attacks for the theft of personal and sensitive information is a phishing attack. It is stated that in the security reports that have been published, a web page has been found that has made millions of new phishing scams in recent years. In such a critical situation, the identification of these web pages is crucial. In this study, a data set and a comparative analysis were made using machine learning classification algorithms in the literature. The analysis results will be compared to the performance of the analysis of the data in the classification algorithm used in Phonetic Counterfeiting trials.

Keywords — Classification algorithm; machine learning; Multilayer perceptron; Random Forest; IBk; Naive Bayes

I. GIRIŞ

Günümüzde web sitelerinin sayısı ve kullanımı büyük miktarda artmıştır . Bunun yanı sıra artan bu web site sayısı ve kullanım miktarı kötücül amaçlı kişileri de harekete geçirmiştir. Saldırganlar kötü amaçlı web siteleri ile internet kullanıcılarına zarar vermektedirler. Saldırganlar kötü amaçlı web sitelerini farklı şablonlarda kullanıcılara çekici kılarak sunmaktadırlar. Bu siteler ile ilgili mcafee'in raporuna göre [1], 2016 yılının 5. çeyreğinde 500 milyonu aşkın yeni kötücül url tespit edilmiştir. Bu raporda, her geçen yıl kötücül web siteleri url sayısında sürekli artış olduğu görülmektedir ve internet kullanıcılarına bu yöntemle ile zarar vermektedir. Literatürde bu alanda çözüm üretmek amacıyla çıkan bir çok çalışma bulunmakla birlikte, güvenlik raporları halen kesin bir çözüm elde edilemediğini göstermektedir [1]-[2]. UCI makine öğrenme deposu (machine learning repository) içinde bulunan kimlik avı sahteciliği yapan web site veri kümesi (phishing websites data set)[3] kullanılmıştır. Multilayer Perceptron, Random Forest, IBk ve Naive Bayes algoritmaları veri kümesi ile uygulanarak, veri kümesi üzerindeki performans ve doğruluk karşılaştırmaları analiz edilmiştir.
Bu çalışmada[4] makine öğrenmesi (machine language) kullanılarak veri kümesinin incelenmesi için Waikato üniversitesinde açık kaynak kodlu olarak JAVA dili üzerinde geliştirilmiştir ve GPL lisansı ile dağıtılan WEKA programı kullanılmıştır.

II. KONUYLA İLGİLİ YAPILMIŞ ÇALIŞMALAR

Kazemian ve Ahmed [5], yaptıkları çalışmada 100.000 web sayfasını bir örümcek (crawler/spider) aracılığı ile indirerek, web sayfalarının özellik vektörüne dönüşümünü

gerçekleştirmişlerdir. Web Uygulama Sınıflandırıcısı (Web Application Classifier-WAC) adı verilen araç ile bu vektörler giriş olarak alınıp Makine Öğrenmesi algoritmaları uygulanmıştır. Uygulanan Makine Öğrenmesi algoritmaları sırası ile; k-En Yakın Komşu, Doğrusal Destek Vektör Makineleri (Support Vector Machines, SVM), Radyal Tabanlı Fonksiyon (Radial Basis Function-RBF) Çekirdekli Destek Vektör Makineleri ve Naif Bayes'dir. Web sayfaları 50, 100, 500, 5.000 ve 100.000 sayıları ile 4 algoritma bazında test edilmiş ve en iyi sonuç Radyal Tabanlı Fonksiyon Çekirdekli Destek Vektör Makineleri (RBF-SVM) algoritması ile alınmıştır.

Li ve diğ. [6], Minimum Kapsayan Top-tabanlı Destek Vektör Makinesi (Minimum Enclosing Ball-based Support Vector Machine - BVM) adını verdikleri çalışmada ilk olarak veri kümesi oluştururken 12 özelliğe göre özellik vektörleri çıkartılmıştır. Sonrasında Destek Vektör Makineleri ve önerilen yöntem olarak Top-tabanlı Destek Vektör Makinesi'nin karşılaştırmalı analizi gerçeklenmiştir. Yapılan deneyler sonucunda Top-tabanlı Destek Vektör Makinesi'ninDestek Vektör Makineleri'ne göre doğruluk oranı ve performans konusunda daha iyi sonuçlar verdiği gözlemlenmiştir.

Moghimi ve Varjani [7], internet bankacılığında kimlik avı (phishing) saldırılarını tespit etmek için kural bazlı yeni bir yöntem sunmaktadır. Önerdikleri özellik kümesi, sayfa kaynağı kimliğini değerlendirmek için 4 özellik ve sayfa kaynak elemanlarının erişim protokollerini tanımlamak için 4 özellikiçermektedir. Önerilen özellik kümesindeki sayfaların URL'leri ile içerikleri arasındaki ilişkiyi belirlemek için string eşleştirme algoritması kullanılmıştır. Yapılan deneylerde, internet bankacılığındaki kimlik avı sahteciliği sayfalarını tespit etmek için önerilen yöntemin %99.14 doğruluk oranına (true-positive) sahip olduğu gözlemlenmiştir.

III. SINIFLANDIRMA UYGULAMASI VE DENEY

Kimlik avı sınıflandırılması uygulamasında toplam 11055 adet veri (data sayısı) bulunmaktadır. Öznitelik olarak ise (attribute) 30 tane attibute bir tane de class attribute içerisinde yer almakta olup toplamda 31 tane özellik bulunmaktadır.

Tablo 1: Phishing websites [3] veri kümesi öznitelikleri.

Having_IP_Address	Submitting_to_email
URL_Length	Abnormal_URL
Shortining_Service	Redirect
having_At_Symbol	on_mouseover
double_slash_redirecting	RightClick
Prefix_Suffix	popUpWidnow
having_Sub_Domain	Iframe
SSLfinal_State	age_of_domain
Domain_registeration_length	DNSRecord
Favicon	web_traffic
port	Page_Rank
HTTPS_token	Google_Index
Request_URL	Links_pointing_to_page
URL_of_Anchor	Statistical_report
Links_in_tags	Result
SFH	

Çalışmada 30 öznitelik sayınında bir azaltmak için öznitelik seçici metodu olan CFS kullanılmıştır. CFS, diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan öznitelikleri seçer. CFS arama metotları içerisinden; Best First, Exhaustive Search, Genetic Search, Linear Rorward Selection, Random Search, RankSearch, Scatter SearchV1 ve Subset Size Forward Selection 9 adet ve aynı özniteliği seçmiştir.

- Prefix_Suffix: Alan adı ön ek veya son ek ile ayrılmışmı (- işareti ile ayrılmışmı) (Alan adı '-' işareti içeriyor ise -1, içermiyor ise 1),
- having_Sub_Domain: URL'de alt alan adı veya çoklu alt alan adı kullanılmışmı (Alan adı kısmında nokta işareti sayısı 1 ise 1, nokta işareti sayısı 2 ise 0, diğer durumlarda ise -1),
- SSLfinal_State: HTTPS protokolünün varlığı (https kullanılıyor ve sağlayıcı güvenilir ve sertifika yaşı 1 yıl ve üstü ise 1, https kullanılıyor ve sağlayıcı güvenilir değil ise 0, diğer durumlarda -1),
- Request_URL: Web sayfası farklı alan adlarından farklı nesneleri çekiyormu (Harici web sitelerinden çekilen nesne istek URL'lerinin yüzdesi 22'den az ise 1, yüzdesi 22 ve 61 arasında ise 0, diğer durumlarda ise -1),
- URL_of_Anchor: HTML çapa etiketi (<a> etiketi) kullanımının URL'de varlığı (Çapa etiketlerindeki URL varlığının yüzdesi 31 değerinin altında ise 1, yüzdesi 31 ve 67 arasında ise 0, diğer durumlarda -1),
- Links_in_tags: Meta, Script ve Link etiketlerindeki bağlantıların oranı (Meta,Script ve Link etiketlerindeki bağlantıların yüzdesi 17 değerinden az ise 1, yüzdesi 17 ve 81 arasında ise 0, diğer durumlarda -1),
- SFH: Sunucu Form Tutucusu/İşleyicisi (Server Form Handler) içinde boş string değişkeni barındırılıyormu (SFH 'about:blank' veya string değeri boş ise -1, farklı bir alan adına referans veriyor ise 0, diğer durumlarda1),
- web_traffic: Alexa web site trafik durumu (Alexa sırası 100.000 altında ise 1, üstünde ise 0, hiç liste de yok ise-1),
- Google_Index: Google tarafından indeksleniyormu (Google tarafından indeksleniyor ise 1, diğer durumda -1),
- Result: Son parametre ise kimlik avı olarak işaretlenip işaretlenmediğini belirten sınıf alanıdır. (Kimlik avı sahteciliği olarak işaretlenmiş ise -1, iyicil olarak işaretlenmiş ise 1 değeri alır),

Bu çalışmada kullanılan metrikler aşağıda verilmiştir.

- Doğruluk(accuracy) [8]: Doğru olarak sınıflandırılanörneklerin toplam örnek sayısına oranıdır.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Kesinlik(precision) [8]: Pozitif olarak etiketlenen örneklerin sayısının pozitif olarak sınıflandırılan toplam örneklere oranıdır.

$$precision = \frac{TP}{TP + FP}$$

- Duyarlılık(recall) [8]: Pozitif olarak etiketlenen örneklerin gerçekten pozitif olan örneklerin toplam sayısına eşittir.

$$recall = \frac{TP}{TP + FN}$$

- F-ölçütü(f-measure):[8] Kesinlik ve duyarlılık değerleri-nin harmonik ortalamasıdır.

$$f - measure = 2 * \frac{precision * recall}{precision + recall}$$

- Ortalama mutlak hata(mean absolute error) [9]: Mutlak hataların ortalamasıdır. Hata, tahmin edilen değer ile gerçek değer arasındaki farktır.

CFS metotu ile seçilen özniteliklerin sıralama algoritmalarında yapılan analizinin Kesinlik, Duyarlık, F-ölçütü sonuçları Tablo-1'de gösterilmiştir.

Algortimalar	Precision	Recall	F-Measure
Naive Bayes	0.926	0.925	0.925
IBl	0.935	0.935	0.935
Random Forest	**0.954**	**0.954**	**0.954**
Multilayer Preceotion	0.951	0.951	0.951

Tablo-1

Sınıflandırma işleminin doğru ve hatalı yapılanların sayısı Tablo-2'de verilmiştir.

Algortimalar	Kimlik Avı Var/Hata	Kimlik Avı Yok/Hata
Naive Bayes	4386/512	5856/301
IBl	4486/412	5768/389
Random Forest	**4591/307**	**5884/273**
Multilayer Preceotion	4551/347	5904/253

Tablo-2

Yapılan hata türleri ve miktarları Tablo-3'de verilmiştir.

Algortimalar	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Naive Bayes	0.0917	0.2312	18.587%	46.547%
IBl	0.0725	0.2692	14.681%	54.187%
Random Forest	**0.0745**	**0.1966**	**15.101%**	**39.581%**
Multilayer Preceotion	0.073	0.2012	14.790%	40.506%

Tablo-3

IV. SONUÇLAR

KAYNAKLAR

[1]McAfee Inc. "McAfee Labs Threats Report-February2017". https://www.mcafee.com/us/resources/reports/rp-quarterly-threats-mar-2017.pdf

[2]McAfee Inc. "McAfee Labs Threats Report-February2015". http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q4-2014.pdf(26.03.2016).

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

[5]Kazemian HB, Ahmed S. "Comparisons of machine learning techniques for detecting malicious webpages".Expert Systems with Applications, 42(3), 1166-1177, 2015.

[6]Li Y, Yang L, Ding J. "A minimum enclosing ball-based support vector machine appro ach for detection of phishing websites". Optik, 127(1). 345-351. 2016.

[7]Moghimi M, Varjani AY. "New rule-based phishing detection method". Expert Systems with Applications, 53, 231-242. 2016.

[8]O. Kaynar, Y. Görmez, M. Yıldız, and A. Albayrak, "Makine Öğrenmesi Yöntemleri ile Duygu Analizi Sentiment Analysis with Machine Learning Techniques," no. September, 2016.

[9]https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance