Araştırma Makalesi / Research Article

K-NN, NN ve Feature Selection yöntemleri ile firewall verilerinin sınıflandırması

Sinan DEMİR[1]*
Zafer ASLAN[2]

Geliş tarihi / Received: 02.01.2023
Düzeltilerek geliş tarihi / Received in revised form: 10.01.2023
Kabul tarihi / Accepted: 14.01.2023
DOI: 10.17932/IAU.ABMYOD.2006.005/abmyod_v17i66003

Öz

Günümüzde internet kullanımının yaygınlaşması, internet güvenliği konusunun önemini artırmıştır. Kişisel bilgilerin, şifrelerin ve diğer hassas bilgilerin korsanlarca ele geçirilmesi veya sahte siteler aracılığıyla hile yapılması, internet kullanıcıları için ciddi riskler oluşturmaktadır. Güvenli internet kullanımı için, kullanıcıların bilinçli olmaları ve güvenliğe dair önlemleri almaları gerekir. Örneğin, şifreleri sık sık değiştirmek, güvenli bağlantıları kullanmak ve elektronik cihazları (bilgisayar, telefon, tablet vb.) güncel güvenlik yazılımları ile koruma altına almak gerekir. Bunun yanında ise kullanmakta olduğumuz cihazları Firewall (Güvenlik Duvarı) teknolojisi ile koruma altına almak güvenlik konusunda büyük önem arz etmektedir. Bu çalışma içerisinde, Fırat Üniversitesinin firewall cihazından elde edilen 65.532 adet log kaydının NN (Neural Network) ve K-NN (K-Nearest Neighbor) algoritmaları kullanılarak sınıflandırma işlemi uygulanmıştır. Bununla birlikte feature selection teknikleri ile de veri seti içerisindeki kolonların önemi ve benzerlik oranları belirlenmiştir. Neural Network algoritmasında "Adam" fonksiyonu optimizasyonunda %98,46, K-NN algoritmasında k değeri 20 iken en başarılı sonuç manhattan'da %99,08 olarak belirlenmiştir. Daha önce literatürde aynı veri seti ile yapılmış olan SVM çalışmasında ise dört SVM tekniği arasında en başarılı teknik %98,5 olarak Sigmoid fonksiyonun da ulaşılmıştır.

Anahtar Kelimeler: *Firewall, Makine Öğrenimi, Siber Güvenlik, Derin Öğrenme*

[1]İstanbul Aydin Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği,34295, Küçükçekmece, İstanbul *ORCID ID 1: http://orcid.org/0000-0002-0753-7244 sinandemir4@stu.aydin.edu.tr*
[2]İstanbul Aydın Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, 34295, Küçükçekmece, İstanbul *ORCID ID 2: http://orcid.org/0000-0001-7707-7370 zaferaslan@aydin.edu.tr*

Classification of firewall data with K-NN, NN and Feature Selection methods

Abstract

The widespread use of the internet today has increased the importance of internet security. The compromise of personal information, passwords, and other sensitive information by hackers or fraudulent websites creates serious risks for internet users. To use the internet securely, users need to be aware and take precautions for their security. For example, they should change their passwords frequently, use secure connections, and protect electronic devices (such as computers, phones, tablets, etc.) with up-to-date security software.In addition, protecting the devices we use with Firewall technology is of great importance for security. In this study, a classification process was applied using the NN (Neural Network) and K-NN (K-Nearest Neighbor) algorithms with 65,532 log records obtained from Firat University's firewall device. Moreover, feature selection techniques were used to determine the importance and similarity rates of the columns in the dataset.In the Neural Network algorithm, the 'Adam' function optimization resulted in 98.46% success, while in the K-NN algorithm, the most successful result was achieved with a k-value of 20 and in the Manhattan distance with a success rate of 99.08%.In a previous study using the same dataset, the most successful technique among four SVM techniques was achieved with a success rate of 98.5% using the Sigmoid function.

***Keywords**: Firewall, Machine Learning, Cyber Security, Deep Learning*

Giriş

Bu bölümde, çalışmamızda kullanılmış olan ve örnek olarak temel aldığımız yazın araştırmalarını ele alacağız. Literatür taramalarını seçerken, öncelikle çalışmasını gerçekleştirdiğimiz konumuz ile benzer olmalarına dikkat ederek, analiz yöntemlerini incelemek için çalışmamıza fayda sağlayacak kaynaklar tercih edilmiştir. Öncelikli olarak Fırat Üniversitesi firewall veri seti üzerinde yapılan çalışmada (Kaya & Ertam, 2018) SVM algoritması yöntemini kullanmış ve belirli bir başarı elde etmiştir. Yapılan çalışma sonucunda dört SVM tekniği arasında en başarılı teknik 98.5% olarak SVM Sigmoid fonksiyonu olarak belirlenmiştir. Performans ölçümleri için sentivity, recall ve harmonic mean skorları performans hesaplamasında kullanılmıştır. Farklı denetimli öğrenme

algoritmaları ve farklı veri analizi metotlarının kullanıldığı (Uçar & Özhan, 2017) çalışma da ise en başarılı performans değerleri denetimli öğrenme algoritmaları tarafından elde edilmiştir. Çalışma içerisinde elde edilen başarılar sonrasında firewall kuralları üzerindeki anormallikleri belirleyip ve yönetme durumları için makine öğrenmesi algoritmalarının başarılı olduğu öne sürülmüştür. Çalışma içerisinde kullanılan farklı tekniklerin arasında ön plana çıkan KNN algoritması için yüksek veri içeren eğitim verileri üzerinde doğası gereği yüksek performans ile çalışması gözlemlenmiştir. (Sunar, Özkan, & Taberner, 2004) Yapmış olduğu çalışma içerisinde NN kullanarak arazi sınıflandırması gerçekleştirmiştir. Bu çalışma içerisinde ise hedef sınıf sayının oldukça yüksek olduğu gözlemlenmiştir. NN kullanılarak birden fazla sınıf tahmini aşamasında fazla seçeneğin bulunması, yüksek performansı etkilemediği ve başarıyla çalıştığı saptanmıştır. Makine öğrenmesi teknikleriyle internet saldırı tespitlerinin karşılaştırmalı analiz gerçekleştirildiği çalışmada ise 2007-2013 yılları arasında makine öğrenme teknikleri ile gerçekleştirilmiş saldırı tespit sistemlerinde en sık kullanılan yöntemin yapay sinir ağları belirlenmiştir (Kaya & Yıldız, 2014) . Özellik seçim yöntemleri arasında çalışma gerçekleştiren (Budak, 2018), filtreleme özellik seçim yöntemleri arasında sıkça kullanılan Fisher Skor yöntemine göre, çalışmasında kullandığı yönteme ait sınıflandırma doğruluk yüzde ortalamasının (%72,59) Fisher Skor yöntemi ortalamasından (%65,74) yüksek olduğunu belirlemiştir. Günümüzde veri madenciliği konusu son yıllarda birçok sektörde kullanılmasına rağmen ülkemizde en çok tıp alanında yaygın olarak kullanıldığı gözlemlenmiştir. (Köktürk, 2012) Bülent Ecevit Üniversitesi Uygulama ve Araştırma Hastanesi Kadın Hastalıkları ve Doğum Polikliniği'ne başvuran erken ve zamanında doğum yapan gebelerden elde edilen veri seti ile sınıflandırma algoritmaları kullanarak K-NN algoritmasında %78,3, NN algoritmasın da ise %90,8'lik sınıflandırma başarısı elde etmiştir. Veri madenciliğinde sınıflandırma ve kümeleme yöntemleri ile ağdaki trafikten elde edilen veri seti üzerinden çalışma yaparak bu çalışma yöntemlerinin birbirleri arasındaki iyi ve kötü yönlerin belirlenerek ağ üzerindeki olumlu ve olumsuz haraketleri birbirinden ayırıp performans artışının sağlanması için K-means-KNN ve K-medoids-KNN yöntemleri ile bir çalışma gerçekleştiren (Çalışkan, 2008) saldırı tespit oranını %99'a yükseltmiştir.

Materyal ve metod

Bu çalışmada Fırat Üniversitesinin Firewall cihazından elde edilmiş olan 65532 adet log kaydının bulunduğu veri setinin içerisindeki kişisel ve önem arz edebilecek tüm bilgiler kaldırılmıştır. Geriye kalan hedef port, nat port, işlem, gönderilen ve alınan byte bilgileri, paketleme için harcanan süre ve paket gönderilen\alınan

bilgileri ile Neural Network ve K-NN algoritmasında sınıflandırma çalışması yapılmıştır. Bununla birlikte ise Feature Selection yöntemleri ile veri seti içindeki kolonların önemi ve benzerlik oranları saptanmıştır.

K-Nearest Neighbor (En yakın komşu)

Veri setinde denetimli öğrenme (supervised machine learning) modeli olarak daha önceden yapılmış olan çalışmalar incelenerek KNN üzerindeki başarılı performanslar dikkat çekmiştir. Bu başarının ardından (Kaya & Ertam, 2018) tarafından yapılan çalışmada SVM ile yaptıkları çalışmanın KNN ile karşılaştırılması amacıyla çalışmalar gerçekleştirilmiştir. Literatürde KNN algoritmasının diğer algoritmalara göre yaygın olarak kullanılma sebebi ise daha hızlı ve kabul edilebilir bir başarı oranına sahip olması olarak açıklanmıştır. (Laribi, 2018) KNN üzerinde öklid, manhattan ve minkowski mesafe formülleri kullanılarak ayrı ayrı hesaplanmıştır. Veri setinde bulunan yaklaşık 65000 öğe üzerinde k değerlerimiz () 25 olarak ele alınarak komşu değerler ile incelenmiş ve performans sonuçları verilmiştir. Tablo 1 Performans sonuçları ‘ında üzere 3 farklı uzaklık hesaplama yönteminde en başarılı yöntemin k=20 iken manhattan uzaklık hesaplama yönteminde % 99,08 olarak belirlenmiştir.

Neural network (yapay sinir ağları)

NN eğitimi sırasında veri dağılımı (eğitim-test) 0.67 eğitim ve 0.33 test olarak ayarlanmış, batch:64 ve epoch:50 olarak kullanılmıştır. Alternatif aktivasyon fonksiyonları deneme aşamasında çıktı katmanı (output layer – dense_5) sabit olarak ‘softmax’ kullanılmıştır. Çıktı olarak dört farklı sınıf için yüzdelikleri elde etme amaçlı olarak ayrı tutulmuş olup diğer tüm katmanların aktivasyon fonksiyonu değiştirilmiştir.

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 11)	132
dense_2 (Dense)	(None, 64)	768
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 4)	68

Şekil 1 *NN modeli özeti*

Şekil 1 NN modeli özeti tablosunda görüldüğü üzere Neural Network katmanlarını inceleyecek olursak;

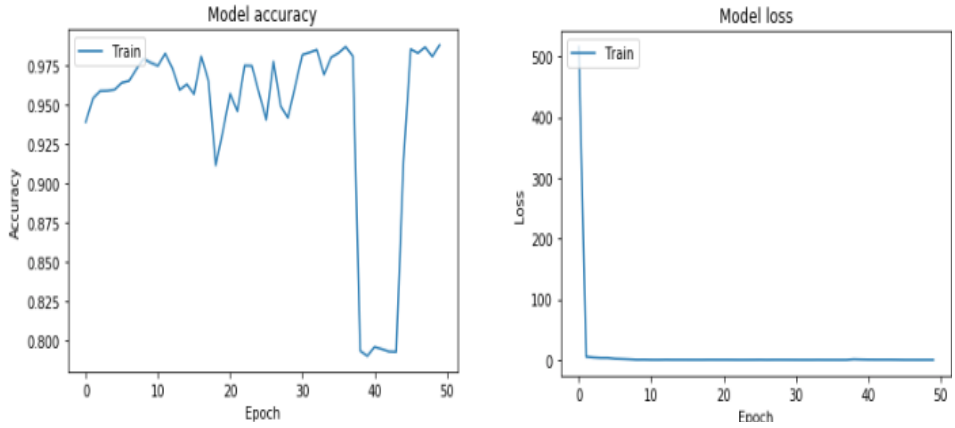
- Giriş katmanı(dense_1), veri setimizden gelen 11 özellik (feature) alanı bulunmasından dolayı bu katmanda 11 adet nöron bulunmaktadır.
- Gizli katman(dense_2,3,4), ağırlık hesaplamaları için kullanılan katmandır.
- Çıktı katmanı(dense_5), 4 adet sınıfımız bulunan veri setimizdeki her bir sınıfa ait bir çıktı değeri üretmesi amacıyla 4 adet çıktı nöronu bulunmaktadır.

Tablo 2 NN performans sonuçları tablosunda görüldüğü üzere Relu, Sigmoid ve Tanh aktivasyon fonksiyonlarının sonuçları gösterilmiştir. NN üzerinde en başarılı performans 'Adam' optimizasyon ve 'Relu' aktivasyon fonksiyonu kullanılarak elde edilmiştir.

Tablo 2 *NN performans sonuçları*

Aktivasyon Fonksiyonları	Optimizasyon Fonksiyonu	
	Adam	SGD
Relu	98,46	57,59
Sigmoid	97,60	77,54
Tanh	97,70	69,23

Resim 1 Adam & Relu Performans Grafiği gösteriyor ki eğitilen modelin başarı ortalaması 95% civarlarında olmuş ve eğitimi 98,46% gibi yüksek bir başarıyla tamamlamıştır.



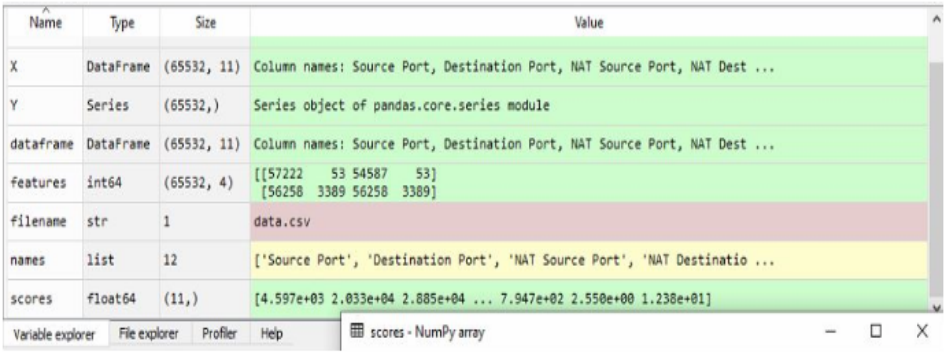
Resim 2 Adam & Relu Hata Grafiğine bakıldığında ise hata değeri ilk epochlar da çok büyük bir öğrenme ile hata değerinin düştüğü görülmekte olup diğer epochlar da ise hata değeri çok küçük değerlere düştüğü görülmektedir. NN ile eğitilen bu modelimizde hata değerimiz ilk epochlar da çok büyük bir öğrenme ile hata değeri düşmüş olup ileri ki epochlarda hata değeri çok küçük değerler ile değişmiştir. NN üzerinde epoch ve batch değerlerini değiştirerek daha farklı, belki daha başarılı sonuçların alınabileceği göz ardı edilmemelidir.

Feature selection (özellik seçimi)

Feature selection, bir veri kümesinde bulunan özelliklerin seçimi veya atlanması işlemidir. Bu işlem, özelliklerin sayısını azaltmak veya veri kümesinin özellikler arasındaki ilişkileri anlamlandırmak için kullanılır. Bu çalışmada feature selection'ın 3 farklı yöntemi kullanılmıştır. Bunlar; Universe Selection, Principal Component Analysis ve Feature Importance' dır.

Universe selection (tek değişkenli seçim)

Feature Selection yönteminde ilk olarak Universe Selection (Tek değişkenli seçim) özelliği gerçekleştirilmiştir. Bu yöntemdeki amaç veri setinde bulunan kolonların önemini bulmaktır. Veri setine baktığımızda 13 tane kolondan 3 tane kolonun en önemli olabileceğini ve Universe Selection yöntemi ile bu 3 kolonla yine aynı sonuca ulaşabileceğimiz anlaşılmaktadır. Resim 3 Universe selection sonuç da görüldüğü üzere bu alanların Source Port, Destination Port ve Nat Destination Port olduğunu söyleyebiliriz.



Resim 3 *Universe selection sonuç*

Principal component analysis (temel bileşen analizi)

Feature Selection da ikinci adım olarak Principal Component Analysis yani Temel bileşen analiz adımını gerçekleştirdim. Buradaki amacımız ise kolonlar arası benzerlik özelliğini belirlemektir. Resim 4 Principal component analysis sonuçları gösterilmektedir.

```
Explained Variance: [9.297e-01 7.023e-02 1.090e-05]
[[ 1.588e-07 -1.173e-05  3.038e-05  6.310e-06  8.049e-01  5.211e-01
   2.838e-01  7.144e-04  6.376e-06  4.500e-04  2.644e-04]
 [ 2.530e-05 -1.800e-04  2.518e-04 -5.841e-05  1.370e-01 -6.286e-01
   7.656e-01  2.857e-04  6.765e-06 -2.597e-04  5.454e-04]
 [-2.616e-01  5.180e-01 -8.142e-01  1.688e-02  6.673e-05 -1.762e-04
   2.429e-04 -5.130e-05 -1.076e-03 -1.345e-04  8.316e-05]]
```

Resim 4 *Principal component analysis sonuç*

Feature ımportance (özellik önemi)

Feature Selection'da son olarak feature importance (özelliğin önemi) adımı gerçekleştirilmiştir. Bu kısımda ise amaç kolonların önem derecesini ölçüp hangi kolonlar ile çalışma gerçekletirilse yine aynı sonuca varılabileceğini bulmaktır. Bu çalışmadaki veri setinde 13 kolondan 2 kolonun önem derecesini belirttiğini Resim 5 Feature Importance Sonuç'nda görebilmekteyiz. Ancak çok fazla kolon olduğunu düşündüğümüz çalışmalarda bu sayı artacaktır.

```
In [41]: from pandas import read_csv
    ...: from sklearn.ensemble import ExtraTreesClassifier
    ...: # Load data
    ...:
    ...:
    ...: names = ['Source Port','Destination Port','NAT Source Port','NAT Destination
Port','Bytes','Bytes Sent','Bytes Received','Packets','Elapsed Time
(sec)','pkts_sent','pkts_received','Action']
    ...: dataframe = read_csv(filename,sep=';')
    ...:
    ...: from sklearn import preprocessing
    ...: le = preprocessing.LabelEncoder()
    ...:
    ...: dataframe['Action'] = le.fit_transform(dataframe["Action"])
    ...: Y = dataframe['Action']
    ...: Action = dataframe['Action']
    ...:
    ...: dataframe = dataframe.drop(columns=['Action'])
    ...: X = dataframe
    ...:
    ...:
    ...: # feature extraction
    ...: model = ExtraTreesClassifier(n_estimators=10)
    ...: model.fit(X, Y)
    ...: print(model.feature_importances_)
[0.05  0.315 0.496 0.073 0.008 0.008 0.003 0.014 0.012 0.006 0.015]
```

Resim 5 *Feature Importance Sonuç*

Tartışma ve sonuç

Yapılan bu çalışma, daha önce (Kaya & Ertam, 2018) tarafından kullanılmış olan Fırat Üniversitesinin Firewall veri seti ile gerçekleştirilmiştir. Yapılan literatür çalışmalarında incelenilen birçok çalışmanın Neural Network (NN), K-Nearest Neighbors (K-NN) ve veri madenciliği yöntemleri kullanılmasından dolayı, (Kaya & Ertam, 2018) tarafından SVM üzerinde yapılan çalışmanın diğer algoritmalar ile karşılaştırılması yapılmış ve bununla birlikte Feature Selection yöntemi kullanılmıştır. Çalışma sonuçlarında görüldüğü üzere K-Nearest Neighbors algoritmasında daha önce yapılmış çalışmalara göre en yüksek başarı %99,08 olarak belirlenmiştir. Neural Network algoritmasında ise "Adam" fonksiyonu optimizasyonunda %98,46 olarak sonucu elde edilmiştir. Bu algoritmaların sonuçlarına bakarak, daha önce literatürde aynı veri seti ile gerçekleştirilmiş olan SVM algoritmasının sigmoid yöntemin de ulaşılmış olan 98.5%'lik sonucuna hemen hemen yakın oldukları gözlemlenmektedir. Daha sonra yapılacak olan çalışmalarda burada kullanmış veri setinden daha büyük veri seti kullanılarak algoritmalar arası kıyaslamalar yapılabilir.

Kaynaklar

[1]Budak, H. (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 21-31.

[2]Çalışkan, S. K. (2008). Gebze: Gebze Yüksek Teknoloji Enstitüsü Mühendislik ve Fen Bilimleri Enstitüsü.

[3]Kaya, Ç., & Yıldız, O. (2014). Makine Öğrenmesi Teknikleriyle Saldırı Tespiti: Karşılaştırmalı Analiz. *Marmara Fen Bilimleri Dergisi*, 89-104.

[4]Kaya, M., & Ertam, F. (2018). Classification of Firewall Log Files with Multiclass Support Vector Machine. Antalya.

[5]Köktürk, F. (2012). *K-En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması.* Zonguldak: Bülent Ecevit Üniversitesi Sağlık Bilimleri Enstitüsü.

[6]Laribi, P. (2018). *Genetik Algoritma ve K-En Yakın Komşu Kullanarak Metin Belgelerinin Sınıflandırılması.* Van: Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik Mühendisliği Anabilim Dalı.

[7]Sunar, F., Özkan, C., & Taberner, M. (2004). Comparison of maximum likelihood classification method with supervised artificial neural network algorithms for land use activities. *International Journal of Remote Sensing, 25*(9), 1733-1748.

[8]Uçar, E., & Özhan, E. (2017). The Analysis of Firewall Policy Through Machine. *Wireless Personal Communications, 96*(10), 1-19.

Tablo 1 *Performans sonuçları*

	K	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Öklid	%	99,01	98,97	98,95	98,93	98,91	97,81	98,98	98,85	98,84	98,83	98,82
Manhattan	%	99,08	99,01	99,00	98,98	98,97	97,81	98,94	98,92	98,91	98,87	98,87
Minkowski	%	99,02	99,00	98,96	98,96	98,95	97,87	98,89	98,87	98,86	98,84	98,83