Şifreli İnternet Trafiğinin Gerçek Zamanlı Sınıflandırılması

Cihangir Beşiktaş[1] Hacı Ali Mantar[2]

[1,2]Bilgisayar Mühendisliği Bölümü, Gebze Yüksek Teknoloji Enstitüsü, Kocaeli
[1]TÜBİTAK BİLGEM, Bilişim ve Bilgi Güvenliği İleri Teknolojiler Araştırma Merkezi, Kocaeli
[1]e-posta: cihangirbesiktas@uekae.tubitak.gov.tr [2] e-posta: hamantar@bilmuh.gyte.edu.tr

Özetçe

İnternet trafik akışlarının uygulama veya protokol bazında sınıflandırılması ağ izleme, hizmet kalitesi, nüfuz tespiti, ağ güvenliği, trend analizi vb. alanlarda etkin bir şekilde kullanılmaktadır. Son zamanlarda bazı P2P uygulamalarının dinamik port numaraları, maskeleme teknikleri kullanmaları port bazlı tespit sistemlerinin yetersiz kalmasına neden olmuştur. Bu engellere alternatif olarak geliştirilen yük bazlı tespit sistemleri ise uygulamaların güvenlik ve takip edilmeme amacıyla şifreleme mekanizmalarına gitmesi ve uygulama katmanında imza arama işlemlerinin çok maliyetli olması nedeniyle yerini makine öğrenmesi ve istatistiki verilere dayanan yaklaşımlara bırakmıştır.

Bu çalışma gerçek zamanlı olarak şifreli internet trafiğini de tespit edebilen makine öğrenmesi tabanlı bir yöntemi ele almaktadır. Bu yöntemde öğrenme aşamasında trafik akışlarının ilk bir kaç paketinin paket boyutlarından uygulamalara özgü paket boyutları vektörü ve normalize edilmiş standart sapma vektörü karakteristikleri çıkartılmaktadır. Yeni bir akışın sınıflandırması ise her bir uygulamanın karakteristiği ile yapılan ağırlıklı kosinüs benzerliği hesabına göre yapılmaktadır. Deneysel sonuçlar önerilen yöntemin yüksek bir başarıma sahip olduğunu göstermektedir.

1. Giriş

Gerçek zamanlı internet trafik sınıflandırması internet servis sağlayıcılarına nüfuz tespiti ile güvenlikte, hizmet kalitesi ile ağın performanslı kullanılmasında, trend analizlerinde, protokol bazlı filtreleme ile farklı kategorilerde internet hizmet peketlerinin oluşturulmasında fayda sağlamaktadır. Ayrıca hükümetlerin koyduğu yükümlülükler çerçevesinde IP veri trafiğinin yasal olarak izlenmesinde de etkin bir rol oynamaktadır.

İnternet trafik sınıflandırılması ile aynı kaynak IP, port, hedef IP, port ve iletim protokolüne (TCP, UDP) sahip olan paketlerin oluşturduğu ağ akışlarının hangi uygulamaya veya protokole ait olduğu tespit edilir. En basit anlamda bu işlem ağ akışlarının IANA tarafından belirlenen protokollerin klasik TCP veya UDP port numaralarına [1] bakılarak gerçekleştirilir. Fakat bazı uygulamaların dinamik port numaraları kullanmaları, bazı uygulamaların da güvenlik duvarlarını aşmak için http, ftp gibi servislerin genelde internete açık olan klasik port numaralarını kullanması port bazlı yaklaşımları yetersiz kılmıştır. Moore ve Papagiannaki [2] port bazlı yaklaşımların %70 başarım sağladığını tespit etmişlerdir.

Port bazlı yaklaşımların yetersiz oluşu araştırma dünyasını yük bazlı yaklaşımlara yöneltmiştir [3, 4].

Yük bazlı yaklaşımlarda ağ akışlarındaki paketlerin uygulama katmanlarına bakılarak derinlemesine paket incelemesi yapılır. Paketlerin yüklerinde belirli imzalar aranarak ağ akışının hangi protokole ait olduğu tespit edilmeye çalışılır. Yük bazlı yaklaşımların ise aşağıda belirtilen sıkıntıları vardır:

- Kendine özgü yapıya sahip olan veya kaliteli dokümantasyona sahip olmayan protokoller için ileri seviyede tersine mühendislik çalışması gerektirmektedir.
- Şifreleme kullanan protokollerin tespit edilmesi mümkün değildir.
- Periyodik olarak protokollerin imzalarının güncellenmesi gerekmektedir.
- Paketler üzerinde yapılan imza arama işlemleri CPU ve bellek kaynaklarını yoğun bir şekilde tüketmektedir.

Yük bazlı yaklaşımların yukarıda listelenen problemlerinden dolayı CPU ve bellek kaynaklarını daha az tüketen ve şifreli trafiğin tespit edilmesine de olanak sağlayan makine öğrenmesi tabanlı istatistiki yaklaşımlar üzerine son zamanlarda birçok çalışma yapılmaktadır. Makine öğrenmesi tabanlı yaklaşımların uygulamaları bir kaç adımdan oluşmaktadır. Önce ağ akışlarının ayırt edilmesini sağlayan özellikler çıkartılır. Bu özellikler maksimum, minimum paket uzunlukları, paket uzunluklarının ortası, ortalaması, standart sapması, paketler arası süreler gibi ağ akışlarına ait davranışlardır. Daha sonra ise öğrenme aşamasına geçilir. Bu aşamada her bir uygulama veya protokol için seçilen özellikler cinsinden bir karakteristik belirlenir. Son olarak ise yeni gelen bir ağ akışının hangi uygulamaya veya protokole ait olduğunu öğrenilen karakteristiklerle karşılaştırma yaparak tespit eden sınıflandırma algoritması uygulanır.

Bu çalışma makine öğrenmesi metodolojisini kullanarak IP trafiğinin gerçek zamanlı olarak sınıflandırılması üzerine yoğunlaşmıştır. Önerilen yaklaşım ağ akışlarının ilk bir kaç paketinin paket uzunlukları ve normalize edilmiş standart sapmaları üzerinde kosinüs benzerliği hesaplaması ile sınıflandırma yapmaktadır.

Bu bildirinin kalan kısmı şu bölümlerden oluşmaktadır. Bölüm 2 makine öğrenmesi tabanlı IP trafik sınıflandırması ile ilgili yapılan çalışmalara yer vermektedir. Bölüm 3 yaklaşımımızın metodolojisini anlatmakta, bölüm 4 yaklaşımımızın deneysel sonuçlarını içermektedir. Bölüm 5 ve 6 da ise gelecek çalışmalar ve sonuçlar anlatılmıştır.

2. Benzer Çalışmalar

Makine öğrenmesi tabanlı IP trafik sınıflandırması çalışmaları metodolojileri açısından üçe ayrılmaktadırlar:

- Güdümlü sınıflandırma
- Güdümsüz sınıflandırma
- Karma sınıflandırma

Güdümlü sınıflandırmada öğrenme aşamasında kullanılan veri

kümesindeki her bir verinin hangi uygulamaya ait olduğu önceden etiketlenir ve bu etiketlere göre karakteristikler çıkartılır. Dolayısıyla öğrenme aşamasındaki veri kümesinin port ve/veya yük bazlı trafik tespiti yapan yazılımlar aracılığı ile veya manuel olarak önceden etiketlenmesi gerekmektedir.
Güdümsüz sınıflandırmada ise veri kümesindeki veriler önceden etiketli değildir ve bu verilerden benzer karakteristiklere sahip olanlar gruplanır. Bu gruplar üzerinde daha sonradan etiketleme işlemi yapılır.
Karma sınıflandırma çalışmaları ise hem güdümlü hem de güdümsüz sınıflandırma yaklaşımlarını birlikte kullanırlar.
Makine öğrenmesi tabanlı IP trafik sınıflandırması çalışmalarının ilk örneklerinden biri McGregor ve diğerlerinin [5] 2004 yılında Beklenti Maksimizasyonu algoritmasını kullanarak gerçekleştirdikleri yaklaşımdır. Bu yaklaşım güdümsüz sınıflandırma yaparak HTTP, FTP, SMTP, IMAP, NTP ve DNS protokollerini tespit etmeye çalışmıştır.
Güdümsüz sınıflandırma çalışmalarına örnek olarak Zander ve diğerlerinin [6] Beklenti Maksimizasyonu algoritmasını kullanan Bayesian sınıflandırıcısı, Bernaille ve diğerlerinin [7] basit K-Means algoritması üzerine yaptıkları çalışma, Erman ve diğerlerinin [8] öklit uzaklığı ve K-Means algoritmasını kullanarak yaptıkları çalışma gösterilebilir.
Güdümlü sınıflandırma çalışmalarında ise Saf Bayes, Saf Bayes Çekirdek Tahmini, Bayesian Ağlar, C4.5 Karar Ağaçları, k-En Yakın Komşular, Sinir Ağları ve Destek Vektör Makineleri genelde tercih edilen algoritmalar olmuştur.
Roughan ve diğerleri [9] NN, LDA ve QDA algoritmaları üzerine yaptıkları çalışmada paket uzunluklarının ortası, varyansı, kök orta karesi, ağ akışlarının süresi, toplam boyutu, toplam paketleri, bağlantıların TCP pencere boyutu, bant genişliği dağılımı gibi özellikleri kullanmışlardır.
Moore ve Zuev [10] ise ağ akışlarına ait 248 özelliği Saf Bayes algoritmasını kullanan sınıflandırıcılarını eğitmek için kullanmışlardır. Bu algoritma ile %65 akış keskinliği yakalamışlardır. Daha sonra [11] de çalışmaları genişletilerek %90 üzerinde akış keskinliği yakalanan Bayesian Sinir Ağı yaklaşımı önerilmiştir.
[9, 10, 11] deki çalışmalar ağ akışlarına ait bir çok özelliği kullandıkları için gerçek zamanda sınıflandırma özellikleri yoktur. Gerçek zamanda sınıflandırma çalışmalarından biri Crotti ve diğerlerinin [12] ağ akışlarının ilk paket uzunlukları, paketler arası geçen süreler ve paket erişim sıralarını kullanarak %91 ağ akış keskinliği yakaladıkları Basit İstatistiksel Protokol Parmakizi metodu olmuştur. Çalışmalarında değişkenlik gösteren RTT, MTU değerlerinin neden oldukları gürültüleri azaltmak için protokol maskesi ve normalizasyon teknikleri kullanmışlardır.
Vektör uzay modelini kullanarak IP trafik sınıflandırması ise Chung ve diğerleri [13] tarafından yapılmıştır. Çalışmalarında özellik olarak paketlerin yüklerinde geçen kelimelerin (16 bitlik sayı) sıklıklarından oluşan yük vektörlerini kullanmışlardır. Paketlerin yüklerini incelediklerinden çalışmaları şifreli trafiğin tespitini sağlayamamaktadır.
Bu çalışmada ise ağ akışlarının ilk bir kaç paketinin uzunlukları ve normalize edilmiş standart sapmaları vektör uzay modelinde özellik olarak kullanılmıştır. İlk bir kaç paketin uygulamaların müzakere fazını oluşturması ve paketlerin yüklerine bakılmaması şifreli trafiğin gerçek zamanda tespit edilmesini sağlayan etkenler olduğu düşünülmektedir.

3. Metodoloji

Bu bölüm IP trafik sınıflandırma yöntemimizi açıklamaktadır. Vektör uzay modeli, uygulamaların paket uzunlukları ile standart sapma vektörlerinin çıkarılması ve bu vektörleri kullanarak gerçeklenen kosinüs benzerliği bu bölümün alt başlıklarında anlatılmaktadır.

3.1. Vektör uzay modeli

Vektör uzay modeli doküman sınıflandırmada kullanılan ve metin dokümanlarını cebirsel olarak temsil eden vektörlerden oluşan bir modeldir. Bu modellemede her bir doküman, barındırdığı metinlerin sıklıklarına göre bir vektöre sahip olur. Dokümanlar sahip oldukları vektörler ile yapılan benzerlik hesapları ile sınıflandırılırlar. Bu modelleme IP trafik sınıflandırmasına da uygulanabilir. Bunun için ağ üzerinden geçen aynı uygulamaya ait akışlar bir doküman gibi ele alınıp ona ait bir vektör oluşturulur. Dokümanları temsil eden metinlerin sıklıkları yerine ise akışların paket uzunlukları ve bu paket uzunluklarının normalize edilmiş standart sapmaları kullanılabilir.

3.2. Ağırlıklı Kosinüs Benzerliği

IP trafik sınıflandırmasında yeni bir ağ akışının protokolü ağırlıklı kosinüs benzerliği ile tespit edilmektedir.
Kosinüs benzerliği iki vektör arasındaki açının kosinüsünün hesaplanmasına dayanır. Vektörler arası açı ne kadar az ise kosinüs değeri de 1' e o kadar yakındır. Dolayısıyla iki vektör birbirine ne kadar benzer ise kosinüs değeri de 1'e o kadar yakındır. Kosinüs benzerliğinin matematiksel tanımlaması aşağıdaki gibidir.
v, *w* iki vektör, *a* bu vektörlere ait ağırklıklandırma vektörü ve *n* vektörlerin eleman sayısı olsun. Bu durumda *a* ağırlık vektörünü kullanarak *v* ve *w* vektörleri arasındaki ağırlıklı kosinüs benzerliği hesabı 1. denklemdeki gibidir.

$$AKB(v,w) = \frac{\sum_{i=1}^{n} a_i v_i w_i}{\sqrt{\sum_{i=1}^{n} a_i (v_i)^2 * \sum_{i=1}^{n} a_i (w_i)^2}} \quad (1)$$

3.3. Çevrimdışı Öğrenme

IP trafiğinden geçen aynı uygulamaya ait akışların vektör uzay modeli kullanılarak karakteristiklerinin çıkartılması makine öğrenmesi tabanlı yaklaşımımızın öğrenme kısmını oluşturmaktadır. Öğrenme aşaması çevrimdışı olarak önceden toplanmış öğrenme veri kümesi ile yapılmaktadır. Bu işlem aşağıdaki aşamalardan oluşur:
1. Veri kümesi üzerindeki her bir ağ akışı port ve yük bazlı sınıflandırıcı ile ya da manuel olarak incelenir ve hangi uygulamaya ait olduğu tespit edilir. Her bir akışın gelen ve giden yönde ilk bir kaç paketinin paket uzunlukları kayıt altına alınır. Bu adımın sonunda her bir uygulama için o uygulamaya ait örneklerin gelen ve giden yönde ilk bir kaç paketine ait paket uzunlukları elde edilmiş olunur.
2. Daha sonra ise her bir uygulama için sahip olduğu örneklerin gelen ve giden yönde ilk bir kaç paketine ait paket uzunluklarının ortalaması alınarak gelen ve giden paket

uzunlukları vektörü, paket uzunluklarının standart sapması hesaplanarak da gelen ve giden standart sapma vektörü hesaplanır. Standart sapma vektörü de normalizasyondan geçirilerek normalize edilmiş standart sapma vektörü elde edilir.
Öğrenme aşamasının akış şeması Şekil 1'de verilmiştir.
Öğrenme aşamasında her bir uygulama için çıkartılan paket uzunlukları ve normize edilmiş standart sapma vektörlerinin matematiksel tanımlamaları aşağıdaki gibidir.
A uygulamasına ait örneklerin *S* kümesinde olduğunu varsayalım. *k* adet örnek içeren *S* kümesindeki herbir örnek ise *n* uzunluğundaki v_i vektörü olsun. Bu durumda *A* uygulamasının paket uzunlukları vektörü 2. denklem ile, standart sapma vektörü 3. denklem ile, normalize edilmiş standart sapma vektörü ise 4. denklem ile hesaplanır.

$$PktUzVek(A) =< a_1, a_2, ..., a_n >$$

$$a_i = \frac{\sum_{i=1}^{k} v_i}{k} \quad (2)$$

$$StdSapVek\ (A) =< s_1, s_2, \ldots, s_n >$$

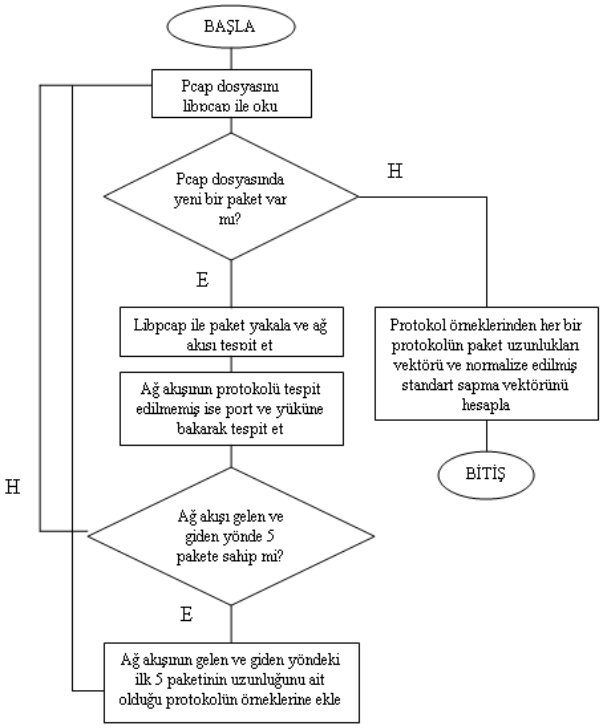
$$s_i = \sqrt{\frac{\sum_{i=1}^{k} (v_i - a_i)^2}{k}} \quad (3)$$

$$NormStdSapVek(A) =< ns_1, ns_2, \ldots, ns_n >$$

$$ns_i = \frac{1}{1 + s_i} \quad (4)$$

3.4. Çevrimiçi Sınıflandırma

Her bir uygulamaya ait karakteristikler çıkartıldıktan sonra yeni bir ağ akışının hangi uygulamaya ait olduğunu tespit etme makine öğrenmesi tabanlı yaklaşımımızın sınıflandırma aşamasını oluşturmaktadır. Bu aşama gerçek zamanda çevrimiçi olarak yapılmaktadır ve aşağıdaki adımlar ile gerçeklenmektedir:
1. Ağ üzerinde yeni bir akış tespit edilir.
2. Bu akışa ait ilk bir kaç paketin paket uzunlukları ile öğrenme aşamasında çıkartılan diğer uygulamaların paket uzunlukları vektörü ve normalize edilmiş standart sapma vektörü arasında ağırlıklı kosinüs benzerliği hesabı yapılır. Normalize edilmiş standart sapma vektörü kosinüs benzerliği hesaplamasında ağırlık olarak kullanılarak, öğrenme paket uzunluğu değişkenlik gösteren vektör elemanlarının benzerlik hesabında daha az etki yapması hedeflenmiştir.
3. Yapılan hesaplama sonucu kosinüs benzerliği en yüksek olan uygulama tespit edilen ağ akışına atanır. Böylece sınıflandırma aşaması tamamlanmış olur.
Algoritma Şekil 2'de verilmiştir.



Şekil 1: Öğrenme akış şeması

```
Yeni bir ağ akışı (X akışı) tespit et

Eğer ağ akışı gelen ve giden yönde 5 pakete ulaşmış ise
    Ağ akışının gelen ve giden paket uzunlukları vektörünü oluştur
    for karşılaştırma yapılacak her bir protokol (Proto(i)) için do
        Sonuç = Kosinus_Benzerliği (X, Proto(i))
        Eğer Sonuç > Maksimum ve Sonuç > Eşik Değeri ise
            Maksimum = Sonuç
            Proto X = Proto(i)

Kosinus_Benzerliği(X, Proto)
    Sonuç = 0, kare_X = 0, kare_Proto = 0
    for gelen ve giden paket uzunlukları vektörlerinin her bir elemanı (el(j)) için do
        sonuç += X_PktUz_el(j) * Proto_PktUz_el(j) * Proto_NStd_el(j)
        kare_X += X_PktUz_el(j) * X_PktUz_el(j) * Proto_NStd_el(j)
        kare_Proto += Proto_PktUz_el(j) * Proto_PktUz_el(j) * Proto_NStd_el(j)
    sonuç = sonuç / (kök(kare_X) * kök(kare_Y))
    return sonuç
```

Şekil 2: Trafik sınıflandırma algoritması

4. Deneysel Sonuçlar

4.1. Test Ortamı

Çalışma 64 bit FreeBSD 8.2 sistemi üzerinde test edilmiştir. C programlama dili ile geliştirilen ve libpcap paket toplama kütüphanesini kullanan test uygulaması hem ağ kartı üzerinden hem de pcap dosyaları aracılığı ile paket toplayabilmekte, öğrenme ve sınıflandırma işlemlerini yapabilmektedir.
Test için Naval Bilgisayar Bilimleri Bölümünün DEEP (Digital Evaluation and Exploitation) çalışma grubunun 2009

Kasım ve Aralık aylarında kayıt ettiği paket izleri [14] veri kümesi olarak kullanılmıştır. Bu paket izleri üzerinde port ve yük tabanlı olarak yapılan incelemede HTTP, HTTPS, SMTP ve FTP protokollerine ait 6485 ağ akışı tespit edilmiştir.

4.2. Performans Metrikleri

Trafik sınıflandırma yaklaşımlarının keskinliklerini karşılaştırmada kullanılan temel performans metrikleri şunlardır:

- *Yanlış Negatif* (YN): Üzerinde çalışılan uygulamaya ait olup yanlış tespit edilen ağ akışlarının yüzdesi.
- *Doğru Pozitif* (DP): Üzerinde çalışılan uygulamaya ait olup doğru tespit edilen ağ akışlarının yüzdesi (100 - YN).
- *Yanlış Pozitif* (YP): Üzerinde çalışılan uygulamaya ait olmayıp üzerinde çalışılan uygulama olarak tespit edilen ağ akışlarının yüzdesi.
- *Doğru Negatif* (DN): Üzerinde çalışılan uygulamaya ait olmayıp üzerinde çalışılan uygulama olarak tespit edilmeyen ağ akışlarının yüzdesi.

Bu çalışmada her bir uygulama için YN, DP, YP ve DN yüzdeleri hesaplanarak performans değerlendirmesi yapılmıştır.

4.3. Test Sonuçları ve Değerlendirme

Veri kümesi kullanılarak gerçeklenen öğrenme aşamasında hesaplanan HTTP, HTTPS, SMTP ve FTP protokollerinin karakteristiklerini oluşturan paket uzunlukları ve normalize edilmiş standart sapma vektörleri Tablo 1, Tablo 2, Tablo 3 ve Tablo 4' de verilmiştir.

Tablo 1: HTTP protokolü vektörleri

Vektör	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Giden Paket Uzunlukları Vektörü	563.34	553.54	556.39	513.86	544.56
Giden Normalize Edilmiş Standart Sapma Vektörü	0.0023	0.0022	0.0022	0.0024	0.0022
Gelen Paket Uzunlukları Vektörü	740.98	621.57	759.85	613.45	724.14
Gelen Normalize Edilmiş Standart Sapma Vektörü	0.0027	0.0029	0.0025	0.0028	0.0025

Tablo 2: HTTPS protokolü vektörleri

Vektör	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Giden Paket Uzunlukları Vektörü	1319.0	1276.5	1117.6	367.37	89.98
Giden Normalize Edilmiş Standart Sapma Vektörü	0.0041	0.0029	0.0021	0.006	0.0059
Gelen Paket Uzunlukları Vektörü	319.94	180.48	600.27	1128.2	991.56
Gelen Normalize Edilmiş Standart Sapma Vektörü	0.002	0.032	0.0023	0.0024	0.0018

Tablo 3: SMTP protokolü vektörleri

Vektör	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Giden Paket Uzunlukları Vektörü	25	162.69	27.36	674.67	46.35
Giden Normalize Edilmiş Standart Sapma Vektörü	1	0.0435	0.1408	0.0029	0.0835
Gelen Paket Uzunlukları Vektörü	14.75	14.22	90.11	251.34	124.39
Gelen Normalize Edilmiş Standart Sapma Vektörü	0.2224	0.0891	0.0291	0.008	0.0041

Tablo 4: FTP protokolü vektörleri

Vektör	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Giden Paket Uzunlukları Vektörü	64.51	64.77	51.77	27.62	26.85
Giden Normalize Edilmiş Standart Sapma Vektörü	0.0215	0.0972	0.0282	0.084	0.078
Gelen Paket Uzunlukları Vektörü	15.98	16.91	20.36	17.76	12.66
Gelen Normalize Edilmiş Standart Sapma Vektörü	0.7445	0.5202	0.0533	0.0411	0.0799

[15] de ilk 4 veya 5 paketin uygulamaların karakteristiklerinin ayırt edilmesinde daha belirleyici olduğunun tespit edilmesi üzerine testlerde ilk 5 pakete ait bilgiler kullanılmıştır.
Normalize edilmiş standart sapma vektörünün sınıflandırma performansına etkisini görmek amacıyla iki ayrı kosinüs benzerliği hesabı yapılmıştır. Bunlardan biri sadece paket uzunlukları vektörünü kullanmakta, diğeri ise normalize edilmiş standart sapma vektörünü benzerlik hesabında ağırlık olarak kullanmaktadır. Bu iki benzerlik hesabına ait sonuçlar Tablo 5 ve Tablo 6'da verilmiştir.

Tablo 5: Sadece paket uzunlukları vektörü ile kosinüs benzerliği sonuçları

	YN	DP	YP	DN
HTTP	%20.99	%79.01	%1.01	%98.99
HTTPS	%29.44	%70.56	%1.04	%98.96
SMTP	%21.22	%78.78	%0.083	%99.92
FTP	%14.24	%85.76	%8.5	%91.5

Tablo 6: Normalize edilmiş standart sapma ve paket uzunlukları vektörü ile kosinüs benzerliği sonuçları

	YN	DP	YP	DN
HTTP	%13.72	%86.28	%0.27	%99.73
HTTPS	%10.55	%89.45	%7.71	%92.29
SMTP	%5.00	%95.00	%0.00	%100.0
FTP	%8.84	%91.16	%9.75	%90.25

Benzerlik sonuçlarından normalize edilmiş standart sapma vektörünün DP değerini ortalama %12 arttırdığı görülmektedir. Ayrıca şifreli yüke sahip olan HTTPS protokolünde %19 gibi yüksek bir keskinlik artışı görülmesi ile yaklaşımımızın şifreli internet trafiğini de başarılı bir şekilde tespit edeceği düşünülmektedir.
SMTP protokolünün YP oranının sıfır olması karakteristiğinin diğer protokollerin karakteristiğine uzak olmasından kaynaklanmaktadır. Ayrıca HTTP ve HTTPS protokollerinin yanlış tespit edilen ağ akışlarının genelde FTP olarak tespit edildiği görülmektedir. Bu da HTTP ve HTTPS ile FTP ağ akışları arasında ortak bir karakteristikte ağ akışlarının bulunduğunu göstermektedir. Bunun sebebi ise HTTP ve HTTPS protokollerinin normalleştirilmiş standart sapma vektörlerindeki değerlerin düşük olmasıdır. Çünkü normalleştirilmiş standart sapma vektörlerindeki değerlerin düşük olması standart sapma değerlerinin yüksek olduğunu göstermektedir. Standart sapma değerleri ne kadar yüksek ise ağ akışlarının birbirlerine benzerlikleri de o kadar uzaktır.

5. Gelecek Çalışmalar

Çalışma aşamasındaki bazı kısıtlar gelecek çalışmalar için yol gösterici olmaktadır.

Bu kısıtlar öğrenme aşamasında çok önemli payı olan port ve yük bazlı sınıflandırıcımızın tespit edebildiği uygulama veya protokollerin sayısının, test verisinde bulunan ağ akışlarının ve protokollerinin sayısının az olmasıdır.

Daha fazla uygulama ve protokol desteğine sahip port ve yük bazlı sınıflandırıcı ve her protokol için bolca ağ akışına sahip test verisi yaklaşımımızın daha geniş çapta değerlendirilmesini sağlayacaktır.

Ayrıca şifreli protokoller için manuel olarak hazırlanacak test verileri yaklaşımımızın şifreli trafiğin tespit edilmesindeki başarısını belirlemede daha etkin bir rol oynayacaktır.

6. Sonuç

Bu çalışmanın sonuçları göstermektedir ki internet trafik sınıflandırmasında ağ akışlarının gelen ve giden yönde ilk bir kaç paketinin paket boyutları vektörleri ve bu vektörler üzerinde yapılan kosinüs benzerliği hesabı yüksek bir başarım sağlamaktadır. Yani ağ akışlarının paket boyutları vektörü protokollerin tespitinde ayırt edici özellik göstermektedir.

Protokollerin normalize edilmiş standart sapma vektörlerinin kosinüs benzerliği hesabında ağırlık olarak kullanılması ise trafik sınıflandırma performansını arttırmıştır.

Ayrıca yaklaşımımızın paket yüklerine bakmadan şifreli internet trafiğini yüksek bir başarım ile tespit edebileceği HTTPS trafiğinin tespitinde gösterdiği yüksek başarım yüzdesinden öngörülmektedir.

7. Kaynakça

[1] IANA, Internet Assigned Numbers Authority, http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml

[2] A. Moore ve K. Papagiannaki, “Toward the accurate identification of network applications”, *Proc. Passive and Active Measurement Workshop (PAM2005)*, Boston, MA, USA, Mart/Nisan 2005

[3] P.-C. Lin, Y.-D. Lin, T.-H. Lee, ve Y.-C. Lai. “Using string matching for deep packet inspection”, *IEEE Computer Practices*, 41(4):23–28, Nisan 2008.

[4] S. Sen, O. Spatscheck, ve D. Wang, “Accurate, scalable in network identification of P2P traffic using application signatures”, *WWW2004*, New York, NY, USA, Mayıs 2004.

[5] A. McGregor, M. Hall, P. Lorier, ve J. Brunskill, “Flow clustering using machine learning techniques”, *Proc. Passive and Active Measurement Workshop (PAM2004)*, Antibes Juan-les-Pins, Fransa, Nisan 2004

[6] S. Zander, T. Nguyen, ve G. Armitage, “Automated traffic classification and application identification using machine learning”, *IEEE 30th Conference on Local Computer Networks (LCN 2005)*, Avustralya, Kasım 2005.

[7] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, ve K. Salamatian, “Traffic classification on the fly”, *ACM Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review*, vol. 36, no. 2, 2006.

[8] J. Erman, A. Mahanti, M. Arlitt, ve C. Williamson, “Identifying and discriminating between web and peer-to-peer traffic in the network core”, *WWW ’07: Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM Press, Mayıs 2007, sayfa 883–892.

[9] M. Roughan, S. Sen, O. Spatscheck, ve N. Duffield, “Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification”, *Proceedings of ACM/SIGCOMM Internet Measurement Conference (IMC) 2004*, Taormina, Sicily, İtalya, Ekim 2004.

[10] A. Moore ve D. Zuev, “Internet traffic classification using Bayesian analysis techniques”, *ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005*, Banff, Alberta, Kanada, Haziran 2005.

[11] T. Auld, A. W. Moore, ve S. F. Gull, “Bayesian neural networks for Internet traffic classification”, *IEEE Transactions on Neural Networks*, no. 1, sayfa. 223–239, Ocak 2007.

[12] M. Crotti, M. Dusi, F. Gringoli, ve L. Salgarelli, “Traffic classification through simple statistical fingerprinting,” *SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 1, sayfa 5–16, 2007.

[13] J.Y. Chung, B. Park, Y.J. Won, J. Strassner ve J.W. Hong, "Traffic Classification Based on Flow Similarity", *IPOM 2009*, LNCS 5843, sayfa 65-77, 2009

[14] DEEP (Digital Evaluation and Exploitation) Paket İzleri, https://domex.nps.edu/corp/scenarios/2009-m57/net/, Department of Computer Science, Naval Postgraduate School, Monterey, CA, 2009

[15] J. Erman, M. Arlitt, A. Mahanti, “Traffic Classification Using Clustering Algorithms”, *SIGCOMM06 Workshops*, Pisa, Italya, 2006