

T.C.
GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ
MÜHENDİSLİK VE FEN BİLİMLERİ ENSTİTÜSÜ

KULLANICI DAVRANIŞ ANALİZİ İLE
NÜFUZ TESPİTİ

Rahim KARABAĞ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

GEBZE
2006

T.C.
GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ
MÜHENDİSLİK VE FEN BİLİMLERİ ENSTİTÜSÜ

KULLANICI DAVRANIŞ ANALİZİ İLE
NÜFUZ TESPİTİ

Rahim KARABAĞ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TEZ DANIŞMANI
Doç. Dr. İbrahim SOĞUKPINAR

GEBZE
2006

ÖZET

TEZ BAŞLIĞI: Kullanıcı Davranış Analizi ile Nüfuz Tespiti

YAZAR ADI: Rahim KARABAĞ

Bilgisayar sistemlerine yönelik saldırılar her geçen gün artmakta ve bu saldırılar sonucunda oluşan zararlar hızla büyümektedir. Birçok kurum günümüzde, bilgi güvenliğinin önemini kavramış ve bu sebeple değişik koruma yöntemlerine başvurmuştur. Güvenlik duvarları, anti-virüs yazılımları, açıklık tarayıcılar ve nüfuz tespit sistemleri gibi teknolojiler bilgi güvenliğini sağlamak amacıyla kullanılan çözümlerin bazılarıdır.

Bu çalışmada kullanıcı davranışlarının analizi ile anormallik tabanlı nüfuz tespit sistemi tasarımı yapılmıştır. Önerilen yöntemde, ağ üzerinden toplanan veriler veri madenciliği teknikleri kullanılarak ağ kullanıcılarına ait istatistiksel bilgiler elde edilir. Elde edilen verilerden kullanıcı davranışları çıkarılır. Kullanıcı davranışlarının sınıflandırılmasında KNN sınıflandırması kullanılır. Bu sonuçlardan kötü niyetli kullanıcılar belirlenerek nüfuz tespiti yapılır. Daha sonra bir alarm üretilir ve bu kullanıcı sistem yöneticisine bildirilir. Sonuçta, ağ içerisindeki kullanıcı kontrol altına alınarak nüfuz işlemlerinin ve aynı zamanda bilinçsizce yapılan ağ kullanım hatalarının önüne geçilir.

SUMMARY

THESIS TITLE: Intrusion Detection Using User Behavior Analysis

AUTHOR: Rahim KARABAĞ

While attacks on computer systems have been increased, loss resulted from these attacks have been grown rapidly. Most of the companies today understand the importance of information security and they have started to use various methods for protection. Technologies like firewalls, anti-virus software, vulnerability scanners and intrusion detection systems are some of the solutions used which are to provide information security.

In this work, an anomaly-based intrusion detection system is designed by using user behavior analysis. In the proposed method, statistical information about the users on the network is gathered from the data collected from the network by using data mining techniques. User behaviors are constituted from this statistical information. kNN classification method is used for clustering the user behaviors. Intrusion detection is performed by using anomaly based analysis on these clusters. If an intruder is detected, an alarm is created and system administrator is informed about this intrusion. As a result, users on the network can be controlled by using the proposed method. So the system can be prevented from intrusions and unwanted network usage errors.

TEŞEKKÜR

Tezime başlamamdan bitirmeme kadar bana devamlı yol gösteren ve destek veren danışmanım Sayın Doç. Dr. İbrahim Soğukpınar' a; yardımlarından dolayı arkadaşlarım Araş. Gör. Türker Akyüz, Dr. Hidayet Takçı ve Araş. Gör. Mustafa Çoşkun' a sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER DİZİNİ

	<u>Sayfa</u>
ÖZET	iv
SUMMARY	v
TEŞEKKÜR	vi
İÇİNDEKİLER DİZİNİ	vii
SİMGELER VE KISALTMALAR DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
ÇİZELGELER DİZİNİ	xi
1. GİRİŞ	1
1.1 Bilgi Güvenliği ve Saldırılar	2
1.2 Nüfuz Tespit Sistemleri – NTS	5
1.3 Veri Madenciliği	5
1.4 Bu Çalışmada Yapılanlar	5
2. GENEL KAVRAMLAR VE KONUYLA İLGİLİ ÇALIŞMALAR	7
2.1 Nüfuz Tespit Sistemleri (NTS)	7
2.1.1 Nüfuz Tespit Sistemlerinin Sınıflandırılması	7
2.1.2 Anormallik Tabanlı Nüfuz Tespit Sistemleri	10
2.1.3 Kötüye Kullanım Tabanlı Nüfuz Tespiti:	12
2.1.4 Nüfuz Tespit Sistemlerinin İncelenmesi	12
2.1.5 Nüfuz Tespit Sistemlerinin Değerlendirilmesi	15
2.1.6 Alınacak Önlemler	17
2.1.7 Testlerde Kullanılan Veri Setleri	18
2.2 Veri Madenciliği	22
2.2.1 Veri Madenciliği Nedir ve Nasıl Yapılır?	22
2.2.2 Veri Madenciliği Teknikleri:	23
2.2.3 Veri Madenciliği İşlevleri	24
2.3 Veri Madenciliği ve Nüfuz Tespit Sistemi	25
2.4 Kümeleme	26
2.5 En Yakın Komşu Kuralı (kNN)	29
3. ÖNERİLEN MODEL	33
3.1 Modelin Yapısı	33
3.2 Kullanıcı Tanımları	35
3.3 Benzerlik	37

	viii
3.4 K-Yakın Komşu (kNN) ile Kümeleme:	37
3.5 Tespit İşlemi	39
4. DENEYSEL ÇALIŞMALAR VE ANALİZ	41
4.1 Ağ Trafikinden Verilerin Elde Edilmesi	41
4.2 Kullanıcılar	45
4.3 Kümeleme	46
4.4 Analiz	47
4.5 Karşılaştırma	50
5. SONUÇ VE ÖNERİLER	52
KAYNAKLAR	53
ÖZGEÇMİŞ	56

SİMGELER VE KISALTMALAR DİZİNİ

KDA-NT	Kullanıcı Davranış Analizi İle Nüfuz Tespiti
NTS	Nüfuz Tespit Sistemi
IDS	Intrusion Detection System
IP	Internet Protocol
DoS	Denial of Service
R2T	Remote to Local
U2R	User to Root
CERT	Computer Emergency Readiness Team
CERT/CC	Computer Emergency Rescue Team / Coordination Center
IDES	Intrusion Detection Expert System
NIDES	Next-Generation Intrusion Detection Expert System
W&S	Wisdom and Sense
NSM	Network Security Monitor
CERIAS	Center for Education and Research in Information Assurance and Security
EMERALD	Event Monitoring Enabling Responses to Anomalous Live Disturbances
DIDS	Distributed Intrusion Detection System
ROC	Receiver Operating Characteristic
MIB	Management Information Base
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
NIS	Network Information Service
NFS	Network File System
LDAP	Lightweight Directory Access Protocol
TCP/IP	Transmission Control Protocol / Internet Protocol
UDP	User Datagram Protocol
RPC	Remote Procedure Call
CDP	Cisco Discovery Protocol
ISP	Internet Service Provider

ŞEKİLLER DİZİNİ

<u>Sekil</u>	<u>Sayfa</u>
1.1. Mayıs-Temmuz 2002 arasında 10 porta yapılan saldırı yoğunluğu	3
1.2 CERT'e raporlanan saldırı sayısı	4
1.3 1980-2000 arası saldırı tiplerindeki ve saldırgan örüntülerindeki değişiklikler	4
2.1 Ağ tabanlı NTS örneği	8
2.2 Normal ve saldırı içeren dağılımın teorik ayrışımı	10
2.3 Örnek ROC eğrileri	16
2.4 DARPA 99 Ağ yapısı	19
2.5 Y. Liao tarafından geliştirilen kNN algoritması	27
2.6 En Yakın Komşu Kuralı uygulaması	30
2.7 K Yakın Komşu Kuralı (k=3)	30
2.8 Voronoi Çizeneği (1)	31
2.9 Voronoi Çizeneği (2)	31
2.10 Voronoi Çizeneği ile En Yakın Komşu İlişkisi	31
3.1 KDA ve NTS ağ yapısı	33
3.2 Kullanıcı davranış analizi ile Nüfuz tespit sistemi modeli	34
3.3 TCPdump paket yapısı ethereal incelemesi	36
3.4 Kullanıcı kümelerinin bulunması ve saldırganın tespiti	39
4.1 Ethereal'da DARPA TCPdump veri analizi	42
4.2 172.16.112.100 kullanıcıasına ait gelen ve giden paketlerin zaman serisi 1s	43
4.3 172.16.112.100 kullanıcıasına ait gelen ve giden paketlerin zaman serisi 10s	44
4.4 172.16.112.100 kullanıcıasına ait protokol istatistikleri	44
4.5 172.16.112.100 kullanıcıasına ait TCP paket uzunlukları istatistikleri	45
4.6 KDA-NTS ağ izleme yönetimi	49

ÇİZELGELER DİZİNİ

<u>Cizelgeler</u>	<u>Sayfa</u>
2.1. Nüfuz Tespit Yöntemlerinin ve Araçlarının Sınıflandırılması	13
3.1: kNN ile yapılan nüfuz tespit sistemleri	35
3.2 Kullanıcı tabloları	37
4.1 K=3 için belirlenen kullanıcı kümeleri	46
4.2 K=5 için belirlenen kullanıcı kümeleri	46
4.3 K=10 için belirlenen kullanıcı kümeleri	46
4.4 K=5 anormallik analiz sonuçları	48

1. GİRİŞ

Bilgisayar sistemlerine yönelik saldırıların artması ile birlikte bilgi güvenliği kavramı, her alanda bir hayli önem kazanmıştır. Bilgi güvenliği; verilerin, ağı ve yönetilebilir ağ cihazlarının içeriden veya dışarıdan gelebilecek saldırılara karşı korunmaya çalışılmasıdır.

Saldırı, bilgisayar sisteminin yetkisiz erişimlerle sisteme yapılan müdahaleler veya kaynakların yanlış kullanılmaya çalışılmasıdır. Günümüzde sadece sistemi bozmaya yönelik saldırılardan çok bilgi hırsızlığı için yapılan saldırılar daha çok ön plandadır. Örneğin bir kişinin banka bilgilerinin elde edilmesi en güncel saldırılardan biridir. Saldırıları genellikle; sistemimizde hatalı ve eksik yetkilendirmeler yapıldığında, kullanıcıların tarafından tahmin edilebilir şifreler kullanıldığında, sistem yanlış yapılandırıldığında veya sisteme dahil edilen yazılımların oluşturduğu açıklıklar bulunduğunda meydana gelir. Başarılı saldırı girişimleri nüfuz olarak tanımlanır. Saldırıların önlenmesinde kullanılan nüfuz tespit sistemleri güvenlik duvarının arkasında, ağ içerisinde çalışırlar.

Son zamanlarda veri madenciliği tabanlı nüfuz tespiti NTS'leri için sıkça kullanılan yöntemlerdendir Daha çok sınıflandırma ve kümeleme gibi tekniklerle tespit işlemini yerine getirir. Nüfuz tespiti için kümeleme metodu kullanılmak istendiği de; ağ trafiği izlenerek kullanıcıların sebep olduğu trafik belli kümelerde gruplanır ve bu yöntemle nüfuz tespiti yapılmaya çalışılır.

Kümeleme sonrasında elde edilen farklı kümeler de normal erişim yapan kullanıcılar yer alırken; Kümesinden farklı davranış gösteren veya sisteme dahil olmayan kullanıcılar da saldırı yapan olarak bulunacaktır. Kümeleme sonrasında kullanıcıların yer aldıkları kümelerden kullanıcıların saldırgan olup olmadıkları anlaşılabilir. Ağ trafiği incelenirken analizler kullanıcı tabanlı yapılır. Anormallik tabanlı bu yaklaşımda bir kullanıcının hangi kümede yer aldığını tespit için k-NN yöntemi kullanılmıştır.

Ağdan toplanan kullanıcı bilgileri kümelendikten sonra kullanıcıların oluşturduğu trafik uygun bir kümeye dâhil edilir. Buradan kullanıcı bilgileri kontrol

edilirken kümelere benzerliklerine bakılır. Bu modelde veri toplama için TCPdump ile veri toplaması yapmış DARPA veri setleri kullanılır, toplanan verileri değerlendirmek için veri madenciliği teknikleri ve kullanıcı davranışlarını çıkarmak için k-NN kullanılır.

1.1 Bilgi Güvenliği ve Saldırılar

Bilgi güvenliği, bilgisayar sistemlerinin ve bu sistemleri oluşturan verilerin, ağın ve makinelerin dışarıdan veya içeriden gelebilecek her türlü saldırılara karşı korunmaya çalışılması anlamına gelir. Nüfuz, sisteme girmeye yetkisi olmayan kişilerin sistemi kırmaya veya kaynaklarını yanlış kullanmaya çalışmasıdır. Nüfuzlar genellikle; sistemin yetkilendirme için şifre koruması kullanmadığı veya zayıf şifreler kullandığı zamanlarda, sistemin doğru yapılandırılmadığı zamanlarda veya sistemin bazı yazılım komutlarının uygulanması ile ortaya çıkan açıklıklar içeren yazılım paketlerini çalıştırdığı zamanlarda meydana gelir.

Saldırı işlemlerini incelersek dört temel kategoride toplanır. 1- Bilgi tarama, 2- hizmet engelleme, 3- Yönetici hesabı ile yerel oturum açma, 4- kullanıcı hesabının yönetici hesabına yükseltilmesi.

1-Bilgi Tarama (Probe/Scan): Bir sunucu yada herhangi bir makinenin geçerli IP adresini, aktif portlarını veya işletim sistemini öğrenmek için yapılan saldırılardır. Örneğin IPSweep, PortswEEP gibi.

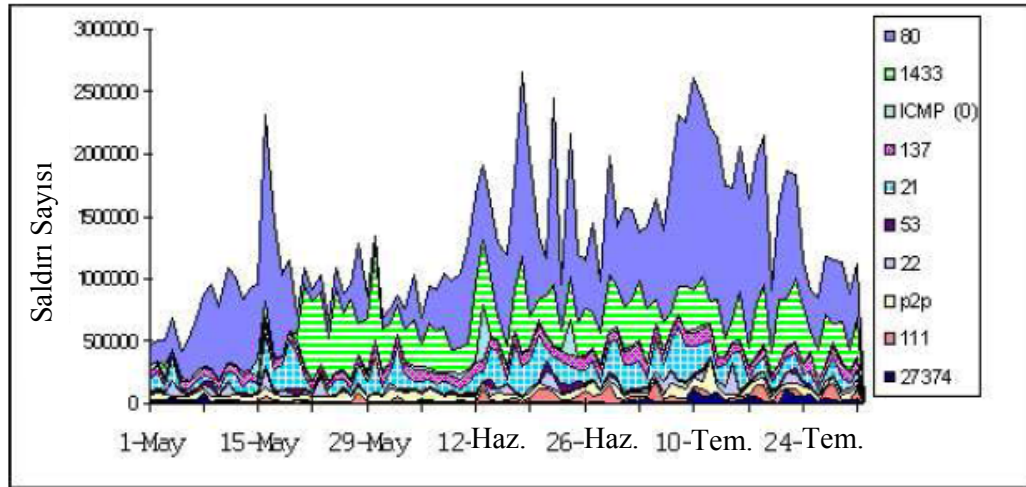
2- Hizmet engelleme(Denial of Service- DoS): Genelde TCP/IP protokol yapısındaki açıklıklardan faydalanır veya bir sunucuya çok sayıda istek yönelterek onu tıkamaya sebep olan saldırılardır. Bunlara örnek verecek olursak Ping-of-Death ölümüne ping, TCP SYN paketi, Sürekli bağlantı istekleri, birçok makine aynı anda hizmet istemesi ile oluşan DoS, Smurf ICMP mesajlarının broadcast ile tüm ağa dağıtılmasıdır, Selfping, TCPreset, MailBomb gibi.

3- Yönetici hesabı ile yerle oturum açma (Remote to Local – R2L): Kullanıcı haklarına sahip olmadığı halde misafir ya da başka kullanıcı olarak izinsiz erişim yapılması. Örneğin SSHTrojan, Guest gibi.

4- Kullanıcı hesabının yönetici hesabına yükseltilmesi (User to Root – U2R):

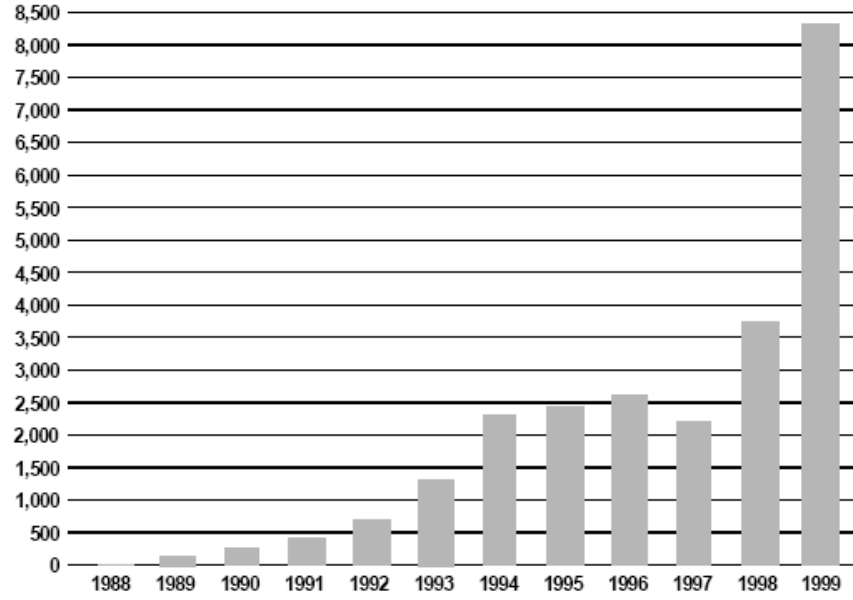
Sisteme girme izni olan fakat yönetici olmayan bir kullanıcının yönetici izni gerektiren işler yapmaya çalışmasıdır. Örneğin Eject, SQLattack gibi.

2003 yılında Yegneswaran tarafından yapılan çalışmada internet üzerindeki 1600 ağdan toplanmış verilerin analizi yapılmıştır. İnternet üzerinde her gün 25000 civarında saldırı gerçekleştiği saptanmıştır. [Yegneswaran et al. 2003] Bu çalışmadan alınan şekil1.1'deki grafikte sadece yoklama (probe/scan) saldırısı göz önüne alındığında Mayıs-Temmuz ayları arasındaki port numaralarına karşı düşen saldırı yoğunluğu verilmiştir. Şekil1.1'de yapılan saldırı yoğunluğunun ne kadar ciddi boyutlarda olduğunu gösterilmektedir.



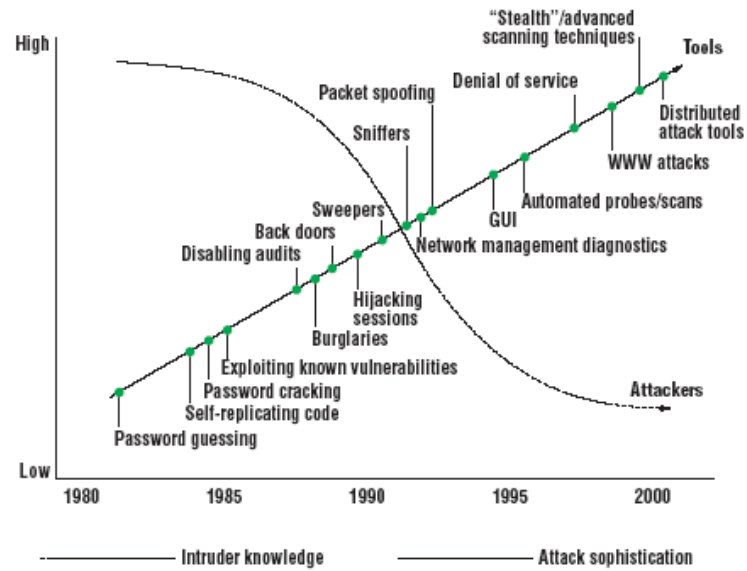
Şekil 1.1. Mayıs-Temmuz 2002 arasında 10 porta yapılan saldırı yoğunluğu
[Yegneswaran et al. 2003].

1980'lerde sadece bilgisayar sistemleri uzmanları tarafından yapılan saldırılar artık otomatikleşmiş saldırı araçları ile çok daha hızlı sürede yapılmaktadır. CERT (Computer Emergency Readiness Team) tarafından 2000 yılında yayınlanan bir rapordan alınan değerler Şekil 1.2' de gösterilmiştir. Saldırı sayılarındaki artışın boyutları daha o senelerden çok ciddi boyutlara ulaşmış ve giderek elektronik ticaret uygulamalarının da artmasıyla fazlaca maddi kayba yol açmaya başlamıştır[Allen et al. 2000].



Şekil 1.2 CERT'e raporlanan saldırı sayısı [Allen et al. 2000].

Şekil 1.3'de ise saldırgan davranışlarındaki değişim ve saldırı tiplerinin yıldan yıla değişimi verilmiştir [McHugh et al.,2000].



Şekil 1.3 1980-2000 arası saldırı tiplerindeki ve saldırgan örüntülerindeki değişiklikler [McHugh,2000].

1.2 Nüfuz Tespit Sistemleri – NTS

Nüfuz tespit sistemleri yukarıda da bahsedilen saldırıları engellemek için geliştirilen güvenlik yazılımlarıdır. Ticari, devlet tarafından geliştirilen ve araştırmacılar tarafından geliştirilen birçok NTS mevcuttur. NTS'ler değişik yaklaşımlar ile sınıflandırılabilirler. Bunlar sistemin veri işleme zamanına göre, çalışma alanına göre ve nüfuz tespit yöntemlerine göre sınıflandırılırlar. NTS İkinci bölümde detaylı olarak anlatılacaktır.

1.3 Veri Madenciliği

Veri madenciliği; büyük miktardaki veri içerisinde anlamlı bilginin çıkarıldığı tekniğe denir. Veri madenciliği ile nüfuz tespiti birbirine yakın konulardır çünkü veri madenciliğinde yapılan işlerin bir kısmı anormal durumların tespiti ile ilgilidir [Takçı ve Soğukpınar, 2002].

Veri madenciliği kullanılarak geliştirilen nüfuz tespit sistemleri anormallik tabanlı yaklaşıma yakındır. Anormalliklerin tespiti için veri madenciliği yöntemleri kullanılabilir. Veri madenciliğinin doğasında yer alan özetlemeler yardımıyla veri indirgeme sayesinde gerçek zamanlı nüfuz tespitinin de yapılması mümkün olabilecektir. Ağ trafiğinden elde edilen normal trafik örüntüleri veri tabanlarında tutulur. Yeni girişimlerle normal trafik karşılaştırılarak nüfuz tespiti yapılır.

1.4 Bu Çalışmada Yapılanlar

Bu tez çalışmasında anormallik tabanlı nüfuz tespit sistemi için k-NN yöntemi ile kullanıcıların davranış analizleri yapılarak kullanıcılar kümelenmiş ve elde edilen sonuçlardan nüfuz tespiti yapılmıştır. Çalışmada, ağ üzerinden toplanan veriler üzerinde işlem yapılmaktadır. DARPA'nın 1999 örnek ağ verileri deneysel çalışma için kullanılmıştır. Ağ verilerinden kullanıcı örüntüleri oluşturulup bu kullanıcılar benzer özellikli olanları bir kümeleme işlemi ile bir araya getirilir. Normal trafik analizleri oluşturulan kümelerle karşılaştırılır. Analiz sonucunda oluşacak sapma miktarı, eşik değeri geçtiği anda saldırı alarmı verilir.

Ağ tabanlı NTS'ler için geliştirilen bu yöntemle kümeleme yöntemleri kullanarak analiz işlem yükü azaltılmıştır. Böylelikle ağ içerisinde oluşacak anormal durumların tespiti hızlanmıştır. Tanımlanmamış saldırıları tespit etme işlemi de daha kısalmıştır. Ayrıca sistem kaynaklarının kullanımı da yapılan istatistiksel özetlemelerle kontrol altına alınmıştır.

2. GENEL KAVRAMLAR VE KONUYLA İLGİLİ ÇALIŞMALAR

2.1 Nüfuz Tespit Sistemleri (NTS)

Bir sisteme yapılan saldırılar genellikle; harici ataklar, bir başkası gibi görünme, imtiyazı kötüye kullanma, gizli kullanıcılar şeklinde ortaya çıkmaktadırlar. Bu saldırıları etkisiz hale getirmek için nüfuz tespit sistemleri kullanılarak saldırının tanınması ve gerekli önlemlerin alınması sağlanır. Nüfuz tespiti, ağ ve ana bilgisayar seviyesinde şüpheli faaliyetlerin tespitinde kullanılan metot ve tekniklerin genel adıdır. Nüfuz tespit sistemlerinde saldırıyı başlamadan engelleme, yeni bir saldırı türü ise bunu öğrenip etkisiz hale getirme ileri için kural koyma, saldırıyı yanıltma gibi teknikler kullanılabilir [Akyüz ve Soğukpınar, 2004].

2.1.1 Nüfuz Tespit Sistemlerinin Sınıflandırılması

2.1.1.1 Veri İşleme Zamanına Göre

Nüfuz tespit sistemleri veri işleme zamanına göre iki gruba ayrılır. 1 Gerçek Zamanlı NTS, 2. gerçek zamanlı olmayan NTS.

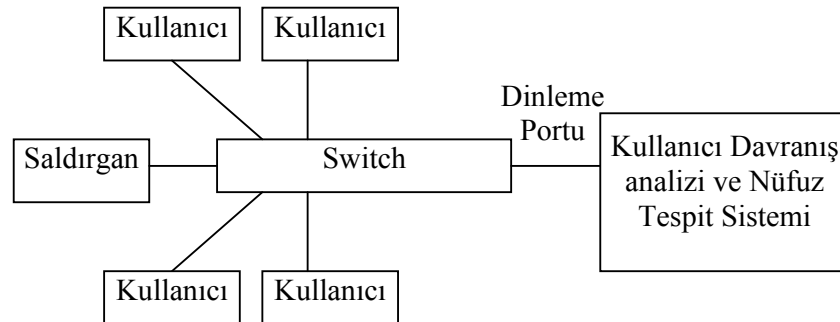
- 1- *Gerçek zamanlı NTS:* Gerçek zamanlı sistemlerde veri ağ içerisinde ki hareketi ile o anda analiz edilir.
- 2- *Gerçek zamanlı olmayan NTS:* Gerçek zamanlı olmayan sistemlerde veri önce depolanır, sonra analiz için ilgili NTS'ye gönderilir. Gerçek zamanlı sistemlerde veri o anda analiz edilir. Bu tip NTS'ler yoğun bilgi akışı olan ağlarda uygulanması zor olan ama aktif olarak cevap üretilmesi gereken durumlarda da tek çözüm olan sistemlerdir. Ticari uygulamaların gerçek zamanlı olması tercih edilir.

Bu çalışmada yapılan NTS modelimizde temel olarak sistem eğitimi için depolanmış veriler kullanılır. Bu verilerden elde edilen özetlenmiş kullanıcı örüntüleri ile gerçek zamanlı veri trafiği analizi yapılır.

2.1.1.2 Çalışma Alanını Göre:

Nüfuz tespit sistemleri çalışma alanına göre üç ana kategori altında incelenir:

- 1- *Konak(Host) tabanlı nüfuz tespit sistemleri:* sadece kendi trafiğini dinleyen uç birimdeki bir yazılımdır. Konak tabanlı NTS'leri iki tür bilgi kaynağı kullanır: işletim sistemi izleme seçenekleri ve sistem günlük dosyaları. Bunların ağ tabanlı sistemlere göre bir avantajı: olayların olduğu yerel sunucuyu izleme yetenekleri sayesinde ağ tabanlı NTS'lerin yakalayamayacağı saldırıları tespit edebilirler. Bu saldırılar çoğu zaman fazla trafik yaratmayan R2L veya U2R saldırılarını içerir.
- 2- *Ağ tabanlı nüfuz tespit sistemleri:* Temel olarak ağ üzerindeki paketleri yakalayıp bunları çeşitli kıstaslara göre değerlendirip ağ trafiği üzerinde yorumlar yapan yapılardır[Coşkun ve Soğukpınar, 2002]. Son yıllarda daha çok ağ tabanlı NTS geliştirilmektedir. Çalışmamız ağ üzerindeki veriler üzerinde işlem yaptığı için burada ağ tabanlı NTS sınıfına girmektedir.



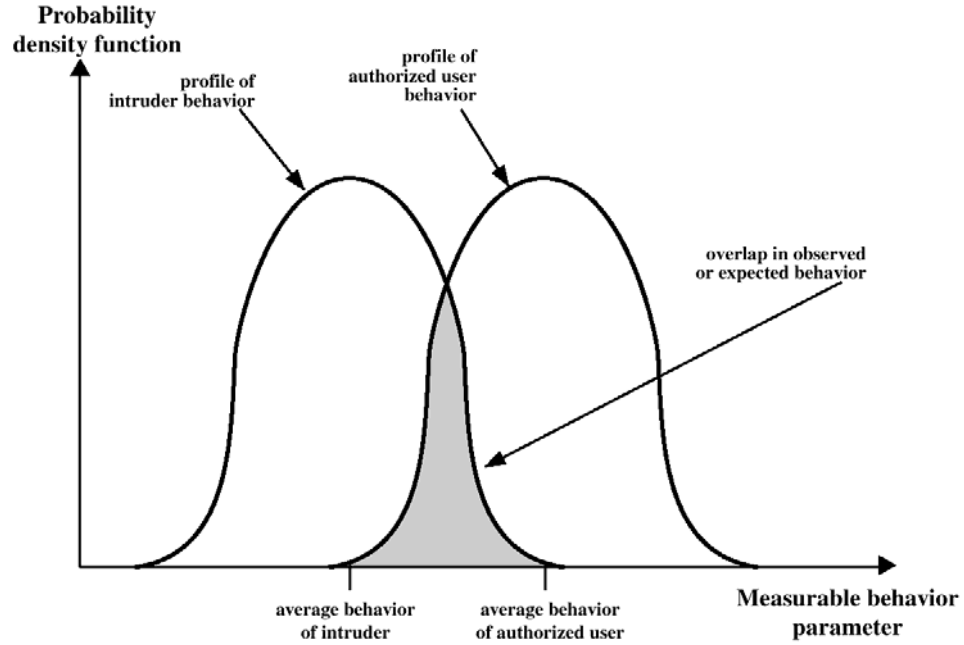
Şekil 2.1 Ağ tabanlı NTS örneği [Karabağ ve ark, 2005]

- 3- *Karma Nüfuz tespit sistemleri:* Son zamanlarda geliştirilen sistemlerde kullanılan karma NTS, ağ verisi üzerinde işlem yaparken aynı zamanda ağ üzerinde ki her bir kullanıcı üzerinde de işlem yapmaktadır. Bu tip NTS'leri işlem yükü çokluğu ile birlikte her sisteme kurulum gerektirmektedir.

2.1.1.3 Nüfuz Tespit Yöntemlerine Göre:

NTS'ler tespit yöntemlerine göre sınıflandırılırsa 2 temel yöntem söz konusudur. Bunlardan ilki anormallik tespitine dayanır. Diğer kötüye kullanım tespiti ya da bazı kaynaklarda tanımlandığı üzere imza-tanıma üzerinden gider.

- *Anormallik (anomaly) Tespiti:* Anormallik (anomaly) normal davranıştan sapma anlamına gelir. Burada normal davranıştan farklılık gösteren davranışların saldırı olarak işaretlenmesi söz konusudur. Normal bir sistemde kullanıcı istekleri tahmin edilebilir istatistiksel değerlerle uyşur. Burada normal davranışın bilinmesi ve modellenmesi esastır. Bu normal davranıştan elde edilen kullanıcı örüntüleri temel alınarak veri trafiğı gözlemlenir bir anormallik varsa tespit gerçekleşir. Bu yöntemin avantajı daha önceden tanınmayan saldırıların keşfedilmesi olasılığıdır. Dezavantajı ise yanlış alarmların (false alarm/ positive) sayısının yüksek olmasıdır[Kemmerer et al. 2002].



Şekil 2.2 Normal ve saldırı içeren dağılımın teorik ayrışımı [Stallings and ,2003]

- *Kötüye Kullanım Tespiti (Misuse Detection)*: Nüfuzları tanımak için daha önceden bilinen örüntülerden faydalanılır. Kötüye kullanım tespitinde saldırı imzalarının tutulduğu saldırı veritabanları bulunur ve bu veritabanlarında yer alan imzalar elle kodlanmak zorundadır. Sistem verileri bilinen imzalarla karşılaştırarak saldırıları tespit eder. Belirli zaman aralıklarında imza veritabanları güncellenmesi gerekmektedir. İlk kez meydana gelen saldırıları tespit etmek mümkün değildir.

2.1.2 Anormallik Tabanlı Nüfuz Tespit Sistemleri

Anormallik tespitinde istatistiksel yöntemler, yapay sinir ağları, veri madenciliği sistemi gibi birçok yaklaşım uygulanabilir. Bunları; kendi kendine öğrenen sistemler ve programlanan sistemler olarak iki sınıfta inceleyebiliriz.

- a) **Kendi-kendine öğrenen sistemler**: Belli bir veri seti ile eğitilerek normal davranışı öğrenirler. Bunlar zaman serilerini kullanmayan sistemler ve zaman serilerini kullanan sistemler ikiye ayrılırlar.

1. Zaman serisi kullanmayan sistemler:

- a. *Kural Tabalı (Rule Based) Sistemler:* Sistem kendisi trafiği inceleyip kurallar oluşturuyor ve saldırı tespiti sırasında bu kurallara göre davranıyorsa bu sınıfa girer.
- b. *Açıklayıcı İstatistikleri (Descriptive Statistics) Kullanan Sistemler:* Kullanıcı örüntüsünün basit istatistiklerle oluşturulup, bu istatistiklerin incelenmesi ile oluşturulan saldırı tespit sistemleridir.

2. Zaman serisi kullanan sistemler:

- a. *Yapay Sinir ağları (Artificial Neural Networks – ANN) yaklaşımı:* Burada sistem önceden bir zaman serisi ile eğitilir ve çalışmaya başladığında, buradan öğrendikleri ile karar verir.

b) Programlanan Sistemler: Kendi kendine öğrenen sistemlerden farklı olarak sistem zaman zaman yönetici tarafından programlanarak güncellemelerinin yapılabildiği tespit sistemleridir. Programlama işlemlerine göre üç grupta ayırabiliriz:

- a. *Açıklayıcı İstatistikleri kullanan sistemler:* Burada yukarıdakinden farklı olarak örüntüler daha önceden tanımlanmıştır. Örüntüleri sınırlayan belli bir eşik değeri, bir istatistiksel büyüklük veya basit bir kural olabilir.
- b. *Bilinmeyeni reddet (Default Deny) yaklaşımı:* Tek başına çok sık kullanılan bir yaklaşım olmasa da, sistem davranışındaki durum-geçiş değerlerini kontrol ederek, farklı bir geçiş görüldüğünde izin vermemek şeklinde bir yaklaşıma sahiptir.
- c. *Sinyal İşleme Tekniklerini Kullanan yaklaşım:* Burada sistem parametreleri sinyal işleme teknikleriyle belirlenir ve saldırı içermeyen normal veri setinden çıkartılan eşik değerleri ile anormallik tespiti yapılır. Bu yaklaşım diğerlerine göre yenidir ve çok geniş bir alanı kapsamaktadır. Özellikle son yıllarda popüler olan temel bileşen

analizi (Principle Component Analysis - PCA) yaklaşımı ile elde edilen özdeğerlerin sistem parameteresi olarak kullanılması bu alandaki çalışmalardan biridir.

2.1.3 Kötüye Kullanım Tabanlı Nüfuz Tespiti:

Kötüye kullanım tespiti (Misuse detection) ya da imza-tanıma dayalı sistemlerde her davranışın bir imzası-karakteri vardır. Bunlar daha önce görülen davranış şablonlarıdır. Eğer gözlenen davranış daha önceden bilinen bir saldırı imzası ile eşleşiyorsa saldırı olarak sınıflandırılır. Daha önce karşılaşılmadıysa saldırı olarak nitelenmez. Avantajı saldırıyı kesin olarak tanıyabilmesidir. Yani yanlış alarm vermezler fakat yeni bir saldırı gelirse bunu sezemezler. Yanlış alarm oranı çok az olduğu için kötüye kullanım tespit sistemleri ticari sistemlerde çokça kullanılırlar. Kendi kendine öğrenen hibrid bir teknik olan Ripper dışında tüm imza temelli teknikler programlanan sistemlerdir. Programlanan kötüye kullanım tespitini aşağıdaki üç şekilde inceleyebiliriz.

- a. **Durum modellemesi:** Burada tüm davranışlar belirli durumlara karşı düşer. Eğer bir davranış daha önceden tanımlı durum ve durum geçişlerine denk düşen hareketler yapıyorsa nüfuz olarak tanınır.
- b. **Uzman sistemler:** İlk Nüfuz Tespit sistemleri kural-tabanlı (rule-based) uzman sistemlerdir. Nüfuz işlemleri programlanan kurallar ile tanımlanır.
- c. **Örüntü eşleme (Pattern Matching):** Sistemde daha önceden tanımlanmış, olmaması gereken bazı sözcüklerin tanınmasına yardım eder. Esnek değildir fakat basittir. Örneğin “parola dosyasını kopyala” komutu görüldüğünde bunun bir saldırı olduğunu en basit şekilde bu yöntem tespit eder.

2.1.4 Nüfuz Tespit Sistemlerinin İncelenmesi

Aşağıda Çizelge 1 'de 2000 yılında yapılan Nüfuz Tespit sistemlerinin sınıflandırması şeması ve bunlara denk düşen gerçekleşmiş sistemler yer almaktadır [Axelsson et al. 2000].

Çizelge 2.1 Nüfuz Tespit Yöntemlerinin ve Araçlarının Sınıflandırılması
[Axelsson et al. 2000]

Anormallik Tabanlı	Kendi-kendine Öğrenen Sistemler	Zaman serisi kullanmayan	Kural tabanlı	W&S
			Açıklayıcı İstatistikler	IDES NIDES EMERALD Haystack
		Zaman serisi kullanan	Yapay sinir ağları	Hyperview
	Programlanan Sistemler	Açıklayıcı İstatistikler	Basit istatistikler	MIDAS NADIR Haystack
			Basit kural	NSM
			Eşik değeri	Computer Watch
		Bilinmeyeni Reddet	Durum Seril	JANUS Bro
Kötüye Kullanım Tabanlı	Programlanan Sistemler	Durum Modellemesi	Durum Geçiş	USTAT
			Petri-net	IDIOT
		Exper sistem		NIDES EMERALD MIDAS DIDS
		Örüntü eşleme		NSM
		Basit kural		NADIR Bro HayStack
İmza esinlenmeli	Kendi kendine öğrenen	Otomatik özellik seçme		Ripper

Çizelge 1 'de bazı sistemlerin birden fazla kategoride yer alması, ilgili yaklaşımların tümünü içerdiğini gösteriyor. Burada ki sistemlerin kısaca aşağıda açıklanacaktır.

Haystack: Amerikan Hava Kuvvetlerinde kullanılan Unisys 1100/60 ana bilgisayarları için tasarlanmış bir saldırı tespit sistemidir. 1988 yılında geliştirilmiştir[Axelsson et al, 2000].

MIDAS: 1988’de National Computer Security Centre (Ulusal Bilgisayar Güvenlik Merkezi) için geliştirilmiştir. Uzman sistem tabanlıdır. [Axelsson et al. 2000].

IDES (Intrusion Detection Expert System): IDES 1988 - 1992 yılları arasında üzerinde çalışılan birçok sistemi içerir. Bu arada birçok sürümü çıkmıştır ve son olarak NIDES (Next-Generation Intrusion Detection Expert System) adını almıştır. [Axelsson et al. 2000]

W&S (Wisdom and Sense – Zeka ve His): Geliştirilmesine 1984 yılında başlanmasına rağmen ilk makalesi 1989’da çıkmıştır. Zekâ kısmı geçmişteki denetleme verilerini inceleyerek normal davranışı oluşturma anlamına gelir. His kısmı ise bunların kural haline getirilip uzman sisteme verildikten sonra anormal davranışların yakalanması anlamına gelmektedir. [Axelsson et al. 2000]

ComputerWatch: AT&T Bell Laboratuvarları tarafından ticari bir ürün olarak 1990 yılında geliştirilmiştir. [Axelsson et al. 2000]

NSM (Network Security Monitor): NSM’de IDES gibi revizyonlardan geçmiştir ve 1990-1994 arasına geliştirilmiştir. NSM ağı dinleyerek, ağın kullanımıyla ilgili bir örüntü geliştirir ve geçerli kullanımı onunla karşılaştırır. Elde edilen veri beklenen bağlantı verisiyle karşılaştırılır ve beklenen aralıkta çıkmayan her veri anormal olarak işaretlenir[Axelsson et al. 2000].

NADIR: 1991-93 yılları arasında Los Alamos Laboratuvarlarında geliştirilmiştir. NADIR haftalık istatistikler tutar. Daha sonra bu istatistikleri uzman sistem kurallarıyla karşılaştırır[Axelsson et al. 2000].

Hyperview: 1992 yılında geliştirilmiş diğer sistemlerden oldukça farklı bir sistemdir. İki ayrı parçadan oluşur. İlki davranışları izleyen ve sınıflayan bir uzman sistem, ikincisi buradan öğrendikleriyle eğitilen yapay sinir ağlarını içeren parçadır[Axelsson et al. 2000].

DIDS (Distributed Intrusion Detection System): 1992 yıllarında geliştirilmiş dağıtık mimarili sistemleri kapsar. Bu tip sistemlerde ağın değişik noktalarından veriler toplanır ve bir merkezde incelenir[Axelsson et al. 2000].

USTAT: 1993-1995 yılları arasında geliştirilmiştir. Durum geçiş analizi yapar. Eğer bir davranış saldırı için tanımlı durum geçişlerini yapıyorsa saldırı olarak sınıflandırılır[Axelsson et al. 2000].

IDIOT: 1994-1996 yılları arasında CERIAS (Center for Education and Research in Information Assurance and Security) geliştirilmiştir. Örüntü tanıma için petri-net'lerin kullanımına dayanır[Axelsson et al. 2000].

JANUS: 1996'da Berkley'de geliştirilmiş bir sistemdir.

EMERALD (Event Monitoring Enabling Responses to Anomalous Live Disturbances): 1997-1998 yıllarında geliştirilmiş ölçeklenebilir, dağıtık bir NTS'dir.

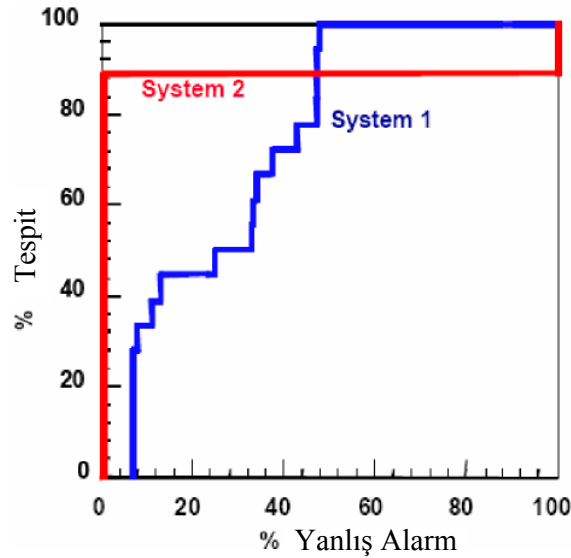
Bro: 1988'de Vern Paxson tarafından geliştirilmiş bir NTS'dir. Gerçek zamanda ağ trafiğini pasif olarak gözlemleyerek saldırı tespiti yapmaktadır. Birçok değişik özelliği kendinde barındırması açısından sıkça söz edilen bir NTS'dir[Axelsson et al. 2000].

Ripper: 1999 yılında geliştirilen bu sistem veri madenciliği yaklaşımını kullanır. Ripper DARPA değerlendirmesine katılmış olan sistemlerden bir tanesidir[Axelsson et al. 2000].

2.1.5 Nüfuz Tespit Sistemlerinin Değerlendirilmesi

Genel anlamda NTS'lerin değerlendirilmesin de saldırının doğru olarak tespit edilmesi, saldırı olmayan bir davranışı saldırı olarak göstermemesi ve saldırıyı doğru olarak göstermemesi gibi ölçümler kullanılır. Bu değerlendirmeler sonucunda NTS'lerin başarımı ortaya çıkarılır. Aşağıda bu değerlendirmeleri detaylı olarak verilmiştir [Mell et al. 2003].

1- Yanlış Alarm (False Alarm) Seviyesi: Yanlış alarm, tespit edilmesi planlanan büyüklük için yapılan yanlış değerlendirmeleri içerir. Yanlış değerlendirmeler iki çeşit olabilir. Biri var olan bir değeri kaçırmak diğeri var olmayan bir değeri varmış gibi tespit etmektir. Yanlış alarm ikincisidir. NTS’de yanlış alarm sayısı ve doğru tespit miktarı birbiriyle ilişkilidir. Yanlış alarmlara izin verildikçe doğru tespit oranı artmaktadır. Sistem parametreleri ikisinin de optimum olduğu noktaya ayarlanmalıdır. Burada yapılacak analiz ve başarımlar hesaplamasında ROC (Receiver Operating Characteristic – Alıcı Karakteristiği) eğrileri kullanılır. Burada örneğin %40’lık bir yanlış alarmı izin verilirse sistemin doğru yakalama şansı %50 olur. Eğer %60’lık yanlış alarm kabul edilirse %100’lük bir doğru tespit şansı vardır. [Mell et al. 2003].



Şekil 2.3 Örnek ROC eğrileri [Mell et al. 2003].

2- Doğru tespit seviyesi: Kaç adet saldırılarının doğru olarak tespit edildiğidir. Yukarıda anlatıldığı gibi doğrudan yanlış alarm ile ilgilidir ve Şekil 2.3’deki y-eksenini oluşturur. Bir NTS’nin var olan saldırıların kaçını yakaladığı önemli bir parametredir. Fakat bunu yaparken yanlış alarm sayısını da çok yükseltmemelidir.

3- Yoğun veri ile başa çıkabilme: Bu kriter ağ tabanlı sistemler için geçerlidir. Trafik yoğunluğu arttıkça tüm sistemler paket düşürebilir hale gelirler. Bazı NTS’ler trafik yoğunluğu çok arttığı zaman saldırı tespitinde tamamen başarısız

olabilirler. Özellikle DoS saldırılarında sistemin güvenli olması için bu önemli bir parametredir.

4- Olayları ilintilendirebilme: Farklı kaynaklardan gelen saldırı bilgilerini birleştirebilme yeteneğidir.

5- Yeni saldırıları tanıyabilme: Bu sadece anormallik tespitine dayalı sistemler için geçerlidir. İmza temelli sistemler yeni gelen hiçbir saldırıyı tanıyamazlar.

6- Saldırıyı tanıyabilme: Saldırının varlığını tespit etmek ve saldırının tipini söyleyebilmek iki ayrı kavramdır. Genelde saldırı tipi, saldırı olduktan sonra ağ yöneticisi tarafından çeşitli kayıtlar incelenerek ortaya çıkartılır.

7- Saldırının başarılı olup olmadığını belirleme: Aslında gerçekleştirilen saldırıların hepsi başarı ile sonuçlanmaz yani sisteme zarar vermez.

2.1.6 Alınacak Önlemler

Nüfuz olduktan sonra alınabilecek önlemler nüfuzun tekrar olmasını engellemeye yarar. Örneğin yazılım açıklarından kaynaklanan nüfuzları önlemek için yazılım yaması yapmak olabilir. Ama asıl önemli olan nüfuz anında önlem alabilmektir. Nüfuzlara karşı iki farklı önlem gerçekleştirilebilir; pasif gözetleme ve erişim engelleme.

a) **Pasif Gözetleme:** Bu yaklaşımda saldırganın davranışları kaydedilir, nüfuzda araya girilmez. Amaç, saldırganın istediği şeyi ve kullandığı yöntemleri öğrenmektir. Burada en büyük sorun, sistemin savunmasız bırakılmasıdır.

b) **Erişimi Engelleme:** Burada amaç saldırganın erişebileceği sistem kaynaklarını sınırlamaktır.

Nüfuza karşı yukarıdaki önlemler alındıktan sonra, nüfuza sebebiyet veren sistemlerin erişimi kaldırılır veya süreçleri sonlandırılır.

2.1.7 Testlerde Kullanılan Veri Setleri

NTS'lerin sınanmasında en zorlu kısım güvenilir veri setleri elde edilmesidir. İnternet ortamından elde edilmiş verilerin çoğu saldırı var ya da yok bilgisi içermez. Herhangi bir ağda bu bilginin edinilmesi için ağ yöneticisinin sistemi gözlüyor olması ve sistem kayıtlarını tutuyor olması gerekir. Bu birçok ISP için çok masraflı ve gereksiz görülen bir iştir. Nüfuzları içeren veri setine tümüyle hâkim olunmak isteniyorsa, bir benzetim yapılarak sentetik test verisi yaratılabilir. Fakat bunun da internet trafiğine benzediğini kanıtlamak zordur. Bu konudaki zorlukları kısaca özetlersek:

- Gerçek veri bulmak zordur.
- Atak sırasında etiketlenmiş veri seti bulmak daha da zordur.
- Verileri saklamak masraflıdır.
- Normal davranışı tanımlamak ve benzetimini elde etmek zordur.

Bütün bu zorluklara rağmen NTS'lerin test edilmesi için bir veri seti şarttır. Bu durumda 2 değişik veri toplama yöntemi kullanılabilir. Bir yöntem yönlendirici MIB'inde (Management Information Base – Yönetim Veri Tabanı) tutulan, giren-çıkan IP paketi ve bunun gibi bazı sayaç değerlerinin SNMP ile sorgulanarak bir makineye aktarılmasından ibarettir. Bu yöntem sıkça kullanılmaktadır çünkü maliyeti düşüktür. Fakat veri 5 dakikada bir sorgulandığından ve 5 dakikanın ortalaması sonuç olarak geri döndüğünden dolayı yeterince hassas değildir. Ayrıca ağ yöneticisi ile işbirliği içinde olunması gerekir aksi takdirde her yönlendirici üzerindeki bilgilere erişim hakkı tanımaz. Bir diğer problem bu yöntemle sadece yönlendirici MIB'inde tutulan verilerin elde edilebilir olmasıdır. Yani sayaç değerleri dışında, örneğin gidip-gelen paketlerin içeriği gibi bilgilere kesinlikle ulaşılamaz. Burada paket içeriğinde belirli bir anahtar kelime nüfuz anlamına geliyorsa, tespit edilmeden sistemden geçer. Bu durumda zaman serilerini kullanan yaklaşımlar ile analiz yapılabilir ve sadece trafik değerlerinde değişikliğe yol açan nüfuzlar tespit edilebilir.

Diğer yöntem ise IP akış ekranlarından veya ağ dinleyicisi (sniffer) programlarından veri toplanmasıdır. IP akış ekranları genelde pahalıdır. Sadece paket

başlıklarına bakıldığından dolayı sınırlı bir bilgi içerir. Yoklayıcı programlar ise en ayrıntılı bilgiyi içerirler ve dolayısıyla çok fazla disk alanına ihtiyaç duyarlar. Ayrıca çok yoğun trafiklerde paketlerin hepsini yakalayamayabilirler.

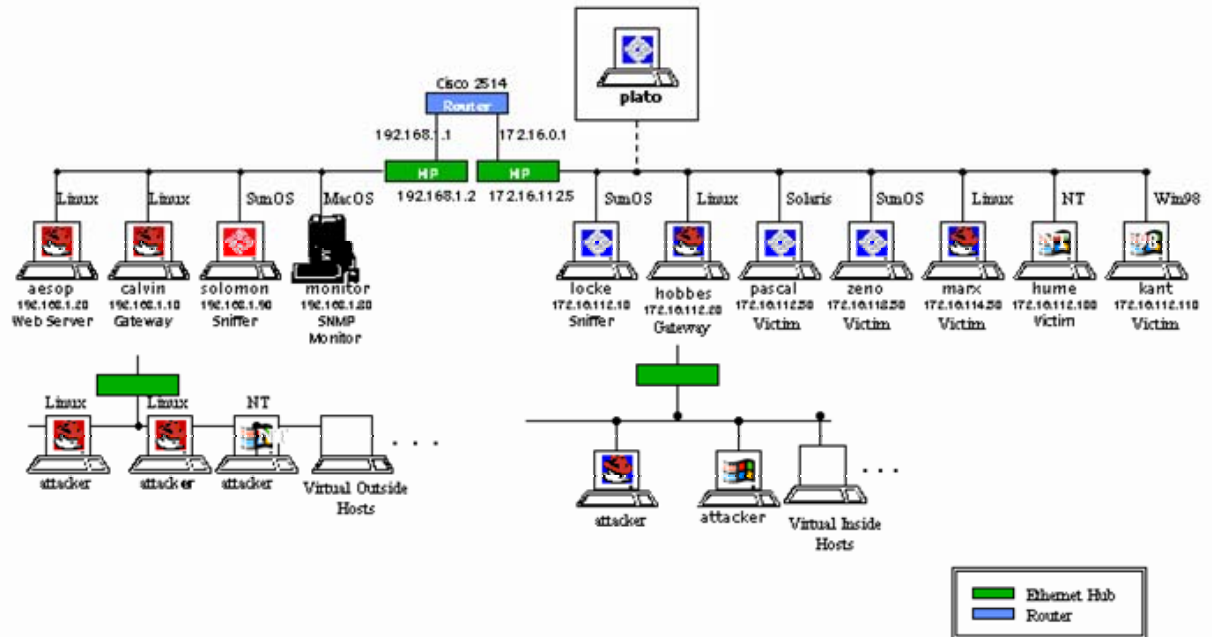
Benzetimi yapılan bir ağ ve en sık kullanılan veri seti DARPA 1998 ve 1999 veri setleridir. Bu veri setleriyle ilgili bilinen problemler olmasına rağmen kontrollü bir çerçevede üretilip, ortak kullanıma açık olan veri setleridir.

2.1.7.1 DARPA Veri Seti Detayları:

DARPA'nın sponsorluğunda MIT Lincoln Laboratuvarlarında NTS'ler için bir karşılaştırma ortamı sunan IDEVAL veri setlerini oluşturulmuştur. Özellikle saldırı tespit sistemlerinin sınanması amacıyla saldırıların hedefi olacak bir ağ ve saldırıları gerçekleştiren başka bir ağ tasarlanmıştır. İç ağda Amerikan Hava Kuvvetlerindeki bir yerel ağın benzetimi gerçekleştirilmiştir. 1999 değerlendirmesinde kullanılan ağ aşağıdaki gibidir. [DARPA, 2006]



Simulation Network 99



Hava kuvvetlerini temsil eden iç ağ 172.16.0.0/16 IP adres uzayına sahiptir. Bu ağ içerisinde dört “kurban” makine bulunmaktadır. Bunların üzerinde SunOS, Solaris, Linux, ve Windows NT çalışmaktadır. (1998 veri setlerinde sadece UNIX makineler kullanılmıştır.) Şekilde görülen trafik oluşturucular yüzlerce sunucuyu ve çeşitli uygulamaları çalıştıran İnternet kullanıcılarını benzetimini yapmaktadır. Protokollerin (HTTP, SMTP, telnet,...) karışımı, trafik yoğunluğunun saatlik değişimleri, 1998 ’de gerçek Hava Kuvvetleri ağından toplanan trafiğe benzer olacak şekilde tasarlanmıştır. Ağ üzerinden 2 noktadan veri toplanmıştır: dört kurban makine ile yönlendirici arasındaki iç ağ dinleyicisi ve yönlendirici ile İnternet arasındaki dış ağ dinleyicisi üzerinden [DARPA, 2006]. Bu veriler inside ve outside dosyalarında ayrı ayrı depolanmıştır.

1999 değerlendirmesi iki aşamadan oluşmaktadır. Veri setleri toplam 5 haftalık bir süreyi kapsamaktadır, ilk 3 haftası eğitim (training) verileri, son 2 haftası da test verileridir. 1. hafta ve 3. hafta saldırı içermemektedir. 2. hafta bazı saldırılar içermekte ve bu saldırılar ayrıca bir dosyada işaretlenmiştir. Bu kısım değerlendirmeye alınacak sistemlerin kendi başarımlarını test etmeleri içindir. Geriye kalan 4. ve 5. haftalarda ise 58 değişik saldırı tipinde 200 adet saldırı mevcuttur. Bu saldırı yazılımları internet ortamından ve hacker sitelerinden toplanmış saldırılardır.

1999 değerlendirmesine 8 grup 18 IDS ile katılmış ve katılan gruplara aşağıdaki dosyalar sağlanmıştır.

- Ağ dinleyicisinden elde edilen dosyalar
- İç ağ dinleyicisinden alınan tcpdump dosyası
- Dış ağ dinleyicisinden alınan tcpdump dosyası
- Solaris Temel Güvenlik Modülü audit dosyaları
- Windows NT olay log dosyaları
- Kurban makinelerden her gece alınan güvenlik kayıtları,

Bu dosyalar şu anda İnternet sitesinden de indirilebilir durumdadır [DARPA]. Değerlendirmenin ilk aşamasında katılımcılara üç haftalık veri dağıtılmıştır. İkinci aşamada, katılımcılara sistemlerini test edebilmeleri için 4. ve 5. hafta test verileri de

dağıtılmıştır. Tüm günler için yukarda belirtilen 5 tip dosya sağlanmıştır. Tahmin edileceği üzere exper sistem tabanlı çalışmalar 2, 3 ve 4'ü kullanırken, anormallik tespiti yapan sistemler 1 numaralı verileri kullanmayı tercih etmişleridir. Veriler 5 hafta için 5'er günlük olmak üzere toplam 25 günlük trafik içerir. Verinin toplam boyutu 7,5 GB civarındadır. Her gün sabah 08:00 ve ertesi gün sabah 06:00 arasında 22 saat veri toplanması sonucu oluşturulmuş verilerdir. Gün başına ortalama 41MB veri düşmektedir[DARPA, 2006]. Bir günde oluşan trafik internet ortamına kıyasla düşük seviyededir fakat dosya boyutları ve bu dosyaların internet ortamından paylaşılacak olması dolayısıyla bu şekilde tutulmuştur. 4 tip saldırı, hizmet engelleme (DoS), Bilgi tarama (probe), yönetici hesabı ile yerel oturum açma (R2L), kullanıcı hesabının yönetici hesabına yükseltilmesi (U2R) saldırıları çeşitli saldırı sitelerinden toplanan kodlarla gerçekleştirilmiş ve sisteme aralıklarla enjekte edilmiştir.

DARPA veri setleri bazı özellikleri dolayısıyla eleştiriler almıştır Bundan başkaca KDD-99 Cup Veri Setleri olarak da bilinen veri seti mevcuttur. Aslında bu veri seti de DARPA 1998 veri setinden bazı özelliklerin ayıklanmış olduğu ve 5. Uluslararası Knowledge Discovery and Data Mining konferansı çerçevesinde düzenlenen bir yarışmada kullanıma sunulan veri setleridir. Sonuçta yine DARPA verileri ile aynı özelliklere sahiptir.

DARPA veri setleri gerçek bir trafiğe benzetilerek oluşturulsa da gerçek trafikten olan bazı verileri içermemektedir. Bundan başka arka plan trafiği, internette sıkça gözlenen paket fırtınalarını içermemektedir. Yanlış protokol gerçeklemeleri sonucu ortaya çıkan bu davranış anormallik içerir. Saldırı olmayan veri setlerinde de bu paket fırtınaları yer alabilirdi. İletişim trafiğinin bilinmemesi ve kararsız oluşu bir başka eksik yanıdır. Bazı günler 10Kbps iken bazı günler 50Kbps çıkmaktadır.

McHugh'ün çalışmasından başka DARPA veri setlerine ait özbenzeşimliliği inceleyen bir başka çalışma 2003 tarihinde yapılmıştır. Burada da Hurst parametresinin günün değişik saatlerinde gündüz ($H=0.7$), gece ($H=0.5$) farklı değerler aldığı yani özbenzeşimli yapının her zaman korunmadığı söylenmiştir [Allen et al, 2003].

2.2 Veri Madenciliği

Veri madenciliği; büyük miktardaki veri içerisinde anlamlı bilgilerin çıkarıldığı tekniğe denir[Takçı ve Soğukpınar, 2002]. Veri madenciliği değişik alanlarda kullanılmaktadır. Son zamanlarda istatistik verileri ile birlikte anormallik tabanlı güvenlik uygulamalarında da kullanılmaktadır.

2.2.1 Veri Madenciliği Nedir ve Nasıl Yapılır?

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir. Veri madenciliği, verilerin içerisindeki desenlerin, bağlantıların, kuralların ve düzensizliklerin çıkarılmasıdır. Elde edilen verilerin istatistiksel ve otomatik olarak keşfedilmesidir. Bu bağlamda kullanılan yazılımlar ve analiz yöntemleri veri madenciliğinin en temel unsurlarıdır.

Veriler arasındaki kuralları, ilişkileri, özellikleri belirlemek için programlar kullanılır. Bu işlemler gerçekleşirken amaç daha önceden fark edilmemiş veri desenlerinin bulunması önceden belirlenen desenlerle karşılaştırılarak farklılıkların bulunmasıdır.

Veri madenciliği işlemi için öncelikle problem tanımlaması yapılmalıdır. Daha sonra veri madenciliğinin kullanılacağı alanlar tespit edilmelidir. Veri madenciliği girdi olarak veri tabanlarında tutulan ham veri üzerinde işlem yapar. Bunun için problemin ihtiyacına göre veri toplama işlemi gerçekleştirilir. Veri toplama işlemi sırasında kaliteli veri setleri oluşturma bilmek için gereksiz bilgilerin çıkarılması, tekrarların ayıklanması, veri uyumunun sağlanması, eksiklerin tamamlanması gibi işlemler yapılarak veri hazır hale getirilir.

Elde edilen veri tabanlarında tutulan verinin eksiksiz geniş ve net veri içermemesi durumunda sorunlar oluşur. Sorunları sıralarsak,

- a. Sınırlı bilgi: veri tabanları başka amaç için tasarlandığında öğrenme bilgisi yeterli olmayabilir.
- b. Gürültü ve eksik değerler: veri özellikleri içerisindeki hatalar ve tamamlanmamış eksik bilgiler gürültü oluşturur. Buda ölçüm yanlışlıklarına sebep olabilir.
- c. Belirsizlik: tahmin yapan bir keşif sistemindeki en önemli husustur. Belirsizlik yanlışlık ve gürültü derecesi ile ilgilidir.
- d. Konu dışı hata: veri tabanında bulunan konu dışı verilerin istatistik kararlarını etkilemesidir.

2.2.2 Veri Madenciliği Teknikleri:

Karar Ağaçları: Görsel olarak birtakım kurallar, şartlar tanımlayarak, kararın verilmesine imkân sağlamaktadır. Çoğunlukla sinir ağları (neural networks) ile birlikte kullanılır.[Han and Kamber, 2006]

Genetik Algoritmalar: Diğer veri madenciliği algoritmalarını geliştirmek için kullanılan optimizasyon teknikleridir. Sonuç model veriye uygulanarak gizli kalmış kalıpları ortaya çıkarılmakta ve bu sayede tahminler yapılabilmektedir. Doğrudan postalama, risk analizi ve perakende analizlerinde kullanılabilir. [Han and Kamber, 2006]

Sinir Ağları: Bu yöntem, belirli bir profile uyuşması için kalıp düzenlerini kontrol etmektedir ve bu süreç içerisinde belli bir öğrenme faaliyeti gerçekleştirerek sistemi geliştirmektedir. Kümeleme, sıralama, tahmin yapma problemler için uygun bulunmaktadır. [Han and Kamber, 2006]

Tahmin Modeli: Geleceğe dair tahminler yapmak üzere kullanılır. Kalıpları ve şekil düzenlerini tanımlamak için kullanılan çeşitli teknikleri ifade eder. [Han and Kamber, 2006]

Kural Modeli: İstatiksel önemine göre veri üzerinden “eğer...öyleyse” kurallarını oluşturmakta faydalıdır. Risk analizi ve risk yönetimi alanlarında kullanılır. [Han and Kamber, 2006]

Bulanık Mantık: Muğlak kavramları içerir. (büyük,küçük,yaşlı,genç) Diğer tekniklerden daha esnektir. [Han and Kamber, 2006]

k-NN (k- Nearest Neighbour) k- en yakın komşu değer: Bir objeyi niteliklerini inceleyerek bir sınıf veya gruba yerleştirir. [Han and Kamber, 2006]

Kural Dönüştürme: Veri madenciliğinin OLAP ve diğer veri analiz uygulamaları ile entegrasyonu için oldukça önemlidir. Ayıklanmış kuralları SQL sorgusuna dönüştürme becerisindedir. [Han and Kamber, 2006]

2.2.3 Veri Madenciliği İşlevleri

Veri madenciliğine işlevleri açısından bakılacak olursa, veri aktiviteleri 3 sınıf altında toplanmıştır. Keşif (discovery), tahmini modelleme (predictive modelling) ve adli analiz (forensic analysis)

Keşif, ne olabileceği konusunda önceden belirlenmiş bir fikir ya da hipotez olmadan veri tabanı içerisinde gizli desenleri arama işlemidir. Asıl amaç, bulunacak desenlerin zenginliği ve bunlardan bilginin kalitesidir.

Tahmini modelleme, veri tabanından çıkarılan desenler, geleceği tahmin için kullanılır. Keşif, verideki desenleri bulmaya yönelikken, tahmini modelleme, bu desenleri yeni veri nesnelerini bulmak için uygular.

Adli analiz, normal olmayan ya da sıra dışı veri elemanlarını bulmak için, çıkarılmış desenleri uygulama işlemidir. Adli analiz, keşifte aranılan genel bilginin tersine, sıra dışı ve özel durumları araştırır.

2.3 Veri Madenciliği ve Nüfuz Tespit Sistemi

Veri madenciliği tabanlı nüfuz tespitinin en önemli avantajı ağ trafiğindeki düzensizliklere bakarak daha önceden meydana gelmemiş saldırıları tespit edebilmesidir. Bunun için ağ üzerindeki veri trafiği eğitilen trafik bilgileriyle karşılaştırır ve düzensizlikler anormallik olarak tespit edilir.

Kullanıcı Davranış Analizi ile Nüfuz Tespit (KDA_NT) modeli, öncelikle ağ üzerindeki trafiği dinleyerek kullanıcıları özelliklerine (kullandığı port, veri miktarı, bağlantı saatleri, kullandığı protokoller vs.) göre kümelere atar. Veri madenciliği yaklaşımında, önce kullanıcıların normal davranışları öğrenilir ardından aktif kullanıcı erişimleri ile normal kullanım örüntüsü karşılaştırılarak tespit yerine getirilir. Bu yaklaşım kendi kendini eğitime tabanlı bir yaklaşımdır. Sistem çalıştığı sürece öğrenme işlemi de sürekli devam eder. Tespit işlemi için her zaman en güncel veri tabanı kullanılır.[Karabağ ve ark. 2005]

Kullanıcılar tarafından alınan veya gönderilen veri paketleri analiz edilerek hangi kullanıcı anormal davranışlar gösteriyorsa o kullanıcı anormal kullanıcılar kümesine dâhil edilir. Bu çalışmada kullanıcı davranışları kullanıcıların sebep olduğu trafiğin istatistiksel özetleri ile sunulur. Analiz sırasında, kullanıcı davranışları saldırı niteliği kazandığı anda daha önceden belirlenen kısıtlamalar getirilir. Böylelikle daha önceden bilinmeyen saldırılar önlenmiş olur.

Veri madenciliği tabanlı kullanıcı davranış analizi şu aşamalardan oluşur.

1. Bilgilerinin Toplanması
2. Kullanıcı Örüntülerinin Oluşturulması
3. Kullanıcıların Kümelenmesi
4. Analizlerinin Yapılması

TCPdump modelin veri toplama aşamasında yer almakta olup modelin veri toplama işlemi istenilen özelliklere göre veri tabanına aktarmak için Ethereal'dan faydalanılmıştır. Bilgi toplama aşamasında, hangi verilerin toplanacağına karar verilir ve o bilgiler haricindeki veriler hesaba katılmaz. Bu işlem seçim olarak da

bilinir. Seçim aşamasından sonraki aşama temizleme aşaması olup bu aşamada kullanım verisi ilişkisiz sahalardan temizlenir. Ardından kullanıcı tabanlı soyutlamalar yapılır. Kullanıcı tabanlı soyutlama için bir yöntem kullanıcı tabanlı verilerin özetlenmesidir. Örneğin, her bir kullanıcının on saniye içerisinde almış olduğu veri paketi boyu bu bilgilerden birisi olabilir.

İkinci aşamada toplanan verilerden kullanıcı örüntüleri oluşturulur. Bu her bir kullanıcı için özet verileri içerir.

Bu kullanıcı verilerine istatistik kuralları ile birlikte kümeleme teknikleri kullanılarak sistemdeki küme yapıları oluşturulur.

Daha önceden oluşturulan bu kullanıcı küme bilgileri ile kullanıcı normal kullanım örüntüleri karşılaştırılarak analiz işlemi gerçekleştirilir. Analiz işlemi sonrasında ağ içindeki kullanıcı trafiğine belirlenen kurallar uygulanır. Küme dışına çıkma veya küme değiştirme nüfuz olarak adlandırılır.

2.4 Kümeleme

Kümeleme benzerlik ve uzaklıklara dayalı olarak yerine getirilebilir. Benzerlik tabanlı kümelemede, küme içindeki elemanlar birbirine benzemektedir. Bu benzerlik küme dışındaki elemanlara benzerlikten daha fazladır. Uzaklık tabanlı kümelemede ise küme içindeki elemanlar birbirine daha yakın küme dışındaki elemanlara ise daha uzaktır.

Benzerlik ve uzaklık ölçütleri: Bir kümeye ait olan kayıt; o kümenin içindeki kayıtlara, dışındaki kayıtlara göre daha benzer veya daha yakındır. Veriler metrik olarak tanımlı ise, karakteristik değerler kullanılarak kümeler ifade edilebilir. Kullanıcı davranış analizi işlemi içinde kullanılan kümeleme k-NN algoritmasıdır. Burada amaç kullanıcının küme dışına çıkmasını kontrol etmektir. Y. Liao tarafından geliştirilen K-NN algoritması Şekil 2.5'te verilmiştir.

Figure 1: Pseudo code for the kNN classifier algorithm for anomaly detection.

```

build the training normal data set  $D$ ;
for each process  $X$  in the test data do
  if  $X$  has an unknown system call then
     $X$  is abnormal;
  else then
    for each process  $D_j$  in training data do
      calculate  $\text{sim}(X, D_j)$ ;
      if  $\text{sim}(X, D_j)$  equals 1.0 then
         $X$  is normal; exit;
    find  $k$  biggest scores of  $\text{sim}(X, D)$ ;
    calculate  $\text{sim\_avg}$  for  $k$ -nearest neighbors;
    if  $\text{sim\_avg}$  is greater than  $\text{threshold}$  then
       $X$  is normal;
    else then
       $X$  is abnormal;

```

Şekil 2.5 Y. Liao tarafından geliştirilen kNN algoritması [Liao and Vemuri, 2002]

Tespit işleminde performans k değerine bağlıdır. Bu değer komşuluğun ölçüsünü vermektedir. Bu değer 5 gibi bir sayı olduğu zaman ($k=5$) normal olarak verilen bir noktaya en yakın beş noktanın da normal olarak kabul edileceği anlaşılır. Ayrıca eşik değerinin seçimi de önemli bir konudur. Yanlış seçilmesi durumunda hata oranı artmaktadır. Eşik değeri yüksek olduğu zaman bazı anormal davranışlar normal davranış göstermekte ve tespit edilmeleri mümkün olamamaktadır.

Yapı olarak sınıflandırmaya benzer fakat sınıflandırmadan temel farkı, grupların önceden belirlenmemiş olmasıdır. Farklı özelliklere göre farklı kümeler oluşabilir. Kümelemeyi sınıflandırmadan ayıran üç temel özellik vardır.

- Oluşacak küme sayısı belirsizdir.
- Kümeler hakkında bir ön bilgi olmayabilir.
- Küme sonuçları dinamiklidir.

Gerçek verilerle oluşturulmuş bir veritabanına kümeleme uygulandığında bazı sorunlarla karşılaşılır. Bunlar:

- Outlier elemanları değerlendirmek zordur.
- Dinamik veriler nedeniyle küme üyelikleri zaman geçtikçe değişebilir.
- Her kümeyi semantik anlamda yorumlamak zor olabilir.
- Bir kümeleme sorununun tek doğru çözümü yoktur.
- Kümeleme için hangi verilerin uygun olduğu belirsizdir.

Outlier sorunu: Veri setinin kalan kısmından çok büyük farklılıklar gösteren verilere denir. Bunlar ya hatadır, ya da sıra dışı gerçektir. Kümeleme algoritmaları outlier verileri belirleyerek dikkate almayabilirler. Ancak burada amaç bazında dikkatli olunmalıdır çünkü bazı durumlarda sıra dışı değerlerin peşinde olunabilir. Ayrıca birçok kümeleme algoritmasında, oluşması istenen küme sayısı önceden verildiği için sorunlar oluşur.

Kümeleme algoritmaları ikiye ayrılır geleneksek algoritmalar ve yeni algoritmalar.

- **Geleneksel algoritmalar:** Küçük boyutlu, sayısal ve belleğe sığan veritabanları için uygundur.

o *Hiyerarşik algoritmalar:* Hiyerarşik algoritmalarda oluşacak küme sayısı belli değildir ve kümeler *agglomerative (bottom-up)* veya *divisive (top-down)* teknik kullanılarak oluşturulur. Dendogram adı verilen, ağaç şeklinde bir veri yapısı kullanılarak kümeler oluşturulur. Kümeler arasında doğal bir yuvarlanmış yapı gösteren, taksonomi benzeri uygulamalar için daha uygundur.

o *Partitional algoritmalar:* *Partitional* algoritmalarda ise oluşacak küme sayısı baştan belirlenir. Önerilen çözümün ne kadar iyi olduğu bir metrik kullanılarak ölçülür. (Minimum Spanning Tree, Squared Error, K-Means, Nearest Neighbor, PAM, CLARA, CLARANS, BEA, Genetik, Neural Networks, v.b.)

- **Yeni algoritmalar:** Veritabanı boyutundan bağımsız olarak sıkıştırılabilen veya özetlene bilen veri belleğe yerleştirilir.

o *Large DB algoritmaları:* Bu algoritmalar örnekleme ve sıkıştırma amacıyla veri yapılarına başvurur. Geniş veritabanlarında kümeleme yapmak için bazı kriterler belirlenmiştir:

- Veritabanının bir kez veya daha az tarama yapılması
- Online çalışabilme özelliği
- Suspendable, stoppable ve resumeable olma
- Veri ekleme veya çıkarma sonucunda güncelleme imkânı
- Kısıtlı bellek ile çalışabilme
- Tarama sırasında farklı teknikler kullanabilme
- Bir kaydın sadece bir kez işlenmesi

(BIRCH, DBSCAN, CURE, v.b.)

o *Kategorik algoritmalar:* Kategorik algoritmalar bellek kısıtlarını aşmak için kullanılır. Geleneksel algoritmalar kategorik veriler üstünde her zaman etkili çalışmazlar. (ROCK, v.b.)

2.5 En Yakın Komşu Kuralı (kNN)

Şekil 2.5’de gösterilen Liao kNN algoritmasında, x_i örüntüleri D_j sınıfları ile doğru olarak ilişkilendirilmiş örnekler bulunmaktadır. Aynı j sınıfı ile ilişkili olan x örüntülerinin birbirlerine benzer olduklarını söyleyebiliriz. Aynı sınıf örüntüleri genellikle bir bölgede kümeleşirler. Bu durum, sınıflandırılmamış x örüntülerini **en yakın komşusu** olan j sınıfına dahil edileceği işaret edilir.

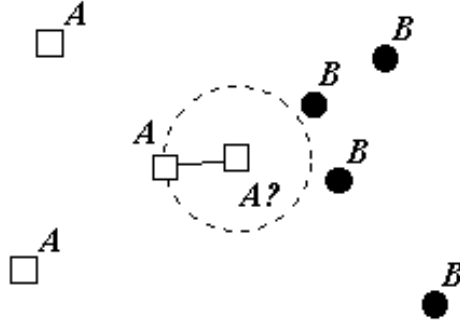
Basit sınıflandırma kurallarından bu biçimindeki olanı en yakın komşu yaklaşımı k-NN olarak adlandırılır ve matematiksel olarak aşağıdaki gibi ifade edilir.

$$D(x_i) = j \quad (j \in [1, m]) \quad (2.1)$$

$$D(x) = D(\min(\text{dist}(x, x_i))) \quad (i \in [1, n]) \quad \forall x_i \text{ için} \quad (2.2)$$

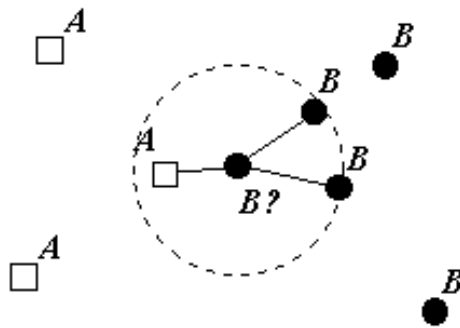
En yakın komşu yaklaşımı x örüntüsünün sınıfını, x örüntüsüne en yakın olan örüntüsünün sınıfı olarak belirleme yaklaşımıdır. Sınıfı belirlenen örüntü ile komşu

örüntü aynı sınıfa ait değiller ise hata söz konusudur. Bu yaklaşım sadece en yakın komşu ile sınıflandırma yapar, önceden sınıflandırılmış diğer örüntüleri önemsemez.



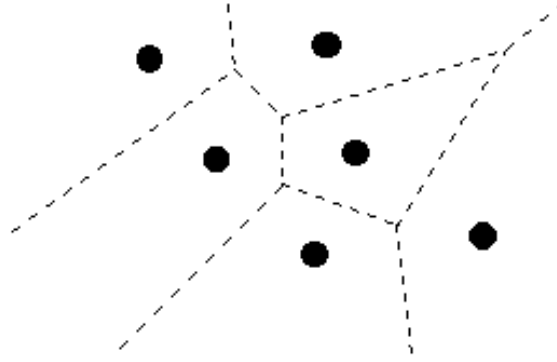
Şekil 2.6: En Yakın Komşu Kuralı uygulaması

k-Yakın Komşu Kuralı: Önceden sınıflandırılmış örneklerin sayısı arttıkça, sınıf belirleme işleminin başarımı artar; Fakat, çoğu zaman tek uzaklıktaki komşuluklar yetinilmez. Bir noktanın k komşusuna bakılarak sınıfının belirlenmesi yaklaşımı k -yakın komşu kuralı olarak adlandırılır. k sayısı; hatalı sınıflandırmayı azaltmak için büyük seçilmeli (*önceden sınıflandırılmış örnek sayısına göre*) küçük tutularak da yakın komşuların etkileri artırılmalıdır.

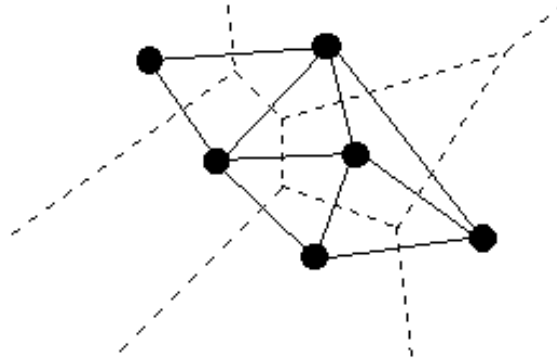


Şekil 2.7: K Yakın Komşu Kuralı ($k=3$)

Voronoi Çizeneği Yaklaşımı: Bir noktaya en yakın bölgelere bölünmesi ile elde edilen çizeneğe **Voronoi Çizeneği** denilir. Noktaları birleştiren doğru parçalarının ortalarından geçen dikmeler yardımıyla bu çizenek elde edilebilir.

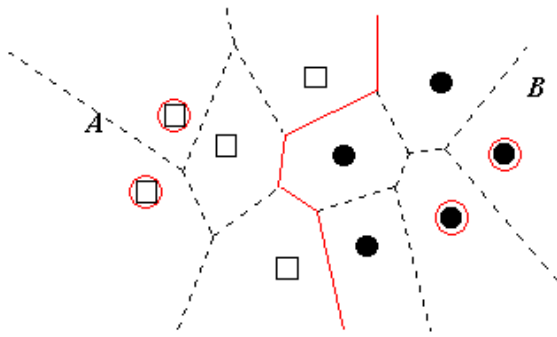


Şekil 2.8: Voronoi Çizeneği (1)



Şekil 2.9: Voronoi Çizeneği (2)

En yakın komşu kuralı ile sınıfı belirsiz bir örüntü çizenekteki bölgesine göre sınıflandırılabilir. Sınıflandırma sırasında değişik sınıflara ait bölgeler arası sınırlar önem kazanır.



Şekil 2.10: Voronoi Çizeneği ile En Yakın Komşu İlişkisi

x örüntüsü ve Y sınıfına ilişkin iki durum ortaya çıkabilir: Birincisi, x örüntüsü ve Y sınıfının dağılımına tamamen hakim bir istatistiksel dağılıma sahiptir. İkincisi; x örüntüsünün ve Y sınıfının dağılımı hakkında bilgimiz yoktur fakat daha önceden sınıflandırılmış örneklerle mevcuttur.

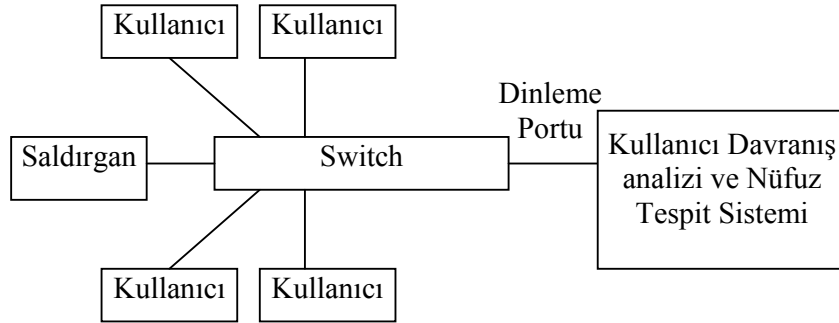
En yakın komşu kuralı ikinci durumla ilgilidir. Bu tür problemlerin sınıflandırılması parametrik olmayan istatistik yöntemler olarak adlandırılır.

3. ÖNERİLEN MODEL

Yapılan çalışmada ağ trafiğinden toplanan veriler veritabanına aktarılır. Oluşturulan veritabanı üzerinden kullanıcı davranış örüntüleri oluşturulur. Bu kullanıcı bilgileri k-NN algoritması ile kümelere aktarılır. Eğitim işlemi bittikten sonra network trafiği gerçek zamanlı dinlenerek kullanıcı davranışları oluşturulan kümeler ile karşılaştırılır. Analiz işlemi sonrası küme değiştiren kullanıcı veya küme dışına çıkan kullanıcı nüfuz girişimi olarak tespit edilir ve alarm mekanizması devreye girer. Yapılan model aşağıda detaylı olarak anlatılmıştır.

3.1 Modelin Yapısı

Modelimiz temel olarak iç ağ ortamında merkezi bir anahtar üzerinde çalışır. İç ağda oluşan network trafiği kullanılan anahtar (switch) üzerinden tek bir uca toplanarak KDA-NT yöntemi tarafından analiz edilir.

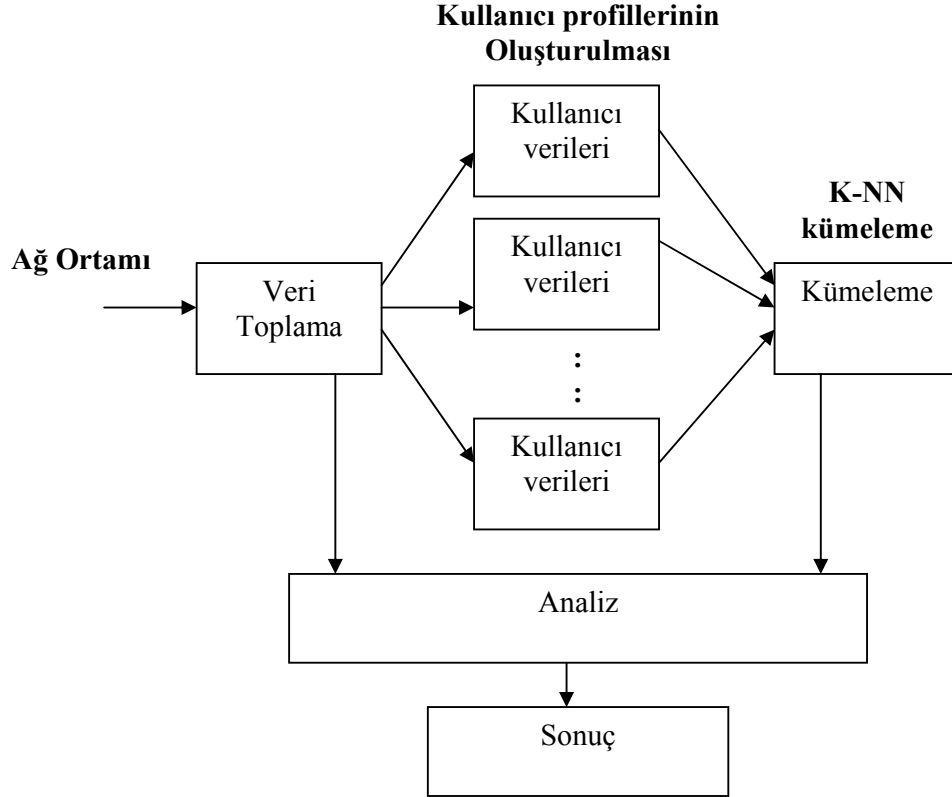


Şekil 3.1 KDA ve NTS ağ yapısı

Nüfuz tespitinde daha önceki yöntemlerden farklı olarak KDA-NT yönteminde kullanıcı temelli ağ analizi yapılmaktadır. Modelimizde veri toplama işlemi gerçekleştirildikten sonra bu verilerden kullanıcı örüntüleri oluşturulur. Bu kullanıcı örüntüleri ayrı bir tabloda depolanır. Yapılan kullanıcı tablosu k-NN algoritması ile belirli kümelere birleştirilir ve oluşturulan küme değerleri ayrı bir tabloda tutulur. Bu eğitim işlemi tamamlandığında yöntemimiz gerçek zamanlı olarak ağı verisi dinlenerek nüfuz tespit işlemi gerçekleşir. KDA-NT yönteminin blok diyagramı Şekil 3.1’de gösterilmiştir. Ağdan toplanan veri kullanıcı bilgilerini içerir. Kullanıcı

profilleri oluşturulurken her bir kullanıcının network trafiğinde ki kullanım istatistiklerinden faydalanılır. Kullanıcı bilgilerinin tutulduğu bir dizi ile tanımlanır.

$$Kullanıcı: X(OZ_1, OZ_2, OZ_3, \dots, OZ_M) \quad (3.1)$$



Şekil 3.2 Kullanıcı davranış analizi ile Nüfuz tespit sistemi modeli

kNN kümelemeyi 2002 Liao ve Verumi tarafından yapılan nüfuz tespit sisteminde ağdan toplanan sistem çağrılarını kümelerken kullanmıştır. İki uygulamanın mukayesesi Çizelge 3.1’de gösterilmiştir. Burada sistem çağrıları kümelenirken yaptığımız çalışmada ağ kullanıcıları kümeleme işlemine tabi tutulur. Temelde her iki sistemde anormallikler için eğitim verileri ile karşılaştırmayı kullanır. Yaptığımız modelde OZ_{ij} kullanıcı özelliklerini verir ve bu özellikler Z gibi Zamanda geriye dönük olarak alınan frekans sayısı ile belirlenir.

Çizelge 3.1: kNN ile yapılan nüfuz tespit sistemleri:

Terim	Sistem çağrılarına kNN sınıflandırılmasıyla nüfuz tespiti	Kullanıcı Davranışlarına kNN sınıflandırılmasıyla nüfuz tespiti
X	Test işlemi	Kullanıcı
D_j	j . eğitilmiş işlem	Karşılaştırma kümesi j.
N	Toplam işlem Sayısı	Kullanıcı Sayısı
M	farklı sistem çağrılarının toplam sayısı	Kullanıcı için seçilen özellik sayısı
n_i	i. sistem çağrısının geçtiği işlem sayısı	i özelliğinin geçtiği kullanıcı sayısı
f_{ij}	i. sistem çağrısının j işlemindeki frekansı	j kullanıcısı için i özelliğinin frekansı
T	Eşik Değer	küme farkı hesabı için kullanılan eşik
a_{ij}	J işlemindeki i. Sistem çağrısının ağırlığı	j kullanıcısı için i özelliğinin ağırlığı
Z		Zamanda geriye dönük olarak alınan frekans sayısı
OZ_{ij}		j kullanıcısı i. özellik frekans ortalaması

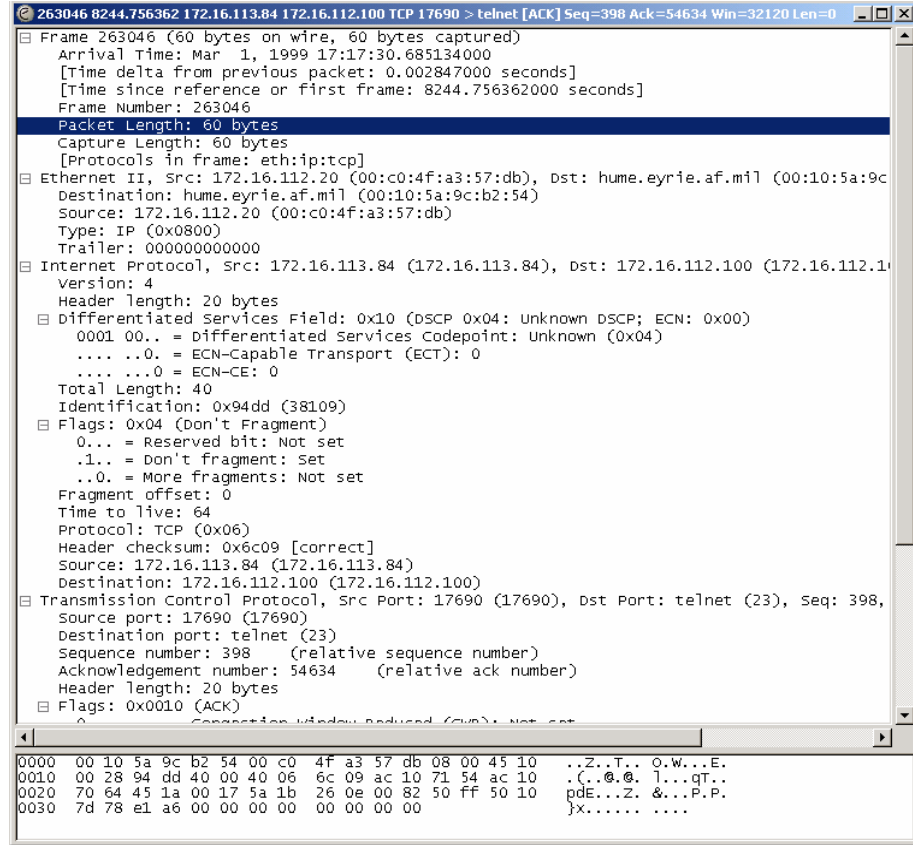
3.2 Kullanıcı Tanımları

Kullanıcı MAC adresi ile tanımlanan IP adresi ile gösterilir. Kullanıcının davranışları, ağ trafiği içerisinde oluşturduğu trafik sinyallerinden çıkarılır. Özellik için zaman serileri kullanılır. Özellik çıkarımı şimdiden geriye dönük olarak çıkarılır ve çalışma saatleri esas alınarak son 1 günlük bağlantıların istatistikleri kullanılır. Gelen paketlerin porta göre bağlantı sayıları bulunur. TCP, SMTP port , SSH port, FTP port, HTTP port, gibi bağlantı sayıları kullanıcıyı tanımlar. Ağ trafiğinden elde edilen bir TCPdump verisi örneği aşağıdaki gibidir.

```

0000  00 10 5a 9c b2 54 00 c0 4f a3 57 db 08 00 45 10  ..Z..T..O.W...E.
0010  00 28 94 dd 40 00 40 06 6c 09 ac 10 71 54 ac 10  .(..@.@.l...qT..
0020  70 64 45 1a 00 17 5a 1b 26 0e 00 82 50 ff 50 10  pdE...Z.&...P.P.
0030  7d 78 e1 a6 00 00 00 00 00 00 00 00 00 00 00  }x.....

```



Şekil 3.3 TCPdump paket yapısı Ethereal incelemesi

Yukarıdaki veri detaylı incelendiği zaman bu verileri özetlemeleri ile kullanıcı tablosu oluşturulur. Özellikler seçilirken kullanıcı trafiğinde ki değişimlerde göz önünde bulundurularak bir hafta belirli zaman aralıklarına bölünmüştür. Örneğin bir kullanıcının mesai saatinde oluşturduğu ağ verisi mesai dışında oluşturduğundan çok farklı olacaktır. Özellik hesaplamasında Z, zamanda geriye dönük olarak alınan frekans sayısı f_{ij} saniyedeki j kullanıcısı için i özelliğinin frekansıdır. Aşağıda gösterilen formülde, Z adet frekans ortalaması j kullanıcısının i. özelliğini verir.

$$OZ_{ij} = \frac{\sum_{k=1}^Z f_{ij}}{Z} \quad (3.2)$$

Kullanıcı özelliklerinin tanımlayıcı olduğunu belirlerken her bir özelliğe bir ağırlık tanımlanır. Bu tanımlama da A matrisinde tutulur. Burada N Kullanıcı sayısı,

n_i , i özelliğinin geçtiği kullanıcı sayısı, OZ_{ij} özellik frekans ortalamasıdır. A matrisinde her bir özelliğin kullanıcılardaki ağırlıkları olan a_{ij} değerleri tutulur.

$$a_{ij} = OZ_{ij} \times \log\left(\frac{N}{n_i}\right) \quad (3.3)$$

Çizelge 3.2 Kullanıcı tabloları

MAC tanımlı Kullanıcı	OZ_{ij}	OZ_{ij}	.	.	.	.	Norm
1- (00:10:5a:9c:b2:54)172.16.112.100	2750	1500	0	0	8672	...	10546,10748
2							

3.3 Benzerlik

Kümesi belirlenemeyen X kullanıcısı k yakın komşu algoritmasına tabi tutulur. Bu kullanıcı için elde edilen eğitim dokümanından kullanıcının profili oluşturulur. K adet yakın komşusuna bakılır ve bu komşuların çoğunluğun olduğu kümeye dahil edilir. Her komşu için benzerlik ölçümü yapılır. Benzerlik ölçümü cos. değeri ile hesaplanır.

$$Sim(X, D_j) = \frac{\sum_{t_i \in (X \cap D_j)} x_i \times d_{ij}}{\|X\|_2 \times \|D_j\|_2} \quad (3.4)$$

t_i : X ve D deki i. ortak özelliği

x_i : X kullanıcısındaki t_i özelliğinin ağırlığı

d_{ij} : D kullanıcısındaki t_i özelliğinin ağırlığı

$\|X\|_2$: X 'in Normu $\|X\|_2 = \sqrt{x_1^2 + x_2^2 + x_3^2 + \dots}$

$\|D\|_2$: D 'nin Normu $\|D\|_2 = \sqrt{d_1^2 + d_2^2 + d_3^2 + \dots}$

3.4 K-Yakın Komşu (kNN) ile Kümeleme:

Elde edilen kullanıcı örüntüleri daha önceden de anlattığımız k-NN algoritması ile kümeleme işlemine tutulur. Burada k için kullanıcı yoğunluğuna bağlı olarak

tespit edilir ve ona göre ölçüm yapılır. Elimizde $\mathbf{X}$ örüntüleri $\mathbf{D}_j$ sınıfları ile doğru olarak ilişkilendirilmiş örnekler bulunmaktadır. Aynı j sınıfı ile ilişkili olan $\mathbf{X}$ örüntülerinin birbirlerine benzer olduklarını söyleyebiliriz. Aynı sınıf örüntüleri genellikle bir bölgede kümeleşirler. Bu durum, sınıflandırılmamış $\mathbf{X}$ örüntülerini **en yakın k komşusu** olan $\mathbf{D}_j$ sınıfına dahil edileceğine işaret eder. $\mathbf{D}_j$ kümesi :

$$D_j (OZ_1, OZ_2, OZ_3, \dots, OZ_M) \quad (3.5)$$

D_j kümeleri için kullanıcı verileri aşağıda belirtilen kNN kümeleme algoritmasından geçirilir ve ait olduğu kümeler bulunur.

```

D={t1, t2, ... tm}           // Küme elemanları kullanıcılar
A                               // elemanlar arasındaki uzaklıkların tutulduğu matris
T                               // Eşik değeri
SK                              // oluşan kümeler
Xl := {a1, a2, ... am};      // kullanıcı
k = 5;                         // k adet yakın komşu bulunur

I den n kullanıcı için hesapla //n kullanıcı sayısı
ai= sim(ti,txi)              // X. kullanıcı için A i kullanıcısının uzaklık matrisi

A uzaklık matrisinden en büyük k adet komşuyu bul
Benzerlik averajını bul sim_avg=(X,D)

J adet Küme ile karşılaştır.
Eğer sim_avg > Eşik değeri se // max benzerlik eşik değerden büyükse
SKj = SKj U Xi              // Xi kullanıcısı SKj kümesine dahil edilir.
Sonuç:
SK = {X1,X2,X3}           //SK kümesi kullanıcı listesi

```

Kümeleme algoritması çalıştırıldığında sistem yöneticisinin tanımladığı tüm kullanıcılar kendilerine en yakın komşularıyla bir araya getirilerek kümeleri oluşturulur ve bu küme değerleri kullanıcı tablolarına eklenir. Artık sistemde kullanıcı ile birlikte bulunduğu kümesi de tutulur. Sistem her an güncel verilerle

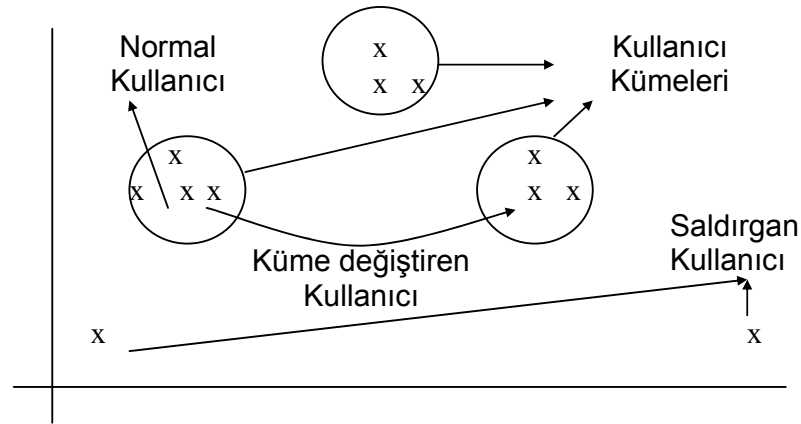
analiz işlemini yaparak küme değerlerinin yenilenmesini sağlar. Böylelikle analizde en güncel veri kullanılmış olmaktadır.

Bu işlem sonrasında her bir kullanıcının kümeleri hesaplanır ve küme merkezi her seferinde tüm kullanıcıların ortalama değerleri ile bulunur.

$$D_j (\text{Ort}(\text{OZ}_1), \text{Ort}(\text{OZ}_2), \dots, \text{Ort}(\text{OZ}_M)) \quad (3.6)$$

3.5 Tespit İşlemi

Her bir kullanıcı verileri ait olduğu küme normal alınarak kontrol edilir. X , Eğitim verisi içerisinde bulunmayan bir kullanıcı ise anormal olarak işaretlenir ve sistem yöneticisine bildirilir. Anormallik tespiti bir kullanıcının kendi kümesi arasındaki uzaklığın bir eşik değeri aşıp aşmamasına bakılarak tespit edilir. Eşik değeri aşılsa normal dışı uygulama olacaktır ve kullanıcının kendi kullanım yetkileri dışına çıktığı anlamına gelecektir. $\text{Sim}(X, D_j) \leq T$ T : Eşik değeri Eşik değeri kullanıcı merkezine en uzak kullanıcı değerine bağlı olarak sistem yöneticisi tarafından verilir. Bu tip anormal davranışlar gözlemlendiği anda uyarı sistemi devreye girerek sistem yöneticisine bildirilir.



Şekil 3.4 Kullanıcı kümelerinin bulunması ve saldırganın tespiti

Anormallik tespiti için kullanılan algoritma aşağıda verilmiştir. Burada gelen X paketi bulunduğu D kümesi ile karşılaştırılır.

Eğitim verisinden D kümeleri belirlenir,

her X kullanıcısı test verisini al,

Eğer X bilinmeyen bir kullanıcıysa

X anormaldir;

Değilse

X her kullanıcı ile D kümesi için eğitim verisini al

$Benzerlik(X, D_i)$ hesapla

Eğer $Benzerlik \geq Eşik$ değeri se

X normaldir,

Değilse

X anormaldir,

Tasarlanan modelde gerçek ağ verileri yerine DARPA'nın 99 yılında hazırlamış olduğu gerçek ağ trafiği simülasyon verileri kullanılmıştır. Bu veriler 5 haftalık bilgiyi kapsamakta ve biz bu bilgilerden TCPdump-inside iç ağdan toparlanan veriler üzerinde işlemimizi gerçekleştirdik.

4. DENEYSEL ÇALIŞMALAR VE ANALİZ

Yaptığımız çalışmada DARPA veri setlerini kullanarak ağ ortamında oluşan trafiğe kullanıcı kümeleme işlemi uygulanmıştır. Kullanıcı trafik yoğunluğunu gösteren bir zaman serisi elde edilip buradan saldırıları ve hatalı ağ kullanımlarını tespit edilmiştir. Bunun için iç ağ dinleyicisi dosyalarından, 5 haftalık verilerden saldırı içermeyen 1. ve 3. hafta ve saldırı içeren test haftası 5. haftaya ait dosyalar alındı.

KDA-NT yöntemimiz ağdan topladığı verileri kullanıcılara göre hesaplamalardan geçirerek her bir kullanıcının örüntüleri oluşturulup veritabanına aktardı. Bu bilgiler daha sonra kümeleme işlemine tabi tutulup küme tabloları oluşturuldu. Bu küme tabloları analiz işlemlerinde anormallik tespiti için kullanılacak. Bu veriler gerçek zamanlı olarak sürekli yenilenmektedir.

Sisteme yeni bir kullanıcı girmesi yönetici tarafından gerçekleştirilmektedir. Yetkisiz kullanıcıların böylelikle önlenmiş olmaktadır. Sistem yetkilisi gerektiğinde ağ izleme yönetimi programı ile ağ da oluşan trafiği inceleyebilir ve ilerisi için yeni kurallar koyabilir.

4.1 Ağ Trafiğinden Verilerin Elde Edilmesi

KDA-NT yöntemine uygun test verileri elde etmek için iç ağdan veri toplama işlemi gerçekleştirilmesi gereklidir. Ağ dinleyicisi verisi ile güvenilir zaman serisi oluşturmak çok detaylı bir işlemdir. Biz burada deney veri seti olarak DARPA'nın 1999 yılında NTS'ler için bir karşılaştırma ortamı sunan IDEVAL veri setleri kullandık. Ağda bulunan bilgisayarlar üzerinde SunOS, Solaris, Linux, ve Windows NT çalışmaktadır. Veri setleri toplam 5 haftalık bir süreyi kapsamaktadır, ilk 3 haftası eğitim verileri, son 2 haftası da test verileridir. Burada dinleyici veri seti TCPdump veri formatında olup, ağa giren ve çıkan tüm paketleri tüm ayrıntılarıyla tutmaktadır. Öncelikle bu çok fazla disk yeri tutan bir durumdur. Dinleyici verileri ağ üzerinde, çift yönlü gelip-giden tüm paketleri içerdiğinden, bir bağlantının iki yönlü paketleri, yönlendirme paketleri ve diğer tüm gereksiz trafik filtrelenmelidir. Her bir gün için bulunan TCPdump veri seti ethereal aracılığıyla açılmış ve filtrelenmiştir.

Ethereal'da filtrelenen veriler sistemimizin kullanımı için tablolara alınır. Bu tablolar kullanıcı verileri olarak gruplanır.

Ethereal' da gösterilen TCPdump verisini incelersek özet görüntüde ilk sütunda paket numarası ikincisinde paket zamanı bunu sistem zamanıyla veya ağı dinlemeye başladığı zamandan itibaren başlayacak şekilde gösterilir. Daha sonra paket adres bilgisi gösterilir. Bu kaynak IP adresi veya seçime bağlı olarak MAC adresi ve hedef IP adresi veya MAC adresidir. Bir diğer alan da protokol bilgisidir paketin TCP, UDP, ARP, TELNET gibi protokol bilgisini gösterir. Son sütunda ise paketin içeriğini açıklayıcı bilgiler gösterilir. Bunlar paket boyu, port numarası, ACK, SEQ no, pencere no işaret bayrak bilgileri, kullanılan komut bilgileri gibi...

The screenshot shows the Ethereal (Wireshark) interface. The top menu bar includes File, Edit, View, Go, Capture, Analyze, Statistics, and Help. Below the menu is a toolbar with various icons for file operations, capture, and analysis. A filter bar at the top of the packet list shows the filter 'ip.dst==172.16.112.100'. The packet list table has columns for No., Time, Source, Destination, Protocol, and Info. The selected packet is No. 60896, Time 3269.911125, Source 172.16.113.105, Destination 172.16.112.100, Protocol TCP, and Info 14972 > smtp [ACK] Seq=1 Ack=1 Len=0 MSS=1460. The packet details pane shows the following structure:

- Frame 20 (90 bytes on wire, 90 bytes captured)
- Ethernet II, Src: SunMicro_83:4a:82 (08:00:20:83:4a:82), Dst: 3com_9c:b2:54 (00:10:5a:9c:b2:54)
- Internet Protocol, Src: 172.16.112.10 (172.16.112.10), Dst: 172.16.112.100 (172.16.112.100)
- User Datagram Protocol, Src Port: ntp (123), Dst Port: 1064 (1064)
- Network Time Protocol

The packet bytes pane shows the raw data in hexadecimal and ASCII:

```

0000 00 10 5a 9c b2 54 08 00 20 83 4a 82 08 00 45 00  ..Z..T.. .J...E.
0010 00 4c d5 14 40 00 ff 11 6d fc ac 10 70 0a ac 10  .L..@... m...p...
0020 70 64 00 7b 04 28 00 38 cb 1c cc 00 04 f0 00 00  pd.{(.8 .....
0030 00 7a 00 02 7b f5 c0 a8 01 0a ba 85 0c d6 9e 4f  .z..{... .....0
0040 90 00 ba 85 0d f2 32 f1 a9 fc ba 85 0d f3 ae 05  .....2. ....
0050 20 00 ba 85 0d f2 32 f1 a9 fc ba 85 0d f3 ae 05  .....2. ....

```

The status bar at the bottom shows the file path 'C:\darpa\week1\1dst172.16.112.100tcp' and the packet number and time '325 MB 22:00:10'.

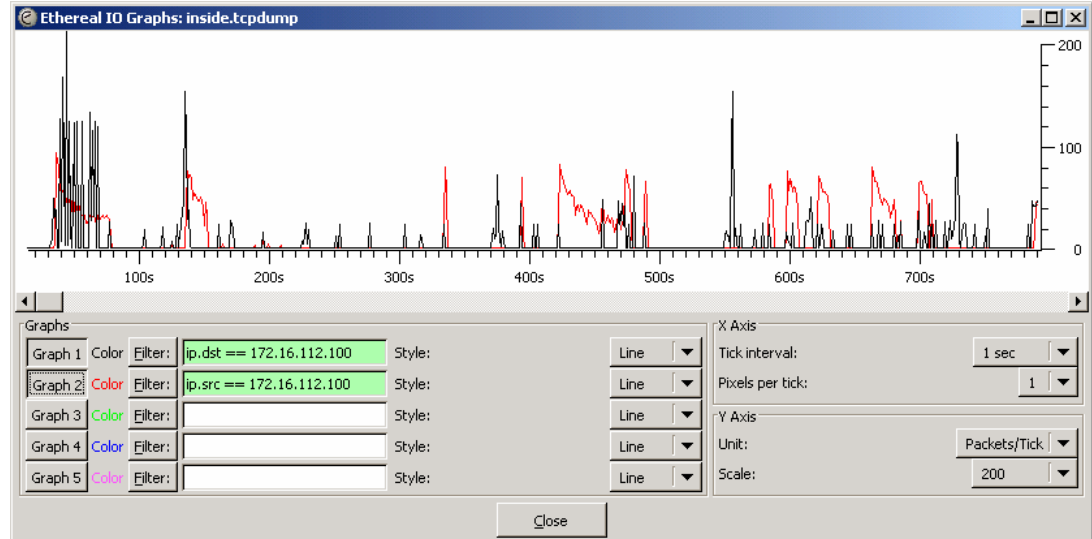
Şekil 4.1 Ethereal'da DARPA TCPdump veri analizi

İlk olarak ele alınan 1 ve 3. hafta pazartesi (saldırı içermeyen veriler) tcpdump.inside dosyası ethereal de açılmalı ve filtre uygulanmalıdır.

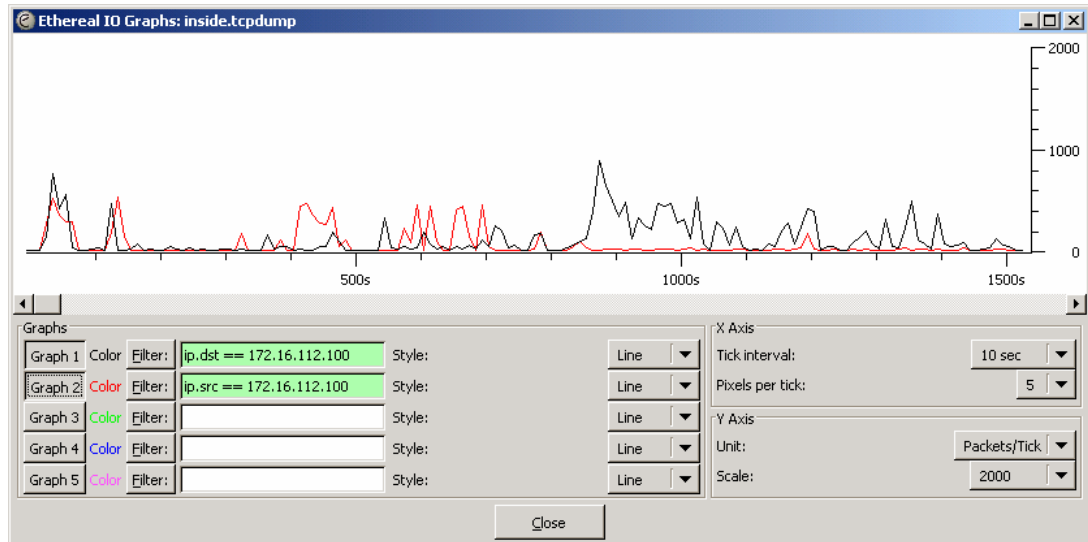
Filtre: *(ip.dst == 172.16.0.0/16) or (ip.src == 172.16.0.0/16) and !ntp and !rip and !loop and !arp and !nbns and !cdp* 'dir.

Bu işlemle ağdaki her bir kullanıcıya gelen ve giden paketler toplanmış ve bu paketlerden rip, loop, ntp, nbns paketleri filtrelenerek analiz işlemleri dışında bırakılmıştır.

Gerçek zamanlı analiz yapılırken bu filtrelemeler kullanılarak TCPdump verisinde işlem yapılır. Bu işlemlerle elde edilen veriler SQL veri tabanına aktarılır. Bu veriler üzerinde kullanıcı sorguları çalıştırılır. Her bir kullanıcıya ait veriler elde edilir. Ethereal'da elde edilen 172.16.112.100 kullanıcısına ait veriler Şekil 4.1'de verilmektedir. Bu verilerin zamana göre gelen ve giden paket sayılarına bakılırsa Şekil 4.2 ve Şekil 4.3'deki grafikler elde edilir. İlk grafikte 1 saniye aralıklarla gelen ve giden paket sayısı ikinci grafikte ise 10s aralıklarla gelen ve giden paket sayısı görülmektedir.



Şekil 4.2 172.16.112.100 kullanıcısına ait gelen ve giden paketlerin zaman serisi 1s



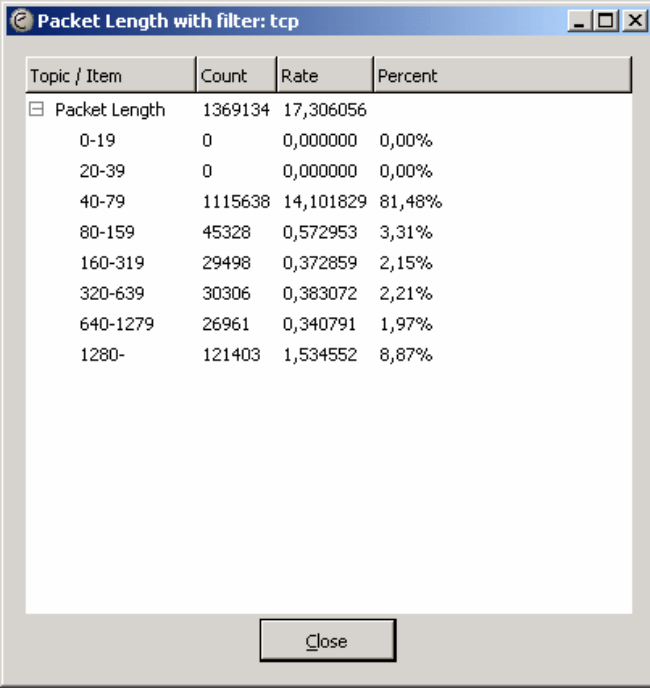
Şekil 4.3 172.16.112.100 kullanıcısına ait gelen ve giden paketlerin zaman serisi 10s

Bu kullanıcı verisini detaylı olarak tablolarsak her bir protokolün tüm trafik içerisinde yüzdelik dilimleri ve paket büyüklükleri özetlenmektedir.

Ethereal: Protocol Hierarchy Statistics							
Protocol	% Packets	Packets	Bytes	Mbit/s	End Packets	End Bytes	End Mbit/s
Frame	100,00%	1372959	287304270	0,029	0	0	0,000
Ethernet	100,00%	1372959	287304270	0,029	0	0	0,000
Internet Protocol	99,00%	1359236	286166394	0,029	0	0	0,000
Transmission Control Protocol	92,79%	1273903	275079424	0,028	586773	35206380	0,004
User Datagram Protocol	6,15%	84380	11022036	0,001	0	0	0,000
Domain Name Service	1,82%	24944	2873126	0,000	24944	2873126	0,000
Routing Information Protocol	0,55%	7594	554364	0,000	7594	554364	0,000
Network Time Protocol	0,79%	10781	970290	0,000	10781	970290	0,000
NetBIOS Name Service	0,14%	1863	204714	0,000	1863	204714	0,000
NetBIOS Datagram Service	0,02%	268	72175	0,000	0	0	0,000
SMB (Server Message Block Protocol)	0,02%	268	72175	0,000	0	0	0,000
Syslog message	0,01%	106	19055	0,000	106	19055	0,000
Simple Network Management Protocol	2,83%	38824	6328312	0,001	38824	6328312	0,001
Internet Control Message Protocol	0,07%	953	64934	0,000	953	64934	0,000
Configuration Test Protocol (loopback)	0,54%	7390	443400	0,000	0	0	0,000
Data	0,54%	7390	443400	0,000	7390	443400	0,000
Address Resolution Protocol	0,37%	5100	294984	0,000	5100	294984	0,000
Logical-Link Control	0,09%	1233	399492	0,000	0	0	0,000
Cisco Discovery Protocol	0,09%	1233	399492	0,000	1233	399492	0,000

Şekil 4.4 172.16.112.100 kullanıcısına ait protokol istatistikleri

Ayrıca toplanan veriler üzerinde kullanıcıyı tanımlayıcı diğer istatistik bilgileri de elde edilebilir. Aşağıda kullanıcıya ait TCP paketi uzunluk istatistikleri görülmektedir.



Topic / Item	Count	Rate	Percent
☐ Packet Length	1369134	17,306056	
0-19	0	0,000000	0,00%
20-39	0	0,000000	0,00%
40-79	1115638	14,101829	81,48%
80-159	45328	0,572953	3,31%
160-319	29498	0,372859	2,15%
320-639	30306	0,383072	2,21%
640-1279	26961	0,340791	1,97%
1280-	121403	1,534552	8,87%

Şekil 4.5 172.16.112.100 kullanıcısına ait TCP paket uzunlukları istatistikleri

4.2 Kullanıcılar

Kullanıcıların karakteristik özelliklerine bakarak kullanıcı tanımlaması yapılır. Kullanıcı profilleri oluşturulurken her bir kullanıcının network trafiğinde ki kullanım istatistiklerinden faydalanılır. Kullanıcı bilgilerinin tutulduğu bir dizi ile tanımlanır. Kullanıcı özellikleri gelen paket istatistiklerinden hesaplanmıştır. Kullanıcıyı tanımlayıcı Port numaraları esas alınarak paket frekansı hesaplanır. Kullanıcı karakteristiği belirleyen özellikleri ne kadar çoğaltırsak modelin güvenilirliği o derecede artacaktır. Bunun dezavantajı ise işlem yükü artacağından kullanılan PC'nin hızlı işlem gücüne sahip olması gerektirmesidir. Burada sistemden bilgi çekerken öncelik kullanıcıyı tanımlayıcı karakteristik özelliğini verecek bilgiler olmasına dikkat edilmelidir. Kullanıcı için ölçülen gelen paket ve çıkan paket bilgileri, hafta içi çalışma saati değerleridir.

4.3 Kümeleme

Kullanıcı verileri k-NN kümeleme algoritmasından geçirilerek kümeler bulunur ve benzer kullanıcılar bir kümede toplandı. Kümeleme işlemi gerçekleşirken k sayısı 3, 5, 10 olarak üç farklı şekilde test edilmiştir. Kümeleme algoritması çalıştırıldığında sistem yöneticisinin tanımladığı tüm kullanıcılar kendilerine en yakın komşularıyla bir araya getirilerek kümeler oluşturuldu ve bu kümeler kullanıcı tablolarına eklendi. Artık elimizde kullanıcı ile birlikte bulunduğu kümesi de bulunur. Sistem her an güncel verilerle analiz işlemini yaparak küme değerlerinin yenilenmesini sağlar. Böylelikle analizde en güncel veri kullanılmış olacaktır. Yapılan kümeleme işlemi sonuçları Çizelge 4.1, 4.2 ve 4.3’de gösterilmiştir.

Çizelge 4.1 : K=3 için belirlenen kullanıcı kümeleri

Küme	Kullanıcılar	Kullanılan Portlar
A	1, 11	20, 21, 25
B	2, 3, 9, 6, 13, 15	23, 25, 79
C	5, 8, 7, 10, 12, 14, 4	25, 79, 2140
D	16	2565, 31337
E	17, 18, 20, 21, 23, 22	2140, 2023, 5000
F	19	1999

Çizelge 4.2 : K=5 için belirlenen kullanıcı kümeleri

Küme	Kullanıcılar	Kullanılan Portlar
A	1, 5, 8, 11, 7, 10, 12, 14, 4	20, 21, 25, 79, 2140
B	2, 3, 9, 6, 13, 15	23, 25, 79
C	16	2565, 31337
D	17, 18, 20, 21, 23, 22	2140, 2023, 5000
E	19	1999

Çizelge 4.3 : K=10 için belirlenen kullanıcı kümeleri

Küme	Kullanıcılar	Kullanılan Portlar
A	1, 5, 8, 11, 7, 10, 12, 14, 4 , 2, 3, 9, 6, 13, 15	20, 21, 25, 79, 2140, 23, 25, 79
B	16	2565, 31337
C	17, 18, 20, 21, 23, 22	2140, 2023, 5000
D	19	1999

Çizelge 4.1, 4.2 ve 4.3'den görüldüğü gibi k sayısı değiştikçe küme grupları değişmektedir. Bu kullanıcı yoğunluğuna ve birbirlerine olan benzerlik değerlerine göre en uygun değer belirlenir. K sayısı küçük alındığında bazı kullanıcıların farklı bir kümeye ayrıldığı görüldü. Kullanıcı karakteristiğini belirleyecek çok yüksek olmayan çok da düşük olmayan bir değer uygun sonuç vermektedir. Bizim uygulamamızda $k=5$ değerinde uygun çözüm bulunmuştur. K değeri her ağ trafiğinde farklı olabilir. Sistem yöneticisi kullanıcı yoğunluğuna göre uygun değeri belirler.

Kümeler oluşturulduklarında 16 ve 19 no'lu kullanıcıların hiçbir kullanıcı ile benzeşemediği görüldü bu kullanıcılar ayrı bir küme gibi değerlendirilmektedir. Diğer kullanıcılarda davranışlarına göre kendilerine uygun kümelere yerleştirilmiştir. K değerine bağlı olarak bazı kullanıcılar değişik kümelere yer almıştır. Bu kullanıcıların kümeleme işleminde yakın komşu değerlendirmesine göre değişmektedir.

4.4 Analiz

KDA-NT yöntemimiz eğitildikten sonra eş zamanlı kullanıcı analizleri yapılır. Kullanıcı davranışları belirlenen esasları sağlamıyorsa anormallik gösterdiği için nüfuz tespit sistemi devreye girecektir. Anormal kullanıcı sistem yöneticisine bildirilir. Sistem yöneticisi anormal olarak tespit edilen kullanıcıların hareketlerini detaylı incelemeye alır. Belirli süre incelendikten sonra hala aynı konum devam ediyorsa kullanıcı saldırgan konumuna alınır ve bu kullanıcıdan gelen paketler engellenir.

Yapılan uygulamada anormallik tespiti için algoritma çalıştırıldığında Çizelge 4.4'deki sonuç bulunmuştur. Burada görünen 6 kullanıcı tespit edilmiştir. Üç kullanıcı sisteme izinsiz girmiş eğitim verisi bulunmayan kullanıcılardır. Diğerleri kendi küme özellikleri dışına çıkan kullanıcılar.

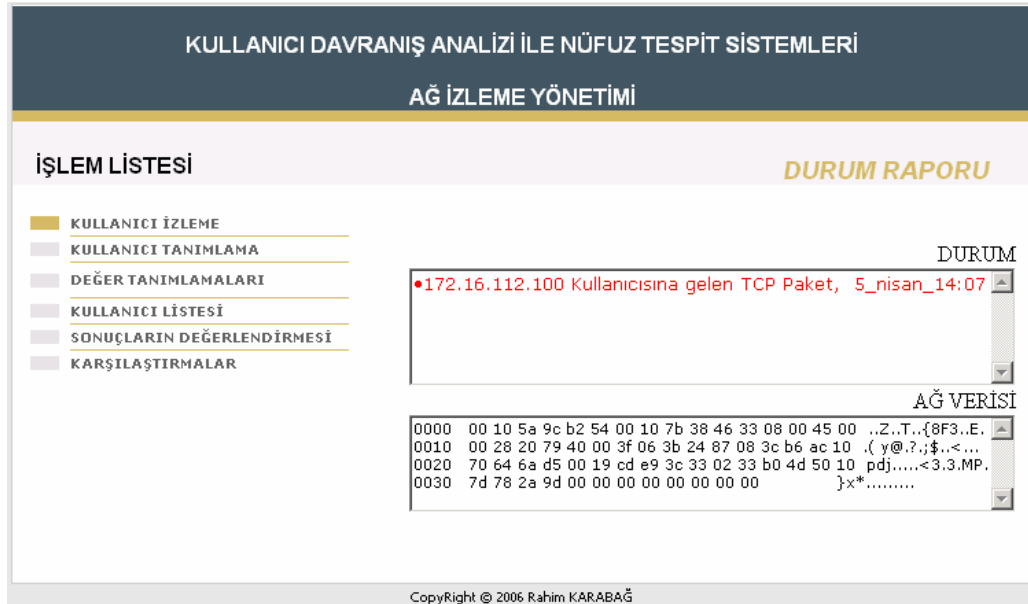
Çizelge 4.4 : K=5 anormallik analiz sonuçları

No	Kullanıcı(IP No)	Küme	Benzerlik	Onay
1	172.16.112.100	A	0,87	Normal
2	172.16.112.149	B	0,05	Anormal
3	172.16.112.194	B	0,996	Normal
4	172.16.112.20	A	0,001	Anormal
5	172.16.112.207	A	0,85	Normal
6	172.16.112.50	B	0,995	Normal
7	172.16.113.105	A	0,85	Normal
8	172.16.113.204	A	0,85	Normal
9	172.16.113.50	B	0,995	Normal
10	172.16.113.84	A	0,86	Normal
11	172.16.114.148	A	0,82	Normal
12	172.16.114.168	A	0,993	Normal
13	172.16.114.169	B	0,959	Normal
14	172.16.114.207	A	0,85	Normal
15	172.16.114.50	B	0,24	Anormal
16	172.16.115.5	C	0,82	Normal
17	172.16.115.87	D	0,853	Normal
18	172.16.116.194	D	-	-
19	172.16.116.201	D	-	-
20	172.16.116.44	E	-	-
21	172.16.117.111	D	0,87	Normal
22	172.16.117.132	D	0,852	Normal
23	172.16.117.52	D	0,91	Normal
24	172.16.118.10		X	Anormal
25	172.16.113.80		X	Anormal
26	172.16.117.103		X	Anormal

Sistem yöneticisi tarafından kullanıcı istatistikleri dağılımına bakılarak bir eşik değeri tanımlanır. Küme ortalamaları ile kullanıcı değerleri arasındaki benzerliğin en az 0,87 olduğu görülmüştür. Kullanıcı davranış değişimleri de düşünülerek sistemimizde eşik değeri 0,8 alınmıştır. Yukarıdaki tabloda kullanıcı kontrolleri yapılarak anormallikler tespit edilen değerler gösterilmiştir. Sisteme daha önce girişi yapılmayan kullanıcılar için değerlendirme yapılmaz, bu kullanıcı doğrudan anormal olarak tespit edilip sistem yöneticisine bildirilir.

Analiz işlemi ile tespit edilen kullanıcılardan kayıtlı olmayan kullanıcıları tespit işlemi %100 gerçekleşmiştir. Analizin yapıldığı gün veri setimizde 539663 paket bulunmaktadır. Bunun içerisinde 85 saldırı paketi yerleştirilmiştir. Bunların 19 tanesi dış ağ paketi olduğu için filtrelenerek işlem dışı bırakılmıştır. Kayıtlı

kullanıcılardan tespit edilen kullanıcılar incelendiğinde 172.16.112.20 IP'li kullanıcı 53 numaralı portuna 228223 paket geldiği görüldü. Eğitim verisinde 53 nolu port hiç kullanılmadığından anormallik tespiti doğrudur fakat saldırı niteliği taşımayan yanlış kullanım mevcuttur. 172.16.114.50 kullanıcısı incelendiğinde 12 saldırı bulunup bu kullanıcı için 124643 paket veri gelmiştir. Tüm kullanıcılar değerlendirildiğinde doğru tespit edilen saldırı oranı 6/7 %85 çıkmaktadır. Tüm kullanıcılara bakılarak hatalı tespitler 4/24 %16,6 olarak belirlenir. Bu hata oranı 1 kullanıcı yanlış olarak saldırı işaretlenmiş 3 kullanıcıda tespit edilememiştir. (172.16.112.50, 172.16.112.100, 172.16.113.50) tespit edilemeyen kullanıcıların verileri incelendiğinde bu kullanıcılarda paket trafiği daha yoğun olduğu gözlenmektedir. Bu da yoğun ağ trafiği olan kullanıcılar için daha belirgin kullanıcı özellikleri tanımlanmalıdır. Doğru tespit işlemi ise 22/26 %85 çıkmaktadır.



Şekil 4.6 KDA-NTS ağ izleme yönetimi

Tcpdump ile ağdan toplanan veriler sistemimizde SQL veri tabanında tutulur. Veri tabanında sorgular ve analizler yapılmakta veriler güncellenmektedir. Uygulamamız web arayüzlü ağ izleme yönetimi programı ile çalışır. Burada analiz işlemleri sonuçları zamanla durum alanında verilir. Gelen durum mesajları alttaki text alanında detaylı olarak görüne bilir. Küme normallerinin T eşik değeri kadar dışına çıkma alarm durumu olarak sistemde görünecektir.

4.5 Karşılaştırma

Yapılan çalışma kullanıcı verilerinin haftalık istatistiksel özetlerinin tutulması NADIR sistemine benzer.[Axelsson et al. 2000] Fakat burada NADIR'dan farklı olarak kurallarla karşılaştırma değil anormallik aranır. Gerçek zamanlı analiz ve bilinmeyen reddetme ile BRO nüfuz tespit sistemine benzer. [Axelsson et al. 2000] Çalışmada haftalık kullanıcı verileri toplandığından bir ağ güvenlik izlemesi yapılmıştır. USTAT modeli gibi durum geçişleri kontrol edilir fakat yapılan çalışmada bu kullanıcı verileri kümelenerek yapılır. USTAT'da daha önceden tanımlı durum geçişleri kontrolü yapılır. [Axelsson et al. 2000]

Bu çalışmada Liao k-nn algoritması ile aynı algoritma kullanılır. Liao da sistem çağrılarını gözlenerek anormallikler ile nüfuz tespiti yapılırken bu çalışmada, kullanıcıların oluşturduğu zaman serileri üzerindeki kümeler ile anormallik tespiti yapılır. [Liao and Vemuri 2002]

Avantajları:

- a. Kullanıcıların ağ kullanımındaki davranışının istatistik bilgileri elde edilir.
- b. Tüm iç ağ trafiği kontrol altına alınmış olur.
- c. Kullanıcı istatistikleri belirli kuralla değil kendi eğitim verileri ile karşılaştırılır.
- d. Saldırı olmayan fakat yanlış kullanım hataları tespit edilebilmektedir.
- e. Daha önceden tanımlanmayan saldırılar tespit edilebilmektedir.
- f. Ağda oluşabilecek trojan ve benzeri ağ trafiği oluşturan virüsleri tespit edilebilmektedir.

Dezavantajları:

- a. Hızlı işlem gücüne sahip bir bilgisayarla kapasitesi yüksek bir hafıza birimi gerekmektedir.

- b. Ağ trafiğinin tamamının görüntülenmesi için yüksek kapasiteli bir ağ portu kullanılmalıdır.
- c. Kullanıcı kullanım istatistiklerindeki değişimler sistemimizi olumsuz etkilemektedir.
- d. Sitemdeki kullanıcı davranış tanımlamaları özellik sayısı artıkça performansı azalacaktır.

5. SONUÇ VE ÖNERİLER

Bilgisayar sistemlerinin tam olarak korunmasını sağlamak amacıyla yöntem gerçek zamanlı olarak çalışmaktadır. Bir bilgisayar sistemini korumanın en iyi yolu çeşitli güvenlik araçlarını birlikte kullanmaktır.

Bu çalışmada önerilen yöntem, bilinen Nüfuz tespit sistemlerine veri madenciliği tekniklerinin uygulanmasını ve gerçek zamanlı kullanıcı analizinin yapılmasını sağlamaktadır. Önerilen yöntemde bir yandan nüfuz işlemleri tespiti yapılırken bir yandan da kullanıcı örüntüleri üzerinde istatistiksel bilgileri de elde edilmiş olur. Gerçek zamanlı analiz ile nüfuz işlemi gerçekleşmeden saldırgan etkisiz hale getirilir. Yöntem tamamen kullanıcı denetimli çalışacağından ağda bulunan yetkisiz kullanıcılar engellenmiş olacaktır. Anormallik tabanlı sistemimizde daha önceden bilinmeyen saldırı türleri tespit edilebileceği gibi kullanıcının da sistem kaynaklarını yanlış veya sisteme zarar verecek şekilde kullanması da engellenir. Bu işlem virüs, trojan ve ağda oluşturulacak saldırı niteliği taşıyan diğer programları da engelleyecektir.

KDA-NT yönteminin analizi ve kararlılığı zamanla kullanıcı özellik tanımlamaları değiştirilerek arttırılabilir. Aynı zamanda iç ağda çalışan bu yöntemi ağdaki farklı noktalara yerleştirerek dağıtık sisteme geçip ağ üzerindeki etkinliği arttırılabilir. Dağıtık oluşturulacak mimaride daha etkin saldırı önlemleri gerçekleştirilebilir.

KAYNAKLAR

- [1] Akyüz Türker, İbrahim Soğukpınar, “AKAT Akıllı Açıklık Tarayıcıları”, TBD 21. Ulusal Bilişim Kurultayı, 2004 ODTÜ-Ankara Türkiye
- [2] Allen J., A. Christie, W. Fithen, J. McHugh, J. Pickel, “State of the Practice of Intrusion Detection Technologies”, Tech. Rep. CMU, 2000.
- [3] Axelsson S., “Intrusion Detection Systems: A Survey and Taxonomy”, Technical Report Dept. of Computer Eng., Chalmers University, March, 2000.
- [4] Barbara D., J.Couto, S.Jajodia, L.Popyack, N. Wu, “ ADAM Detecting Intrusion by Data Mining”, IEEE Workshop on Information Assurance and security, 2001
- [5] Beghdad Rachid, “Modelling and solving the intrusion detection problem in computer networks”, Computers & Security, vol:23, no:8, Sayfa: 687-696, 2004
- [6] Coşkun Mustafa, İbrahim Soğukpınar, “Dağıtık Saldırı Tespit Sistemleri için bir Model”, 19. Bilişim Kurultayı, İstanbul-Türkiye, 2002.
- [7] DARPA Intrusion Detection Evaluation, web sitesi, “<http://www.ll.mit.edu/IST/ideval/index.html>” , 2006
- [8] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques Second Edition”, Morgan Kaufmann Publishers, 2006
- [9] Jeffrey M. Stanton, Kathryn R. Stam, Paul Mastrangelo and Jeffrey Jolton, “Analysis of end user security behaviors”, Computers & Security, vol.24 no:2 Sayfa.124-133, 2005,
- [10] Karabağ R., H.Takcı, T.Akyüz, "Kullanıcı Davranış Analizi ile Nüfuz Tespit Modeli (KDA-NTM)" ABG2005, Ağ ve Bilgi Güvenliği Sempozyumu, 9-12 Haziran 2005, İstanbul

[11] Kemmerer R.A., G. Vigna, "Intrusion Detection: A Brief History and Overview", IEEE Computer Special Issue on Security and Privacy, vol. 35, no.4, Sayfa. 27-30, Apr. 2002

[12] Liao Yihua, V.Rao Vemuri, "Use of K-Nearest Neighbor Classifier for Intrusion Detection", Computer and Security vol:21 no:5 Sayfa:439-448,2002

[13] Manavoğlu Eren, Dmitry Pavlov, C.Lee Giles, "Probabilistic User Behavior Models", Third IEEE International Conference on Data Mining ICDM03-Computer Society 2003

[14] McHugh John, Alan Christie, and Julia Allen, Software Engineering Institute, "http://www.cert.org/archive/pdf/IEEE_IDS.pdf" CERT Coordination Center, IEEE SOFTWARE sayfa 42-51, Eylül/Ekim 2000

[15] Mell P., V. Hu, R. Lippmann, J. Haines, M. Zissman, "An Overview of Issues in testing Intrusion Detection Systems", NIST Technical Report, 2003.

[16] Orebaugh Angela , "Ethereal Packet Sniffing", Syngress Publishing, 2004

[17] Salvatore J.Stolfo, S.Hershkop, Ke Wang, O.Nimeskern, Chia-Wei Hu, "Behavior Profiling of Email", 1st NSF/NIJ Symposium on Intelligence & Security Informatics (ISI 2003). Tucson, Arizona,USA, 2003

[18] Sequeria K., Mohammed Zaki, "ADMIT Anomaly Based Data Mining for Intrusions", eighth ACM SIGKDD International Conference on Knowledge discovery and data mining table of contents, sayfa386-395, 2002

[19] Stallings W., "Network Security Essentials", Prentice Hall, 2003

[20] H.Takçı, İ.Soğukpınar, "Saldırı Tespitinde Yeni Bir Yaklaşım",19.TBD Bilişim Kurultayı, 3-6 Eylül 2002, İstanbul Türkiye 2002

[21] Tien Lu Chang, A.P.Boedihardjo, p.Manalwar, “Exploiting Efficient Data Mining Techniques to Enhance IDS”, IEEE International Conference on Information Reuse and Integration, Las Vegas, Nevada, West Point New York US, 2005

[22] Yegneswaran V., P. Barford, J. Ullrich, “Internet Intrusions: Global Characteristics and Prevalence”, ACM SIGMETRICS, 2003.

[23] Verwoerd T. and Ray Hunt, “Intrusion detection techniques and approaches”, Computer Communications, Vol.25, No:15, Sayfa: 1356-1365, 15 September 2002

ÖZGEÇMİŞ

1977 yılında İstanbul Üsküdar’da doğdu. İlk ve orta okulu 1983-1991 yılları arasında sırasıyla Sokullu Mehmet Paşa İlkokulu ve Burhan Felek Lisesi’nde tamamladı. Liseye 1991 yılında Haydarpaşa Endüstri Meslek Lisesinde Elektronik bölümünde başladı. 1992 Yılında Haydarpaşa Teknik Lisesi Bilgisayar bölümüne geçti ve 1995’de mezun oldu. Üniversiteye 1996’da Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümünde başladı. 1998 yılında İstanbul Üniversitesi Bilgisayar Mühendisliği bölümüne yatay geçiş yaptı ve 2001 yılında buradan mezun oldu. İstanbul Üniversitesinde 1999 ve 2000 yıllarında öğrenci asistan olarak çalıştı. 2002 yılında Gebze Yüksek Teknoloji Enstitüsü Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü’nde yüksek lisans eğitime başladı. Halen Gebze Yüksek Teknoloji Enstitüsü Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü’nde Araştırma Görevlisi olarak çalışmaktadır.