

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

AKIŞ ÇİZGESİ TABANLI NÜFUZ/SIZMA TAHMİN YÖNTEMİ

ÖZNUR ERDUR SEVER
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GEBZE
2013

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

AKIŞ ÇİZGESİ TABANLI NÜFUZ/SIZMA
TAHMİN YÖNTEMİ

ÖZNUR ERDUR SEVER
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
PROF. DR. İBRAHİM SOĞUKPINAR

GEBZE
2013



GEBZE YÜKSEK
TEKNOLOJİ ENSTİTÜSÜ

YÜKSEK LİSANS JÜRİ ONAY FORMU

GYTE Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 17/06/2013 tarih ve 2013/32 sayılı kararıyla oluşturulan jüri tarafından 27/09/2013 tarihinde tez savunma sınavı yapılan Öznur ERDUR SEVER'in tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof.Dr. İbrahim SOĞUKPINAR

ÜYE

: Doç.Dr. Hacı Ali MANTAR

ÜYE

: Yrd.Doç.Dr. Serdar Süer ERDEM

ONAY

GYTE Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

ÖZET

Bilgi Teknolojilerinin gelişimi ile birlikte sistem ve sisteme dâhil olan varlıkların içinde bulunduğu tehditler artmaktadır. Bu sebeple bilgi güvenliğinin sağlanması daha da önem kazanmaktadır. Günümüzde tasarlanan bilgi sistemlerinin altyapısı genellikle bilgisayar ağlarına dayanmaktadır ve bu da bilgi sistemlerini saldırılara daha açık hale getirmektedir.

Ağ güvenliğinin sağlanmasında Nüfuz Tespit Sistemleri (Intrusion Detection Systems, IDS) önemli bir araçtır. Nüfuz Tespit Sisteminde uygulanan teknikler ile saldırılar, ancak kısmen ya da tamamen gerçekleştikten sonra saptanabilmektedir; buna bağlı olarak saldırının kontrol altında tutulması ya da durdurulması zor olmaktadır. Bu nedenle yeni geliştirilecek olan IDS sistemlerine saldırıyı tahmin etme özelliği dâhil edilmelidir. Son zamanlarda üzerinde çalışılan bu yeni sistemler Nüfuz Tahmin Sistemleri (Intrusion Prediction Systems, IPS) olarak adlandırılmaktadır.

Bu çalışmada Bilgisayar ağlarına olan nüfuz/sızmaların önceden tahmin edilmesine yönelik bir yöntem önerisinde bulunulmuştur. Önerilen yöntem test edilerek sonuçlar verilmiştir.

Anahtar Kelimeler: Sızma Tahmin Sistemleri, Grafa Dayalı Sızma Tahmini, İçerik Tabanlı İmza Çıkartılması, Hibrit İmza Tanımlaması.

SUMMARY

Along with the improvement of Information Technologies, the threat; that the system and the included entities are in, has been increasing. Therefore ensuring information security gains more importance. The infrastructure of the information systems designed lately, is generally built upon computer networks; which makes information systems more prone to attacks.

Intrusion Detection Systems are important tools in providing network security. The attacks in Intrusion Detection Systems (IDS) can only be detected after they occur partially or fully. And because of this, taking attacks under control or ceasing them is difficult. For this reason, attack prediction feature should be included to the new IDS systems to be designed. These new systems studied on recently are named as Intrusion Prediction Systems (IPS).

In this research; a method is suggested to predict intrusions through computer networks. Experimental results have been presented after testing the method.

Keywords: Intrusion Prediction Systems, Graph Based Intrusion Prediction, Content Based Signature Generation, Hybrid Signature Definition.

TEŞEKKÜR

Başta, yüksek lisans eğitimim süresince bana bilgisi ile yol göstererek yardımlarını esirgemeyen değerli hocam Prof. Dr. İbrahim SOĞUKPINAR'a,

Eğitimimin her aşamasında bana her konuda destek olan ve hiçbir fedakârlıktan kaçınmayan sevgili eşim Y. Müh. Süleyman Zafer SEVER'e,

Beni yetiştiren ve hayatım boyunca her konuda desteklerini benden eksik etmeyen annem Melahat ERDUR ve babam Feridun ERDUR'a,

Eğitimimi önemseyen ve devamlılığında katkısı olan Yaltes A.Ş.'deki yöneticilerim ve iş arkadaşlarıma içten teşekkürlerimi sunarım.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	iv
SUMMARY	v
TEŞEKKÜR	vi
SİMGELER ve KISALTMALAR DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
TABLolar DİZİNİ	xi
1. GİRİŞ	1
2. NÜFUZ TAHMİN SİSTEMLERİ ve İLGİLİ ÇALIŞMALAR	5
2.1. İstatiksel Yöntemler	5
2.1.1. Dinamik Bayes Ağları	7
2.1.2. Saklı Markov Modeli	9
2.1.3. Saklı Yarı-Markov Modeli	10
2.2. Veri Madenciliği Yöntemleri	11
2.2.1. Sınıflandırma Yöntemi	12
2.2.2. Kümeleme	13
2.2.3. Birliktelik Kuralı ve İlişki Analizi	13
3. AKIŞ ÇİZGESİ TABANLI NÜFUZ/SIZMA TAHMİN YÖNTEMİ	14
3.1. İmza Sınıfları	14
3.1.1. Düşüm Tabanlı İmzalar	14
3.1.2. Kenar Tabanlı İmzalar	15
3.1.3. Hibrit İmzalar	15
3.2. Yöntemin Tanımı	15
3.3. Yöntemin Matematiksel Modeli	15
3.3.1. Genel Tanımlar ve Notasyon	15
3.3.2. Düşümlerin Tanımlanması	17
3.3.3. Düşüm Skorlarının Hesaplanması	17
3.3.4. Kenar Skorlarının Hesaplanması	18
3.3.5. Kenar İmzalarının Çıkartılması	20
3.3.6. Çıkarılan İmzaların Nüfuz/Sızma Tahmininde Kullanılması	21

	<u>Sayfa</u>
3.3.7. Yöntemin Uyarlanması	22
4. DENEYSEL SONUÇLAR, ANALİZ ve KARŞILAŞTIRMA	26
4.1. Yöntemin Uygulanması	26
4.2. Deneysel Sonuçlar	28
4.3. Eşik Değerinin Belirlenmesi	31
4.3.1. Normal	33
4.3.2. Neptune	34
4.4. Özet Analiz Sonuçları	36
4.5. Detaylı Analiz Sonuçları	37
4.5.1. DoS (Denial of Services)	37
4.5.2. R2L (Remote to Local)	37
4.5.3. U2R (User to Root)	39
4.5.4. Probe	39
4.5.5. Normal	40
4.6. Karşılaştırma	41
5. SONUÇLAR ve ÖNERİLER	44
KAYNAKLAR	45
ÖZGEÇMİŞ	49

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler ve Açıklamalar

Kısaltmalar

Ø	: Boş Küme
ADR	: Attack Detection Rate – Saldırı Tespit Oranı
CCM	: Conjunction of Combinational Motifs
DARPA	: Defense Advanced Research Projects Agency
DoS	: Denial of Service
FN	: False Negative - Normal davranış olarak algılanan, atak kayıtlarının sayısı
FP	: False Positive - Atak olarak algılanan, normal davranış kayıtlarının sayısı
FPR	: False Positive Rate - Atak olarak algılanan, normal davranış kayıtlarının oranı (yanlış alarm oranı)
GYTE	: Gebze Yüksek Teknoloji Enstitüsü
IDS	: Intrusion Detection System – Nüfuz Tespit Sistemi
IPS	: Intrusion Prediction System – Nüfuz Tahmin Sistemi
KDD	: Knowledge Discovery in Databases
R2L	: Remote to Local
ROC	: Receiver Operating Characteristic – Alıcı Karakteristiği
TN	: True Negative - Tahmin edilen normal davranış sayısı
TP	: True Positive - Tahmin edilen atak sayısı
U2R	: User to Root

ŞEKİLLER DİZİNİ

<u>Sekil No:</u>	<u>Sayfa</u>
1.1: Ağ güvenliğine genel bakış.	3
2.1: Tıpta teşhis koymada Bayes Ağı örneği.	5
2.2: Bayes Ağı gösterimi.	6
2.3: Bayes Ağları ile istatistiksel yöntemler arası ilişkiler.	7
2.4: Birinci derece Markov Modelini gösteren Bayes Ağı.	8
2.5: İki zamanlı Dinamik Bayes Ağı modeli.	8
2.6: Dinamik Bayes Ağı modeli.	8
2.7: Saklı Markov Modeli.	9
2.8: Saklı Yarı-Markov Modeli.	10
2.9: Saklı Yarı-Markov Modeli detayları.	11
3.1: Nüfuz/Sızma akış çizgesi.	16
3.2: Nüfuz/Sızma akış çizgesinde düğüm ve kenarlar.	20
3.3: Tam bağlı zayıf nüfuz/sızma akış çizgesi.	20
4.1: İdeal ve kötü performans göstergesi olan ROC eğrileri.	32
4.2: Normal akışlar için ROC eğrisi.	34
4.3: Neptune atak akışları için ROC eğrisi.	35

TABLÖLAR DİZİNİ

<u>Tablo No:</u>	<u>Sayfa</u>
4.1: KDD Cup 99 veriseti öznitelikleri.	27
4.2: Atak tipleri analizinde kullanılan öznitelikler.	29
4.3: Normal akışlar için analiz parametreleri.	33
4.4: Neptune atak akışları için analiz parametreleri.	35
4.5: Akış çizgesi tabanlı nüfuz tahmin sistemi özet analiz sonuçları.	36
4.6: DoS grubu atak tipleri detaylı analiz sonuçları.	37
4.7: R2L grubu atak akışları birinci kısım detaylı analiz sonuçları.	38
4.8: R2L grubu atak akışları ikinci kısım detaylı analiz sonuçları.	38
4.9: U2R grubu atak akışları detaylı analiz sonuçları.	39
4.10: Probe grubu atak akışları detaylı analiz sonuçları.	40
4.11: Normal akışlar detaylı analiz sonuçları.	40
4.12: Akış çizgesi tabanlı nüfuz tahmin sistemi analiz sonuçları.	41
4.13: A neuro-genetic based short-term forecasting framework for network intrusion prediction system analiz sonuçları.	42
4.14: An adaptive automatically tuning intrusion detection system analiz sonuçları.	43

1. GİRİŞ

Bilgi Teknolojilerinde tasarlanan sistemler zamanla daha çok iletişime ihtiyaç duymaktadır; bu nedenle bilgisayar ağlarının sistem tasarımında kullanımı artmış ve buna bağlı olarak ağ güvenliği önemli bir konu haline gelmiştir. Bilgisayar ağlarında güvenlik için tehdit olarak nitelendirilen unsurlara şunlar örnek verilebilir:

- Ağdaki veri paketlerinin dinlenmesi,
- Hizmet dışı bırakma saldırıları,
- Parola saldırıları,
- Yazılımdaki zayıflıkların kötüye kullanılması,
- Zararlı kodlar,
- Virüsler,
- Truva atları,
- Solucanlar.

Bilgisayar ağlarında yukarıda listelenen tehditlere karşı ağ güvenliğini sağlamada kullanılan yöntemler özetle şöyle sınıflandırılabilir:

- Anti Virüs Yazılımlar; bilgisayar virüslerinin sisteme yerleşmesini önlemek ya da sisteme yerleşmiş virüsleri ortadan kaldırmak için kullanılan yazılımlardır.
- Güvenlik Duvarları; bilgisayarları ve bilgisayar ağlarını dışarıdan gelebilecek tehditlere karşı filtrelemeye benzer bir yöntemle koruyan sistemlerdir
- Veri Şifreleme; verilerin gönderici ile alıcı haricinde başka kişiler tarafından anlaşılmasını önlemek amacıyla belli kurallar dahilinde değiştirilmesidir.
- Nüfuz Tespit Sistemleri; bilgisayar sistemlerinde gerçekleşen olayları izleyerek güvenlik sorunu olmaları açısından değerlendirmektir. Bir sorunla karşılaştıkları zaman çeşitli alarmlar üretebilirler ya da saldırıyı önlemek için ağın işleyişinde değişiklikler yapabilirler.

Nüfuz Tespit Sistemleri üzerinde oldukça çalışılan ve geliştirilen bir sistemdir. Nüfuz Tespit Sistemlerini geliştirmeye yönelik çalışmalar ve yöntemlerle ilgili çalışmalara ileriki örnekleri vermek mümkündür.

Lee ve Stolfo sistem çağrılarının örnekleri üzerinde çalışmak ve küçük kural kümeleri ile normal veri içindeki birbiri ardına gelenleri tanımlamak için veri madenciliği yaklaşımı keşfetmişlerdir. Çalışmalarında belirlenen kuralları ihlal eden birbiri ardına gelen sistem çağrıları dizilerini izleme ve tespit amacıyla anormal olarak kabul etmişlerdir [1].

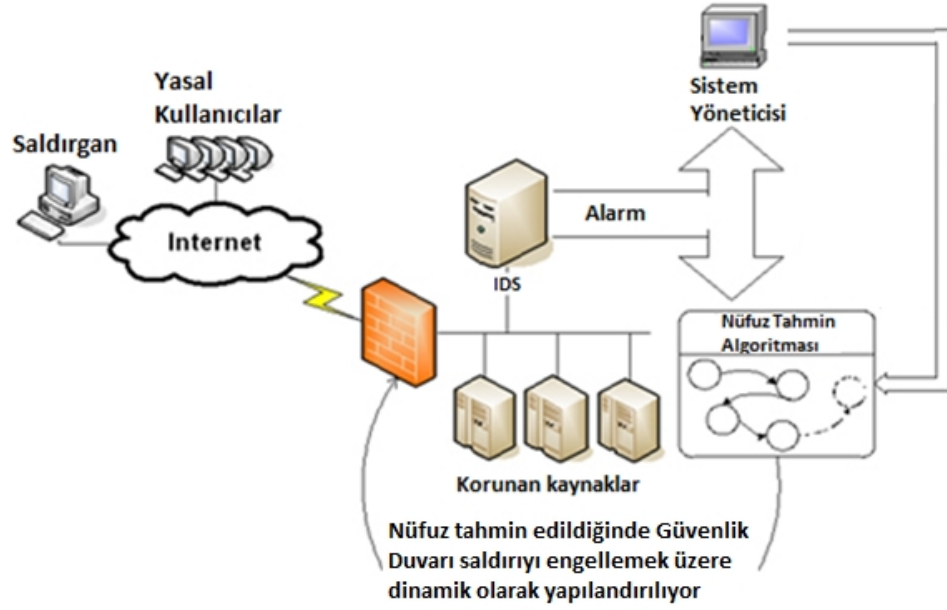
Warrander ve arkadaşları görünmeyen olayları değerlendirmek ve modellemek için Saklı Markov Modeline dayanan bir yöntem önermişlerdir [3].

Yeung ve Ding ve Lee ve Xiang, anormallik tespit etmek için bilgi teorisine dayanan ölçümler kullanmışlardır [5], [6].

Liao ve Vemuri K-en yakın komşu sınıflandırmasını kullanmışlar ve Hu ve arkadaşları, program davranışını modellemek için güçlü Destek Vektör Makinesi (Support Vector Machine) uygulamışlar ve sistem çağrısı verilerine göre her işlemi normal ve normal olmayan olmak üzere sınıflandırmışlardır [7], [8].

Sharma ve arkadaşları anormal olayları tespit etmek için kernel tabanlı benzerlik ölçümleri adapte etmişlerdir [9].

Yapılan çalışmalarda Nüfuz Tespit Sistemlerinde uygulanan yöntemlerin zararlı hareketleri tespit etmek için etkili olduğu gösterilmiştir. Ancak bu yöntemler atakları gerçekleştikten sonra tespit edebilmektedirler ve kısmen veya tamamen atakları gerçek zamanlı olarak engellemeyi zorlaştırmaktadırlar. Bu nedenle, sisteme önemli bir hasar verilmeden uygun bir cevap verilebilmesi ve amacın tipini tahmin etme özelliğine sahip bir sistem tasarlanmalıdır. Bu da Nüfuz Tahmin Sistemi tasarımı ve çalışmalarının yapılmasını gerekli hale getirmiştir. Nüfuz Tahmin Sistemlerinin bilgisayar ağlarındaki örnek kullanımı Şekil 1.1’de gösterilmiştir [38].



Şekil 1.1: Ağ güvenliğine genel bakış.

Bu çalışmada; planlanan atakları önceden farketmeye ve alarm üretmeye yönelik sistem tasarlanmak üzere bir nüfuz tahmin yöntemi önerisinde bulunulmuştur. Amaç; bilgisayar ağında gerçekleştirilen bağlantı kayıtları analiz edilerek bir bağlantı akışının atak ya da normal olduğuna karar vermek ve henüz atak akışı tamamlanmadan alarm vererek uyarıda bulunmaktadır. Böylece henüz sisteme zarar verilmeden atağın önlemesine yönelik önlemler alınabilir.

Önerilen yöntem polimorfik solucanların tespiti için tasarlanan CCM (Conjunction of Combinational Motifs) yönteminin nüfuz tahmini için adapte edilmesi ile oluşturulmuştur [30]. Bir öğrenme verisetindeki atak akışlarındaki bağlantı kayıtları analiz edilerek atak akışlarında ortak görülen bağlantı kayıtlarından imzalar çıkartılmaktadır. Daha sonra çıkartılan imzalar gerçek zamanlı olarak bağlantı kayıtlarının analiz edilmesinde ve bu analiz sonucunda atak akışına ya da normal akışa ait olduğuna karar verilmesinde kullanılmaktadır.

Önerilen yöntemin uygulaması geliştirilmiş ve KDD Cup 99 verisetinde test edilmiştir. Öncelikle; KDD Cup 99 öğrenme veriseti kullanılarak imzalar çıkartılmıştır. Bunun için her atak tipine ait belirleyici öznitelikler belirlenmiş ve bu özniteliklerle yapılan değerlendirme ile atak akışlarında görülen ortak bağlantı katarları çıkartılmıştır. Çıkartılan bu ortak bağlantı katarlarının atak akışlarında ve normal akışlarda görülme oranlarından hesaplanan skorlar ve akış çizgeleri kullanılarak imzalar elde edilmiştir. Sonraki aşamada; öğrenme veriseti analizi sonucu

elde edilen imzalar test veriseti bağlantı kayıtlarının gerçek zamanlı olarak değerlendirilmesinde ve atak ya da normal akışlara ait olduğunun karar verilmesinde kullanılmıştır.

Yöntemin analiz sonuçları benzer şekilde nüfuz tahmin yöntemleri öneren ve önerdikleri yöntemleri KDD Cup 99 verisetinde test ederek analiz sonuçlarını sunan iki çalışma ile karşılaştırmalı olarak sunulmuştur [34], [35].

Tezin sonraki bölümleri aşağıdaki şekilde organize edilmiştir. İkinci bölümde Nüfuz Tahmin Sistemleri ile ilgili yapılmış olan çalışmalara yer verilmiştir. Önerilen yöntem üçüncü bölümde anlatılmıştır. Dördüncü bölümde, deneysel sonuçlar ve karşılaştırmalar verilmiştir. Sonuç ve öneriler son bölümde açıklanmıştır.

2. NÜFUZ TAHMİN SİSTEMLERİ ve İLGİLİ ÇALIŞMALAR

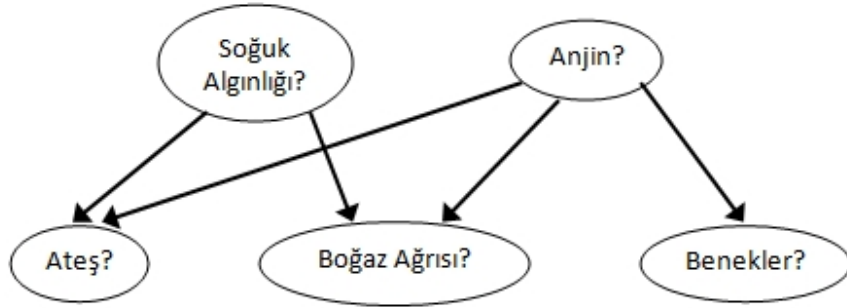
Nüfuz Tahmin Sistemleri teknikleri iki ana başlık altında gruplanabilir:

- İstatistiksel Yöntemler,
- Veri Madenciliği Yöntemleri.

Dökümanın ilerleyen kısımlarında bu teknikler ayrıntılı bir şekilde anlatılacaktır.

2.1. İstatiksel Yöntemler

Bayes Ağları; problemlerde durumlar ve durumlar arası geçişlerin koşullu olasılık ile gösterildiği grafiksel modellerdir. Gerçek hayattaki bir uygulaması olarak tıp biliminde teşhis koyma aşağıdaki gibi bir örnekle gösterilebilir [15].



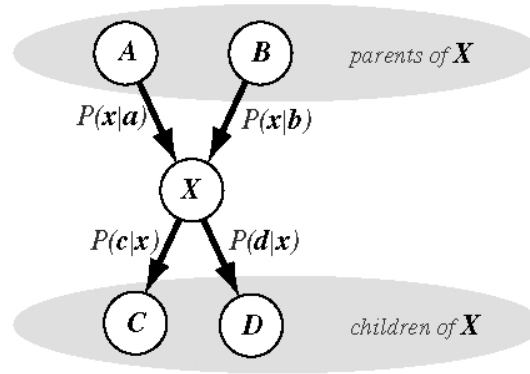
Şekil 2.1: Tıpta teşhis koymada Bayes Ağı örneği.

Bu örnekte raporlanan semptomları değerlendirerek hastaya teşhis koymaya çalışılmaktadır. Bu probleme yaklaşımın doğal bir yolu, olası semptomlarına göre hastalığın etkilerini istatistiksel olarak modellemektir. Sonraki aşamada da gerçek bir hastada görülen semptomlar verildiği zaman en muhtemel hastalıklar için bir dağılım çıkarmaktır. Şekil 2.1’de iki hastalık ve üç semptom içeren bir modelin Bayes ağı görülmektedir. Bayes ağının her bir düğümü rastgele bir değişkene karşılık gelmektedir ve oklar da değişkenler arasındaki istatistiksel ilişki hakkında bilgi vermektedir. Belli uygulamalarda değişkenler arasında çizilen oklardan rastgele

yorumlar yapmak normaldir. Örnek olarak, ateş ve boğaz ağrısı görülmesi direk veya dolaylı olarak soğuk algınlığının sonucu olarak yorumlanabilir [15].

Bayes Ağları; Bilgisayar bilimlerinde veri modelleme ve durum geçişlerini ifade etmede kullanılan yöntemlerden biridir. Yönlü dönuşsüz ağlar (directed acyclic network) ile gösterilir ve her düğüm ayrı bir değişkeni ifade eder.

Formal gösteriminde değişkenler arasındaki koşullu olasılık ilişkilerini aşağıdaki gibi bir grafik yapısı üzerinde gösterir [16].



Şekil 2.2: Bayes Ağı gösterimi.

Yukarıdaki gösterimde; X , A , B , C ve D durumları ifade eder. A ve B , X 'in ebeynleri; C ve D ise X 'in çocuklarıdır. $P(x|b)$, B durumundan X durumuna geçiş olasılığı göstermektedir [16].

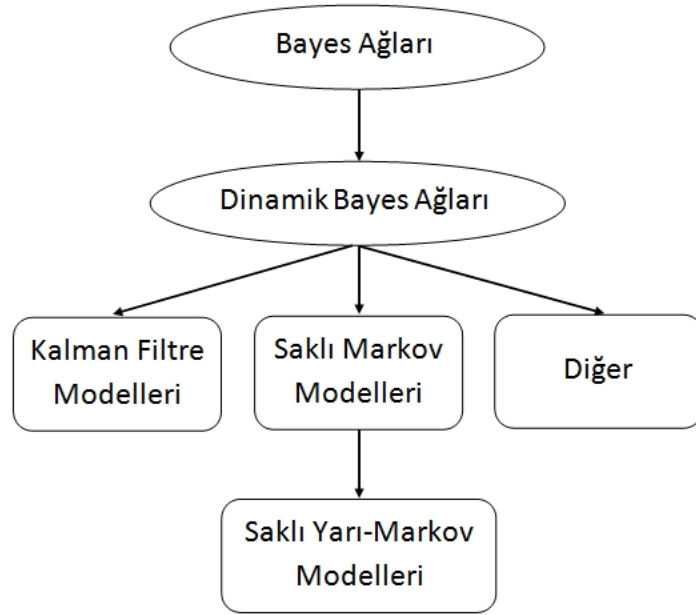
Bayes olasılık hesaplama yöntemine göre $x_1, \dots, x_n$ eleman kümesinin birleşik olasılığı zincir kuralına göre; her değişken ile o değişkeni etkileyen ondan önceki diğer değişkenlerin koşullu olasılıkları çarpımına denk gelecektir.

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1}) \quad (2.1)$$

Markov özelliği; mevcut durum verildiğinde, gelecek durumların geçmiş durumlardan bağımsız olması anlamına gelir. Bayes ağları Markov özelliği göstermektedir. Bu durum her durumun meydana gelme olasılığının sadece ebeynlerinden etkilenmesi anlamına gelir. Bu bağlamda; Bayes ağındaki koşullu bağımsızlık ilişkileri değerlendirildiğinde x_i değişkenini ebeveynleri etkiler ve $Pa(x_i)$ ile ifade edilir.

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | Pa(x_i)) \quad (2.2)$$

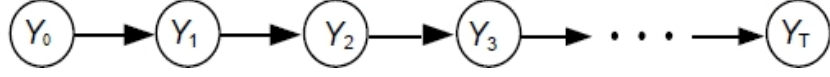
Bayes Ağları; istatistiksel yöntemlerin temelini oluşturmaktadır. Bayes Ağları ile Nüfuz Tahmin Sistemleri tasarlanmasında kullanılan diğer istatistiksel yöntemler arasındaki ilişki Şekil 2.3’de gösterilmiştir.



Şekil 2.3: Bayes Ağları ile istatistiksel yöntemler arası ilişkiler.

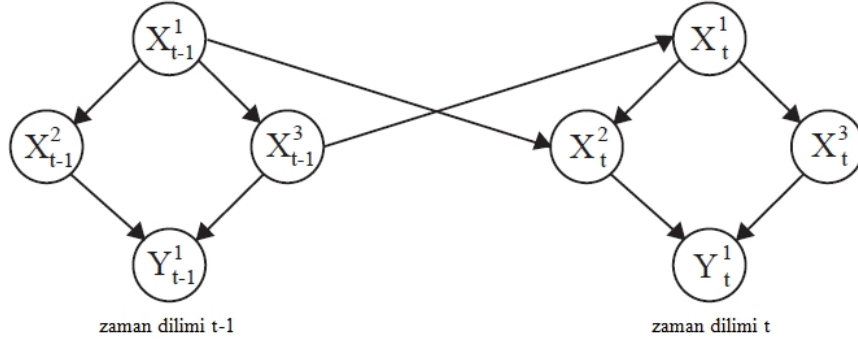
2.1.1. Dinamik Bayes Ağları

Zaman serileri modellemede, farklı zamanlarda belirli değişkenlerin değerleri gözlemlenir. Bir olayın gelecekte başka bir olaya sebep olacağı varsayılır, fakat tersi geçerli değildir. Bu kural, zaman serilerinde Bayes Ağlarının tasarımını basitleştirir. Yönlü oklar zamanda ilerlemeyi gösterir. Her bir değişkene zaman indeksi atanarak (t) sıralı veriler için $\{Y_1, \dots, Y_T\}$ her bir değişkenin sadece bir önceki değişkenden direk etkilendiği Birinci Derece Markov modeli oluşturulur [17]. Birinci Derece Markov modeli aşağıdaki gibi gösterilebilir [17].



Şekil 2.4: Birinci derece Markov Modelini gösteren Bayes Ağı.

Dinamik Bayes Ağları, zamana bağlı olarak değişen değişkenlerin oluşturduğu Bayes Ağlarıdır. Şekil 2.5’de $(t-1)$ anındaki Bayes Ağının düğümlerinin değeri zamanla değişiklik göstererek (t) anındaki Bayes Ağına dönüşmüştür [15].

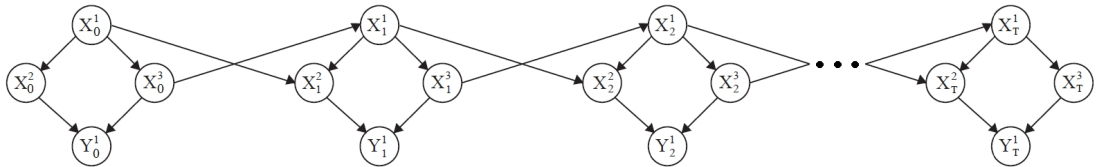


Şekil 2.5: İki zamanlı Dinamik Bayes Ağı modeli.

$(t-1)$ anında X değişkeninin değeri bilinirse; (t) anında X değişkeninin değeri olasılığı; (t) anındaki her X değişkeninin ebevenyleri bilindiği durumdaki olasılıklarının çarpımıdır.

$$P(X_t | X_{t-1}) = \prod_{i=1}^n P(X_t^i | Pa(X_t^i)) \quad (2.3)$$

Eğer işlem, uzun zaman devam ederse grafik aşağıdaki gibi olacaktır [15].



Şekil 2.6: Dinamik Bayes Ağı modeli.

Dinamik Bayes Ağlarında, T anındaki grafiğin birleşik olasılık dağılımı aşağıdaki gibi hesaplanır [15].

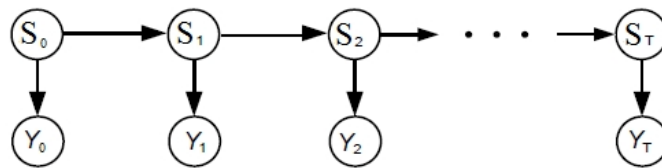
$$P(X_{0:T}) = P(X_0) \prod_{t=1}^T P(X_t | X_{t-1}) \quad (2.4)$$

Li Feng ve arkadaşları yaptıkları çalışmada; bir dizi sistem çağrısının amacının normal erişim ya da saldırı amaçlı olduğunu anlamaya yönelik bir Dinamik Bayes Ağ yaklaşımı geliştirmişlerdir. Bu şekilde nüfuz tehlikesini önceden tahmin ederek uyarılar verilmesini ya da saldırıyı engellemek için önlemler alınmasını sağlamaya çalışmaktadırlar [21], [22].

2.1.2. Saklı Markov Modeli

Mühendislik ve bilimde; birçok uygulama arka arkaya gelen veriler üzerinde çalışmayı gerektirmektedir. Bu uygulamalara örnek olarak konuşma tanıma, birden fazla hedefi izleme, gen analizi, tıbbi izleme, otoyol gözetimi verilebilir. Bu uygulamaların her birinde gözlemlenen veriler, gözlemlenen durumun altında yatan saklı süreçleri belirlemek için kullanılmaktadır. Bu problem durum-uzay model (state-space model) yaklaşımı kullanılarak modellenen filtreleme problemidir. Bir durum-uzay modeli, her bir zaman adımında gözlemlenen durumun altında yatan saklı süreçlerin parametrelerine göre değişen saklı durum vektörüdür.

Durum-uzay modelinin iki bileşeni vardır. Dinamik model; durumun zamanla nasıl değiştiğini açıklar. Gözlemlenme modeli ise; gözlemlerin saklı durumdan nasıl ortaya çıktığını açıklamaktadır [15].



Şekil 2.7: Saklı Markov Modeli.

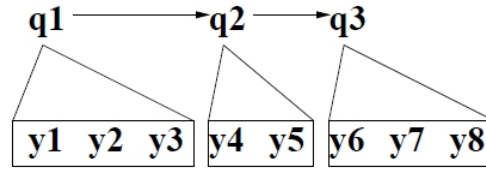
Şekil 2.7’de $Y_{0:T}$, gözlemleri; $S_{0:T}$, saklı durumları ifade eder [15]. Bu grafikte T anındaki ardışık bir durum ve gözlem dizisinin birleşik olasılığı aşağıdaki gibidir [18].

$$P(S_{0:T}, Y_{0:T}) = P(S_0)P(Y_0|S_0) \prod_{t=1}^T P(S_t|S_{t-1})P(Y_t|S_t) \quad (2.5)$$

Kjetil Haslum ve arkadaşları yaptıkları çalışmalarda dağıtık bir Nüfuz Tahmin ve Önleme sistemi tasarlamışlardır. Bu sistemde ağ üzerindeki değişik noktalardaki Nüfuz Tespit Sistemlerinden edinilen bilgileri merkezi bir sistemde Saklı Markov Modeli kullanarak değerlendirerek nüfuzu tahmin etmektedirler [23]. Daha sonra bu çalışmalarına risk değerlendirme özelliği kazandırarak daha akıllı hale getirmişlerdir ve sistemlerinde performans iyileştirmelerinde bulunmuşlardır [24], [25].

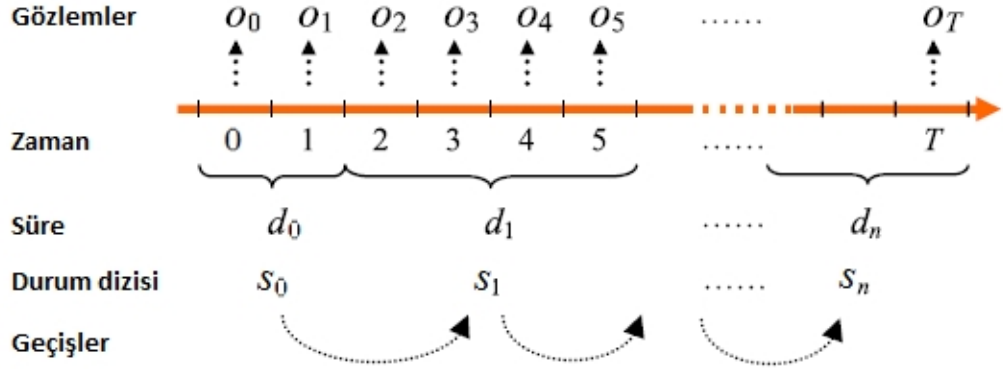
2.1.3. Saklı Yarı-Markov Modeli

Saklı Yarı-Markov Modeli, Saklı Markov Modeli ile aynı yapıdadır. Ancak durumlar Saklı Markov Modelinden farklı olarak Markov özelliği göstermemektedir. Her durumun ortaya çıkardığı sayısı sabit olmayan bir dizi ardışık gözlem mevcuttur. Saklı Yarı-Markov Modeli Şekil 2.8'deki grafik ile gösterilebilir [15].



Şekil 2.8: Saklı Yarı-Markov Modeli.

Saklı Yarı-Markov Modeli, geometrik dağılım göstermeyen bir modele uygulandığında Saklı Markov Modelinden daha isabetlidir.



Şekil 2.9: Saklı Yarı-Markov Modeli detayları.

Şekil 2.9’da zaman eksenini boyunca gözlemler ve durumları incelediğimizde; durumların farklı süreler aldığı ve her durum için ortaya çıkan gözlem dizisinin farklı uzunluklarda olabildiği görülebiliyor [19]. s_i durumu d_i kadar süre alıyor ve bu durumda O_1 ve O_2 gözlemleri oluşuyor.

Bu grafikte T anındaki ardışık durumlar ve her durumda ortaya çıkan değişik sayıdaki gözlemler dizisinin birleşik olasılığı aşağıdaki gibidir [20].

$$P(O_{0:T}, s_{0:n}) = \left[\prod_{i=1}^n P(s_i | s_{i-1}) \right] \left[\prod_{i=0}^n P(O_{d_i} | s_i) \right] \left[\prod_{i=0}^n P(d_i | s_i) \right] \quad (2.6)$$

Zhang Zhengdao ve arkadaşları ise çalışmalarında Saklı Yarı-Markov Model kullanarak sistem çağrılarını değerlendirmişler ve ağ üzerinde nüfuz tahmininde bulunmuşlardır [26].

2.2. Veri Madenciliği Yöntemleri

Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır [3]. Çok fazla veri içinden “öz” bilgiyi çıkarmak veya büyük veritabanlarında bulunan verilerden gizli kalmış örüntüleri çıkarmak olarak da tanımlanabilir. Veri madenciliği, satış ve pazarlama, bankacılık, sigortacılık, finans, telekomünikasyon gibi sektörlerde ticari kazanç sağlamak amacıyla kullanılmaktadır.

Bunun yanında tıp, genetik, antropoloji ve enformatik gibi birçok bilim dalında bilimsel amaçlarla kullanılmaktadır.

Veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases - KDD) adı verilen ana süreç temelde 5 alt süreçten oluşur, bunlardan birisi de veri madenciliğidir. Bu süreçler şöyledir [2]:

- Problemin tanımlanması
- Verilerin seçilmesi ve toplanması
- Verilerin hazırlanması
- Veri Madenciliği
- Yorumlama ve Değerlendirme

Veri madenciliği çalışmalarında kullanılan teknikler 3 temel grupta toplanır:

- Sınıflama
- Kümeleme
- Birliktelik Kuralları ve İlişki Analizi

2.2.1. Sınıflandırma Yöntemi

Sınıflama, veri sınıflarını ve kavramlarını tanımlayan ve ayırt eden bir modelin (ya da fonksiyonun) bulunması işlemidir. Yeni bir nesnenin niteliklerinin incelenmesi ve bu nesnenin önceden tanımlanmış bir sınıfa atanmasıdır. Her bir sınıfın özellikleri önceden belirlenmiş olmalıdır. Sınıf özelliklerinin belirlenmesi bir grup çalışma verisinin analizine dayanır (mesela, sınıf etiketleri belli olan nesneler). Bu model kullanılarak sınıf etiketleri belli olmayan nesneler sınıflandırılabilir [2].

Sınıflandırma yöntemleri şunlardır:

- Karar ağaçları
- Yapay sinir ağları
- K-en yakın komşu algoritması
- Bayes sınıflandırma
- Regresyon analizi

Liao Cheng-Bin çalışmasında anahtarlardan geçen ağ paketlerinin özelliklerini çıkartmış ve bunları önceden belirlenmiş bir ağ paketleri özellikleri kümesinde veri madenciliği sınıflandırma yöntemini kullanmıştır. Bu şekilde oluşabilecek saldırıları tahmin etmiştir [27].

2.2.2. Kümeleme

Kümeleme analizi, sınıflandırmada olduğu gibi sahip olunan verileri gruplara ayırma işlemidir. Sınıflandırma işleminde, sınıflar önceden belirli iken kümelemede sınıflar önceden belirli değildir. Verilerin hangi kümelere ve kaç değişik gruba ayrılacağı eldeki verilerin birbirlerine olan uzaklığa göre belirlenir [3].

2.2.3. Birliktelik Kuralı ve İlişki Analizi

Birliktelik kuralı $X \rightarrow Y$ şeklinde ifade edilir. Bu ifadede X nesneler kümesi, Y ise tek bir nesnedir. Birliktelik kuralı, veriler arasındaki gizli ilişkileri, bağıntıları ve eş zamanlı gerçekleşen olayları ortaya çıkarmayı amaçlamaktadır. Sık tekrarlanan öğeler bulunur ve bu öğelerden güçlü birliktelik kuralları oluşturulur. Bu yöntemin kullanıldığı en yaygın alan alışveriş sepeti analizidir. İlaç sektörü ve sağlık sigortası dolandırıcılığı gibi birçok alanda uygulanmaktadır. Bu alandaki en popüler algoritmalar, Apriori, AprioriTid, SETM ve AIS algoritmalarıdır [3].

Zhi-tang Li ve arkadaşları çalışmalarında veri madenciliği tekniğini kullanarak gerçekleşen saldırıları derecelendirirken ardından nüfuz gelme ihtimalini hesaplıyor [28]. Bir başka çalışmalarında ise yine veri madenciliği tekniğini kullanarak gerçekleşmiş olan saldırılardan bir saldırı grafiği oluşturuyorlar ve bu grafiğe göre saldırıyı gerçekleştirenin saldırı planını hesaplamaya ve bir sonraki saldırısını tahmin etmeye çalışıyorlar [29].

3. AKIŞ ÇİZGESİ TABANLI NÜFUZ/SIZMA TAHMİN YÖNTEMİ

Bu çalışmada, daha önce polimorfik solucan tespiti için önerilen CCM (Conjunction of Combinational Motifs) metodunun uyarlanması ile bilgisayar sistemlerine sızma tahmininde kullanılacak yeni bir yöntem tasarlanmıştır [30].

CCM, yönlü kenarlar ve bağımsız düğümlerin birleşimine dayanan yeni bir hibrit polimorfik solucan tespit imza şeması ve graf tabanlı hibrit imza sınıflarının uygulamasıdır.

Düğüm çıkartılması, önerilen CCM yönteminin ilk aşamasıdır. Önerilen düğüm skoru hesaplama yöntemini kullanarak düğümler, güçlü ve zayıf olarak ayrılmaktadır. Güçlü düğümler, CCM imza kümesinin doğrudan elemanı olarak kullanılmaktadır.

Güçlü düğümler, keşfedilmiş polimorfik solucan akış çizgesinden çıkartılır. Kalan akış çizgesi, zayıf düğümleri içerir ve bu zayıf düğümler güçlü kenarları bulmak için analiz edilir.

Sonuç olarak, elde edilen imza kümesi güçlü düğümlerin ve güçlü yönlü kenarların birleşimi olarak tanımlanır ve polimorfik solucanların tespitinde kullanılır [30].

3.1. İmza Sınıfları

İçerik tabanlı imza yapıları; Düğüm Tabanlı İmzalar, Kenar Tabanlı İmzalar ve Hibrit İmzalar olarak sınıflandırılabilir [30].

3.1.1. Düğüm Tabanlı İmzalar

Düğüm Tabanlı İmzalar, Nüfuz/Sızma tahmini için düğümlerden faydalanır. Tahmin yönteminde keşfedilmiş düğümlerin tamamı kullanılabilceği gibi keşfedilmiş düğümlerin bir alt kümesi de kullanılabilir [30].

3.1.2. Kenar Tabanlı İmzalar

Kenar Tabanlı İmzalar, bir nüfuz/sızma tahmin mekanizması tanımlamak için kenarlardan faydalanır. Yöntem, kenar kümesi E 'nin kenarlarının tamamını kullanılabileceği gibi kenarlarının bir alt kümesini de kullanılabilir [30].

3.1.3. Hibrit İmzalar

Hibrit İmzalar, bir nüfuz/sızma tahmin mekanizması tanımlamak için düğümlerden ve kenarlardan faydalanır [30].

3.2. Yöntemin Tanımı

Bu çalışmada, daha önce polimorfik solucan tespiti için önerilen CCM (Conjunction of Combinational Motifs) metodunun uyarlanması ile bilgisayar sistemlerine nüfuz/sızma tahmininde kullanılacak yeni bir hibrit imza oluşturma modeli tasarlanmıştır [30].

3.3. Yöntemin Matematiksel Modeli

3.3.1. Genel Tanımlar ve Notasyon

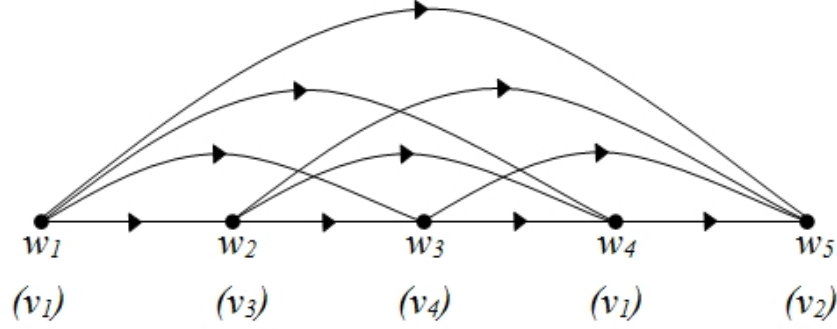
Düğüm kümesi V : V , $n \geq 1$ olmak üzere n düğümden oluşan ve her düğümün şüpheli akış havuzundaki nüfuz/sızma kopyalarının belirli bir bölümünde yer alan bir ortak karakter katarını temsil ettiği bir küme olsun. $V = \{v_1, v_2, \dots, v_n\}$

Kenar kümesi E : E , her kenarın $V \times V$ kümesindeki sıralı bir düğüm ikilisini temsil ettiği, n^2 kenardan oluşan bir küme olsun. $\forall i, j$ için $e_{ij} = (v_i, v_j)$ ve $1 \leq i, j \leq n$ iken, $E = V \times V = \{e_{ij}\}$.

Akış çizgesi X : X , keşfedilmiş düğümler $v_i \in V$ cinsinden bir akış çizgesi olsun. X ; $1 \leq i \leq m$ ve $x_i \in V$ olmak üzere m düğümden oluşur. X te hiçbir $v \in V$ düğümü

yoksa $X = \emptyset$ dir. X , bir komşu düğümler listesi, komşu kenarlar kümesi veya yönlü çizge olarak ifade edilebilir.

Nüfuz/Sızma Akış çizgesi W : W , keşfedilmiş düğümler $v_i \in V$ cinsinden bir nüfuz/sızma çizgesi olsun. W ; $1 \leq n \leq p$ olmak üzere p düğümden (w_i) oluşur ve $1 \leq i \leq p$ olmak üzere $\forall i$ için $w_i \in V$ 'dir. W , bir komşu düğümler listesi, komşu kenarlar kümesi veya yönlü çizge olarak ifade edilebilir.



Şekil 3.1: Nüfuz/Sızma akış çizgesi.

Düğüm skor fonksiyonu $f_{V_{score}} : v_i \rightarrow (V_{score})_i \in R : f_{V_{score}}$, düğüm kümesi V üzerinde her $v_i \in V$ için düğüm skorunu hesaplayan bir fonksiyon olsun ve V_{score} düğüm skorlarının oluşturduğu küme olsun.

Kenar skor fonksiyonu $f_{E_{score}} : e_{ij} \rightarrow (E_{score})_{ij} \in R : f_{E_{score}}$, kenar kümesi E üzerinde her $e_{ij} \in E$ için kenar skorunu hesaplayan bir fonksiyon olsun ve E_{score} kenar skorlarının oluşturduğu küme olsun.

Akış skor fonksiyonu $f_{X_{score}} : X \rightarrow (X_{score}) \in R : f_{X_{score}}$, akış çizgesi X üzerinde akış skorunu hesaplayan bir fonksiyon olsun.

CCM ile imzaların üretilmesinde iki çeşit havuz kullanılmaktadır. Bunları Normal Akış Havuzu ve Şüpheli Akış Havuzu olarak adlandırılır.

Normal Akış Havuzu, nüfuz/sızma içermeyen trafik akışları için tanımlanmaktadır. Nüfuz/Sızma içeren akış havuzlar ise Şüpheli Akış Havuzu olarak tanımlanmaktadır.

3.3.2. D ğ mlerin Tanımlanması

 nerilen n fuz/sızma tahmini y nteminde d ğ mler, atak akıřlarında g r len en uzun ortak baėlantı kayıtları katarıdır.

D ğ mler  ıkartılırken bilgisayar aėındaki baėlantı kayıtları analiz edilir. Atak akıř havuzundaki n atak akıřı  rneėinden en az K tanesinde yer alan, en az α en uzunluėundaki atomik alt baėlantı kayıt katarlarından her birisi bir d ğ m olarak belirlenir.

3.3.3. D ğ m Skorlarının Hesaplanması

D ğ m k mesi V ve N fuz/Sızma akıř k mesi I , oluřturulduktan sonra her d ğ m $v_i \in V$ i in ř pheli veya normal akıřta g r nme olasılıklarına g re skorlar hesaplanır.

Bu olasılıklar Bayes' kuralı kullanılarak hesaplanır. ' $L(v_i) = true$ ' notasyonu d ğ m n sızma akıřında olma durumunun; ' $L(v_i) = false$ ' notasyonu ise d ğ m n normal akıřta olma durumunun g sterimi i in kullanılmaktadır.

Bir d ğ m v_i 'nin sızma akıřında olma olasılıėı $P[L(v_i) = true|v_i]$, normal akıřta olma olasılıėı $P[L(v_i) = false|v_i]$ olur.

Bařlangı ta bir d ğ m n sızma akıřı ve normal akıř elemanı olması olasılıkları eřit kabul edilmektedir, $P[L(v_i) = true|v_i] = P[L(v_i) = false|v_i]$

D ğ m k mesi V ve Akıř k mesi X analiz edilerek her d ğ m v_i i in sızma akıřı ve normal akıřta g r nme olasılıkları oranı hesaplanır. Bu hesaplama ařaėıdaki gibi olur.

$$f'_{V_{score}} = \frac{P(v_i|L(v_i) = true)}{P(v_i|L(v_i) = false)} \quad (3.1)$$

Elde edilen sayılar çok büyük olacağından bu oranın logaritmik değeri kullanılacaktır. Bu nedenle düğüm skor fonksiyonu $f_{V_{score}}$ aşağıdaki denklem ile tanımlanmaktadır.

$$f_{V_{score}} : v_i \rightarrow (V_{score})_i \in R,$$

$$f_{V_{score}}(v_i) = (V_{score})_i = \log \left(\frac{P(v_i | L(v_i) = true)}{P(v_i | L(v_i) = false)} \right) \quad (3.2)$$

Sızma akışında daha sık, normal akışta daha az görünen düğümler iki akışta da ortak olanlara göre daha yüksek skorlara sahiptirler. Bu düğümler güçlü düğümler, diğerleri ise zayıf düğümler olarak tanımlanırlar. Güçlü düğümlerin içerisinde görüldükleri akışların sızma akışı olması ihtimali daha fazladır. Bu nedenle, skorları hesaplanan düğümler, Güçlü ve Zayıf olarak iki kümeye ayrılmaktadır.

Güçlü düğümler V_{strong} kümesini, Zayıf düğümler ise V_{weak} kümesini oluşturmaktadır. $P_{cluster_V}$ prosedürü, $f_{V_{score}}$ fonksiyonunu kullanarak kenar skorlarının hesaplanmasını ve k-means clustering algoritmasını kullanarak V_{strong} ve V_{weak} kümelerin oluşturulmasını sağlar.

3.3.4. Kenar Skorlarının Hesaplanması

CCM imza kümesi güçlü kenarların ve güçlü yönlü kenarların kümesi olarak tanımlanır. V_{strong} kümesinin çıkartılmasından sonra bir sonraki adım $E_{signature_edges}$ yönlü kenar kümesinin belirlenmesidir. $E_{signature_edges}$ nüfuz/sızma akışı grafi I 'daki güçlü kenarlarından oluşur.

CCM imza kümesi güçlü kenarların ve güçlü yönlü kenarların kümesi olarak tanımlanır. V_{strong} kümesinin çıkartılmasından sonra bir sonraki adım $E_{signature_edges}$ yönlü kenar kümesinin belirlenmesidir. $E_{signature_edges}$ nüfuz/sızma akışı grafi W 'daki güçlü kenarlarından oluşur. Bir zayıf ve bir güçlü düğüm veya iki güçlü düğüm içeren kenarlar genellikle güçlüdür çünkü güçlü düğümün kendisi şüpheli akış havuzunda daha sık görülürken, normal akış havuzunda daha seyrek görülmektedir.

Bu tür güçlü kenarlar göz ardı edilir ve güçlü düğümler imza kümesinde tek başına kullanılır. İmza kümesindeki güçlü kenarlar zayıf düğümlerden oluşurlar.

İmza kümesindeki güçlü kenarlar yalnızca V_{weak} kümesindeki zayıf düğümleri içerirler. V_{strong} kümesindeki güçlü düğümler nüfuz/sızma akışı grafi W 'dan silinirler ve akış grafi yönlü ve devirsiz hale getirilir bu yeni akış grafına W_{weak} adı verilir. Bu grafda her düğüm W_i tüm diğer düğümlere W_j , $j > i$ olacak şekilde bağlıdır. W_{weak} , $E_{signature_edges}$ kümesini çıkarmak üzere incelenir.

$E_{signature_edges}$, W_{weak} 'in kenar kümesinin alt kümesidir ve CCM imza kümesinin güçlü kenarlarını içerir. Akış havuzlarındaki yönlü kenar olasılıklarının oranı olarak W_{weak} 'in her bir kenarı için kenar skorları hesaplanır.

Kenar skorları; W_{weak} kümesi elemanlarından; içerisinde bulundukları akışların sızma akışı olma ihtimali daha fazla olanları belirlemek için hesaplanır. Hesaplanan skorlara göre sızma akışlarında daha sık görülenler Güçlü kenarlar ve sızma akışlarında çok fazla görülmeyenler Zayıf kenarlardır.

$P_{cluster_E}$ prosedürü, f_{Escore} fonksiyonunu kullanarak kenar skorlarını hesaplamak ve k-means algoritmasını kullanarak E_{strong} ve E_{weak} kümelerini oluşturmak için tanımlanmıştır.

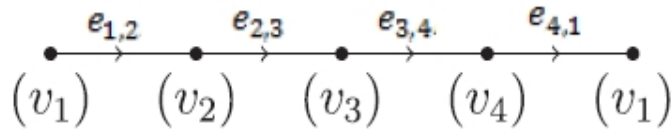
$$f_{Escore} : e_{ij} \rightarrow (E_{score})_{ij} \in R, \quad (3.3)$$

$$f_{Escore}(e_{ij}) = (E_{score})_{ij} = \log \left(\frac{P(e_{ij} | L(e_{ij}) = true)}{P(e_{ij} | L(e_{ij}) = false)} \right)$$

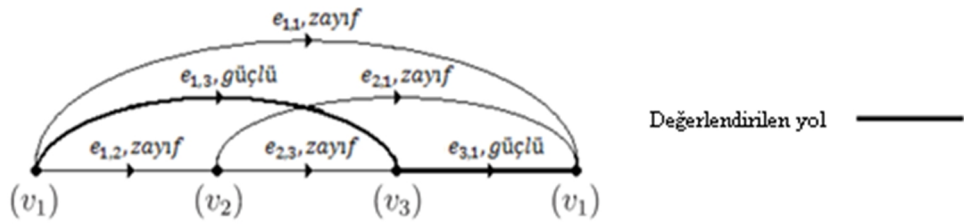
W_{weak} akışında e adet kenar olduğunu varsayarsak ve akış havuzlarındaki akışların ortalama boyutunu L byte olarak varsayarsak, kenar skor hesaplama ve kenar kümelerini belirlemek $O(e \times L)$ 'de tamamlanmaktadır. En iyi durumda tüm bulunan düğümler güçlüdür ve güçlü kenar aramaya ihtiyaç yoktur. “ e ” sifıra eşit ise çalışma zamanı karmaşıklığı da sifıra inecektir. En kötü durumda, bulunan tüm n adet kenar zayıf olacaktır ve tüm $e = n^2$ yönlü kenar adayları nüfuz/sızma akışı grafında görünecektir. Bu durumda çalışma zamanı karmaşıklığı $O(n^2 \times L)$ olacaktır.

3.3.5. Kenar İmzalarının Çıkartılması

W_{weak} 'in başlangıç düğümünden son düğümüne kadar olan maksimum toplam skorlu bulunur. $P_{extract_signature_edges}$ prosedürü ile zayıf kenarlar W_{weak} 'in en büyük yolundan çıkarılır ve kalan güçlü kenarlar $E_{signature_edges}$ kümesi olarak kaydedilerek imza kümesinin bir parçası olur. W_{weak} 'de p adet zayıf düğüm olduğu varsayılarak, $E_{signature_edges}$ kümesini oluşturmak $O(p^2)$ 'de tamamlanacaktır. Aşağıdaki şekilde gösterilen W nüfuz/sızma akışı grafında $W_{i,j} \in W$, W grafının $v_j \in V$ olan i 'inci düğümünü temsil etmektedir. w_i notasyonu W 'nin i 'inci düğümünü temsil etmek için kullanılmaktadır. $n = 4$ düğüm bulunmuştur ve nüfuz/sızma akışı grafı v_1, v_2, v_3, v_4, v_1 olarak çıkarılmıştır. Şekildeki nüfuz/sızma akışı grafı, nüfuz/sızma'nın şablonunu göstermektedir. CCM imzası, güçlü düğümler ve güçlü kenarların birleşimidir.



Şekil 3.2: Nüfuz/Sızma akış çizgesinde düğüm ve kenarlar.



Şekil 3.3: Tam bağlı zayıf nüfuz/sızma akış çizgesi.

Nüfuz/sızma şablonunda güçlü kenarları çıkarmak için sadece zayıf düğümler kullanılacaktır. Nüfuz/sızma şablonunun çıkarılmasından sonraki adım güçlü düğümlerin şablondan çıkarılmasıdır. Çünkü güçlü düğümler CCM imza kümesinde direk kullanılacaktır ve güçlü kenarlara bir katkısı yoktur. Geri kalan grafa ilk düğümden başlayarak son düğümüne kadar en fazla maliyetli yol aranır. Bu en maliyetli

yolda kenar maliyetleri ve kenar skorları hesaplanır ve gerekli kümeler $P_{cluster_V}$ ve $P_{cluster_E}$ prosedürleri ile belirlenir.

Yönlü devirsiz graf W_{weak} , kenar skorları ve kenarların güçleri ile birlikte “Tam Bağlı Zayıf Atak Akış Çizgesi”nde gösterilmiştir. Kenar skorları 3.3.4 bölümünde anlatılan kenar hesaplama fonksiyonu ile hesaplanmıştır. W_{weak} ’in başlangıç düğümünden bitiş düğümüne kadar giden en büyük skorlu yolu bulmak için özyinelemeli bir hesaplama algoritması kullanılmıştır. Bulunan yol üzerindeki güçlü kenarlar ilgili nüfuz/sızma’nın CCM imza kümesinde kullanılmıştır.

$P_{extract_signature_edges}$ prosedürünün her i . iterasyonunda, $(W_{weak})_1$ ’den $(W_{weak})_{i+1}$ ’e maksimum skorlu yol bulunur. W_{weak} ’de $N = 4$ düğüm olduğu için $(W_{weak})_1$ ’den $(W_{weak})_4$ ’e maksimum skorlu yol, $N - 1 = 3$ iterasyonda bulunur. Maksimal yolun kenarları, $i = N$ ’den başlayarak W_{weak} ’in ilk düğümüne ulaşılan kadar $(W_{weak})_i \in W_{weak}$ düğümlerinin ters yönde taranması ile bulunur.

3.3.6. Çıkarılan İmzaların Nüfuz/Sızma Tahmininde Kullanılması

Nüfuz/Sızma tahmininde bulunmak için güçlü düğümler ve güçlü kenarlar kullanılarak çıkarılan hibrit imzalar kümesi kullanılır.

Bilgisayar ağlarında görülen bağlantı kayıtları tek tek değerlendirilerek analiz edilmekte olan akış kümesi $Akis$ olsun,

t anında yapılan bir bağlantı kaydı $x \in Akis$ olsun, öncelikle bu x bağlantı kaydı için analiz edilen atak tipi için Tablo 4.2’deki ilgili öznitelikler kullanılarak düğüm sınıflandırılması yapılır.

Düğüm sınıflandırması şöyle yapılır;

Analiz edilmekte olan atak tipi için önceden çıkartılan düğüm kümesi $V_{AtakTipi}$,

bu kümenin eleman sayısını veren fonksiyon $s(V_{AtakTipi})$ olsun,

$V_{AtakTipi}$ kümesindeki her bir düğüm v_i için düğümü oluşturan bağlantı kaydı sayısını belirten düğüm uzunluğunu veren fonksiyon $l(v_i)$ olsun,

Eğer analiz edilmekte olan bağlantı kaydı x ve öncesinde gelmiş olan ardışık $l(v_i)-1$ bağlantı kaydının oluşturduğu kayıt katarı, v_i 'ye eşit ise bu düğüm bu akışta görülmüş demektir.

Analiz edilmekte olan akışta görülen düğümlerin oluşturduğu küme $(Akis \cap V_{AtakTipi})$, bu kümenin eleman sayısını veren fonksiyon $s(Akis \cap V_{AtakTipi})$ olur,

Nüfuz/Sızma tahmin değerlendirmesinde; x bağlantı kaydı analizinde kullanılan karar kuralı $L(x)$ aşağıdaki gibidir.

$$KararKurali : \begin{cases} L(x) = Nufuz, eger \\ s(Akis \cap V_{AtakTipi}) \geq s(ImzaKumesi) * EsikDeger \\ L(x) = Normal, eger \\ s(Akis \cap V_{AtakTipi}) < s(ImzaKumesi) * EsikDeger \end{cases} \quad (3.4)$$

3.3.7. Yöntemin Uyarlanması

Bu bölümde; önceki bölümlerde düğüm skorlarının hesaplanması, kenar skorlarının hesaplanması ve kenar imzalarının çıkarılmasında anlatılması için verilen matematiksel ifadelerin sözde kodları verilmiştir.

3.3.3 D ğ m Skorlarının Hesaplanması bařlıđı altında anlatılan $P_{cluster_V}$ prosed r  ve s zde kodu ařađıdaki gibidir.

$P_{cluster_V}$: D ğ m skorlarını hesaplar, G  l  ve Zayıf d ğ m k melerini oluřturur

Girdiler: D ğ m k mesi V , Sızma Akıř Havuzu, Normal Akıř Havuzu

 ıktılar: G  l  d ğ m k mesi V_{strong} , Zayıf d ğ m k mesi V_{weak} , D ğ m skorları k mesi V_{score}

- Adım 1: V_{strong} , V_{weak} , V_{score} 'a $\emptyset$ atanması
- Adım 2: V k mesindeki her v_i 'nin "strong" veya "weak" olma durumunu tutan S 'ye $\emptyset$ atanması
- Adım 3: V_{strong} k mesi merkez deđerini tutacak c_{strong} 'a 1 atanması
- Adım 4: V_{weak} k mesi merkez deđerini tutacak c_{weak} 'e 0 atanması
- Adım 5: V k mesindeki her v_i i in
 - Adım 5.1: $f_{V_{score}}(v_i)$ 'nin $(V_{score})_i$ 'a atanması
- Adım 6: V k mesindeki her v_i i in
 - Adım 6.1: $(V_{score})_i$, c_{weak} 'e daha yakın ise $S(v_i)$ 'ye "weak" atanması
 - Adım 6.2: $(V_{score})_i$, c_{strong} 'a daha yakın ise $S(v_i)$ 'ye "strong" atanması
- Adım 7: S deđiřtik e
 - Adım 7.1: $S(v_i) = \text{"strong"}$ elemanları i in c_{strong} hesaplanması
 - Adım 7.2: $S(v_i) = \text{"weak"}$ elemanları i in c_{weak} hesaplanması
 - Adım 7.3: $(V_{score})_i$, c_{weak} 'e daha yakın ise $S(v_i)$ 'ye "weak" atanması
 - Adım 7.4: $(V_{score})_i$, c_{strong} 'a daha yakın ise $S(v_i)$ 'ye "strong" atanması
- Adım 8: V k mesindeki her v_i i in
 - Adım 8.1: $S(v_i) = \text{"strong"}$ ise v_i 'yi V_{strong} k mesine eklenmesi
 - Adım 8.2: $S(v_i) = \text{"weak"}$ ise v_i 'yi V_{weak} k mesine eklenmesi

3.3.4 Kenar Skorlarının Hesaplanması başlığı altında anlatılan $P_{cluster_E}$ prosedürü ve sözde kodu aşağıdaki gibidir.

$P_{cluster_E}$: Kenar skorlarını hesaplar, Güçlü ve Zayıf kenar kümelerini oluşturur

Girdiler: Kenar kümesi $E = (V_{weak} \times V_{weak})$, Sızma Akış Havuzu, Normal Akış Havuzu

Çıktılar: Güçlü kenar kümesi E_{strong} , Zayıf kenar kümesi E_{weak} , Kenar skorları kümesi E_{score}

- Adım 1: E_{strong} , E_{weak} , E_{score} 'a $\emptyset$ atanması
- Adım 2: E kümesindeki her e_{ij} 'nin "strong" veya "weak" olma durumunu tutan S 'ye $\emptyset$ atanması
- Adım 3: E_{strong} kümesi merkez değerini tutacak c_{strong} 'a 1 atanması
- Adım 4: E_{weak} kümesi merkez değerini tutacak c_{weak} 'e 0 atanması
- Adım 5: E kümesindeki her e_{ij} için
 - Adım 5.1: $f_{E_{score}}(e_{ij})$ 'nin $(E_{score})_{ij}$ 'a atanması
- Adım 6: E kümesindeki her e_{ij} için
 - Adım 6.1: $(E_{score})_{ij}$, c_{weak} 'e daha yakın ise $S(e_{ij})$ 'ye "weak" atanması
 - Adım 6.2: $(E_{score})_{ij}$, c_{strong} 'a daha yakın ise $S(e_{ij})$ 'ye "strong" atanması
- Adım 7: S değıştikçe
 - Adım 7.1: $S(e_{ij}) = \text{"strong"}$ elemanları için c_{strong} hesaplanması
 - Adım 7.2: $S(e_{ij}) = \text{"weak"}$ elemanları için c_{weak} hesaplanması
 - Adım 7.3: $(E_{score})_{ij}$, c_{weak} 'e daha yakın ise $S(e_{ij})$ 'ye "weak" atanması
 - Adım 7.4: $(E_{score})_{ij}$, c_{strong} 'a daha yakın ise $S(e_{ij})$ 'ye "strong" atanması
- Adım 8: E kümesindeki her e_{ij} için
 - Adım 8.1: $S(e_{ij}) = \text{"strong"}$ ise e_{ij} 'nin E_{strong} kümesine eklenmesi
 - Adım 8.2: $S(e_{ij}) = \text{"weak"}$ ise e_{ij} 'yi E_{weak} kümesine eklenmesi

3.3.5 Kenar İmzalarının Çıkartılması başlığı altında anlatılan $P_{extract_signature_edges}$ prosedürü ve sözde kodu aşağıdaki gibidir.

$P_{extract_signature_edges}$: Kenar imzalarını çıkarır

Girdiler: Zayıf düğüm kümesi V_{weak} , Kenar skorları kümesi E_{score}

Çıktılar: Kenar imzaları kümesi $E_{signature_edges}$

- Adım 1: $E_{signature_edges}$ 'a $\emptyset$ atanması
- Adım 2: N 'e V_{weak} kümesi eleman sayısının atanması
- Adım 3: V_{weak} kümesindeki her v_i elemanı için v_i 'den v_i 'ye giden maximum skorlu yolun skorunu tutan $N \times N$ boyutunda $MaxSkor$ matrisinin tanımlanması
- Adım 4: $MaxSkor$ kümesindeki her m_{ij} elemanına 0 değerinin atanması
- Adım 5: Maximal yol üzerinde, v_i düğümüne doğrudan bağlantısı olan v_j düğümünün indis değeri j 'yi tutan D kümesine $\emptyset$ atanması
- Adım 6: $MaxSkor$ matrisinin her elemanı m_{ij} için
 - Adım 6.1: $m_{1j} + (E_{score})_{ij}$, m_{1j} 'den küçük ise
 - Adım 6.1.1: m_{1j} 'ye $m_{1j} + (E_{score})_{ij}$ atanması
 - Adım 6.1.2: D kümesinin j . elemanına i değerinin atanması
- Adım 7: i 'ye N değerinin atanması
- Adım 8: i , 1'e eşit değil iken
 - Adım 8.1: D kümesinin i . elemanının değerinin d_i değişkenine atanması
 - Adım 8.2: E kümesinin (d_i, i) . elemanı $e_{d_i, i}$ değişkenine atanması
 - Adım 8.3: $e_{d_i, i}$ "strong" ise $e_{d_i, i}$ 'nin $E_{signature_edges}$ 'a eklenmesi

4. DENEYSEL SONUÇLAR, ANALİZ ve KARŞILAŞTIRMA

4.1. Yöntemin Uygulanması

Yöntemin uygulanmasında; geliştirme, MATLAB programlama dili ile gerçekleştirilmiştir.

Analiz ve testler, KDD Cup 99 veri kümesi kullanılarak yapılmıştır. Bu çalışmada KDD Cup veri kümesinin kullanılmasının sebebi; geçmiş ve günümüzde yapılan sızma tespit ve tahmin çalışmalarında kullanılıyor olması ve bu nedenle önerilen yöntemin sonuçlarının diğer çalışmalarla karşılaştırılmasına olanak sağlamasıdır.

KDD Cup 99 veriseti, DARPA 98 verisetinin bazı özniteliklerinin çıkartılmasıyla (başlangıç tarihi, ip ve port) oluşturulmuştur. KDD Cup 99 veriseti yaklaşık 4.900.000 tane saldırı içermektedir. (KDD, 1999) Saldırı oranı doğal değildir. Yaklaşık %80 oranında saldırı içermektedir.

KDD Cup 99 verisetinde kayıtlar 41 öznitelikten oluşmaktadır. 42. öznitelik ise o kaydın görüldüğü akışın tipini belirtir. KDD Cup 99 veriseti öznitelikleri Tablo 4.1’de verilmiştir.

Tablo 4.1: KDD Cup 99 veriseti öznitelikleri.

1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login	42	attack_type

KDD Cup 99 verisetindeki saldırı tipleri 4 ana başlıkta gruplanmaktadır.

Hizmet Engelleme (Denial of Service – DoS): Bilgisayar ağlarında; bazı hesaplamalar yapılarak veya kaynakların çok fazla meşgul edilmesi ile sistemin gerçek isteklere cevap veremeyecek hale getirilmesidir.

Yönetici Hesabı ile Yerel Oturum Açma (Remote to local – R2L): Bilgisayar ağlarında; kullanıcı haklarına sahip olunmadığı halde ağa izinsiz erişim yapılmasıdır.

Kullanıcı Hesabının Yönetici Hesabına Yükseltilmesi (User to root – U2R): Bilgisayar ağlarında yönetici izni olmayan bir kullanıcının hakkı olmadığı erişimde ve işlemlerde bulunmasıdır.

Bilgi Tarama (Probe): Bilgisayar ağlarında; bilgi toplamak ya da önceden bilinen açıkları bulmak amaçlı gerçekleştirilen saldırılardır.

4.2. Deneysel Sonular

KDD Cup 99 veri setinde ardışık olarak görülen aynı tipteki kayıt katarları, akışlar olarak tanımlanmaktadır. Bir atak tipi analiz edilirken; analiz edilmekte olan atak tipindeki kayıt dizileri atak akışları, diğer tipte olan kayıt dizileri ise normal akışlar olarak kabul edilmektedir.

Düğümlemlerin seçilmesinde her atak tipi için belli öznitelikler kullanılarak analiz yapılmıştır. Aynı tipteki atak akışlarında ortak görünen minimum n uzunluğundaki ardışık kayıt katarları belirlenmektedir. Bu kayıt katarları düğümler olarak tanımlanır. Bu özniteliklerin belirlenmesinde KDD Cup 99 veri setindeki saldırı tipleri ve öznitelikler arasındaki bağlantıları inceleyen çalışmalardan faydalanılmıştır [31], [32], [33]. Analizler sırasında düğümlerin çıkartılması için her bir atak tipi için kullanılan öznitelikler Tablo 4.2’de verilmiştir.

Tablo 4.2: Atak tipleri analizinde kullanılan öznitelikler.

Atak Tipi	Öznitelik	Kayıt Sayısı (Öğrenme)	Atak Grubu
Back	10, 13	2203	DoS
Land	7	21	DoS
Neptune	33, 35, 38, 39	107201	DoS
Pod	5	264	DoS
Smurf	3, 23, 28, 26, 41	280790	DoS
Teardrop	5, 8	979	DoS
Ftp write	5, 9, 23	8	R2L
Guesspassword	3, 4, 35	53	R2L
Imap	3, 26	12	R2L
Multihop	23	7	R2L
Phf	5, 6, 10, 14	4	R2L
Spy	-	2	R2L
Warezclient	10, 5, 1	1020	R2L
Warezmaster	-	20	R2L
Bufferoverflow	14	30	U2R
Loadmodule	3, 24, 36	9	U2R
Perl	14, 16	3	U2R
Rootkit	4, 23, 24	10	U2R
Ipsweep	3, 19, 22	1247	Probe
Nmap	4, 5	231	Probe
PortswEEP	15, 28	1040	Probe
Satan	22, 23, 27	1589	Probe
Normal	3, 34, 35, 39, 38, 26, 25, 30, 29	97277	-

Bölüm 4.4 Özet Analiz Sonuçları kısmında analiz sonuçları özet olarak tek tabloda verilmiştir. Bu tabloda; her bir atak tipi için Saldırı Tespit Oranı, Yanlış Pozitif Oranı ve Yanlış Negatif Oranları görülebilmektedir.

Bölüm 4.5 Detaylı Analiz Sonuçları kısmında ise her atak grubu için içerdiği atak tipleri sonuçları verilmiştir.

Analiz ölçümleri sonuçlarının verilmesi için kullanılan değerlerin tanımları aşağıdaki gibidir.

TP: (Doğru Pozitif / True Positive): Tahmin edilen ve alarm verilen saldırı akışlarının içerdiği bağlantı kaydı sayısıdır.

TN: (Doğru Negatif / True Negative): Alarm verilmeyen ve saldırı olmayan akışların içerdiği bağlantı kaydı sayısıdır.

FP: (Yanlış Pozitif / False Positive): Alarm verilen ancak saldırı olmayan akışların içerdiği bağlantı kaydı sayısıdır.

FN: (Yanlış Negatif / False Negative): Alarm verilmeyen ancak saldırı olan akışların içerdiği bağlantı kaydı sayısıdır.

Doğruluk: Tahmin edilen saldırı akışları ile normal akış kayıtlarının toplamının tüm kayıtlara oranıdır.

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Hata: Tahmin edilemeyen saldırı ve normal akış kayıtlarının toplamının tüm kayıtlara oranıdır.

$$\text{Hata} = \frac{FN + FP}{TP + TN + FP + FN} \quad (4.2)$$

Duyarlılık: Tahmin edilen saldırı akışlarındaki kayıt sayısının tüm saldırı kayıtlarına oranıdır.

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \quad (4.3)$$

Kesinlik: Tahmin edilen saldırı akışlarındaki kayıt sayısının, doğru tahmin edilen saldırı ve saldırı olarak tahmin edilen normal kayıtlar toplamına oranıdır.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \quad (4.4)$$

ADR: (Saldırı Tespit Oranı / Attack Detection Rate): Tahmin edilen saldırıların tüm saldırılara oranıdır.

$$ADR = \frac{TP}{TP + FN} \quad (4.5)$$

FPR: (Yanlış Pozitif Oranı / False Positive Rate): Saldırı olarak tahmin edilen normal kayıtların tüm gerçek normal kayıtlara oranıdır.

$$FPR = \frac{FP}{FP + TN} \quad (4.6)$$

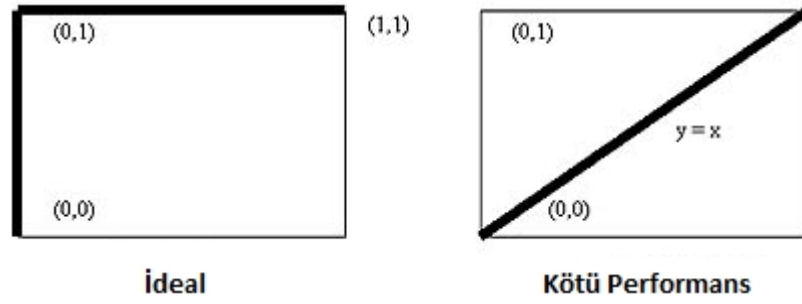
4.3. Eşik Değerinin Belirlenmesi

Nüfuz tespit ve tahmin sistemlerinin başarısını ölçmede kullanılan ADR ve FPR değerleri arasındaki ilişkiyi görselleştirmek için ROC (Receiver Operating Characteristic) eğrisi kullanılır [36]. ROC eğrisi, istatistik karar teorisine dayanır, 1950'lerin başlarında elektronik sinyal tanımlamaları ve radar problemlerinde kullanılmaya başlanmıştır. ROC, basit anlamda doğru pozitiflerin, yanlış pozitiflere olan oranı olarak ifade edilebilir.

ROC eğrisi; değişen parametrelere göre doğru pozitiflerin oranı (ADR) değerlerinin, yanlış pozitif oranı (FPR) değerlerinin bir fonksiyonu olarak çizilmesiyle oluşmaktadır. Her sınıflandırma işleminde yapıldığı gibi, yöntemler kesinlik (yanlış pozitifleri eleme kabiliyeti) ve hassasiyet (doğru pozitifleri tespit etme kabiliyeti) arasındaki dengeyi kurmakla uğraşmaktadır. ROC eğrisi de kesinlik ve hassasiyet arasındaki dengeyi değerlendirmek için kullanılmaktadır.

ROC eğrilerindeki her bir nokta belli bir eşik değerle elde edilen ADR ve FPR değerlerini temsil eder. İdeal bir testin sonucunda elde edilen ROC eğrisi sol üst köşeden (%100 ADR, %0 FPR) geçer.

İdeal ve kötü performans göstergesi olan ROC eğrileri Şekil 4.1'de verilmiştir [39].



Şekil 4.1: İdeal ve kötü performans göstergesi olan ROC eğrileri.

Bölüm 3.3.6’da verilen yöntemin matematiksel modelinde tanımlanan eşik değerin belirlenmesinde ROC eğrilerinden faydalanılmıştır. Eşik Değer aşağıdaki gibi tanımlanabilir.

Eşik Değer: Analiz edilmekte olan bir akışın saldırı akışı olduğuna karar vererek alarm üretmek için ilgili akışta görülmesi gereken en az imza sayısının ilgili atak tipi için çıkartılan tüm imzalara oranı (Bkz. Bölüm 3.3.6)

Tablo 4.2’de verilen öznelilikler kullanılarak öğrenme veriseti üzerinde eşik değeri değiştirilerek yapılan analizlerde elde edilen ADR ve FPR sonuçlarına göre ROC eğrileri çizilmiştir. Elde edilen ROC eğrilerinde sol üst köşeye en yakın noktayı (ADR ve FPR değerleri) veren eşik değeri, ideal eşik değeri olarak kabul edilmiştir. Eğer eşit uzaklıkta birden fazla nokta varsa en yüksek ADR değerini veren eşik değeri, ideal eşik değeri olarak kabul edilir [37]. Sol üst köşenin koordinatı (0, 1) olur. Sol üst köşe ile diğer noktalar arasındaki uzaklık hesaplanırken Öklid uzaklığı formülü kullanılmaktadır.

$P=(p_x, p_y)$ ve $Q=(q_x, q_y)$ noktaları için Öklid uzaklığı formülü aşağıdaki gibidir.

$$\text{Öklid Uzaklığı} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (4.7)$$

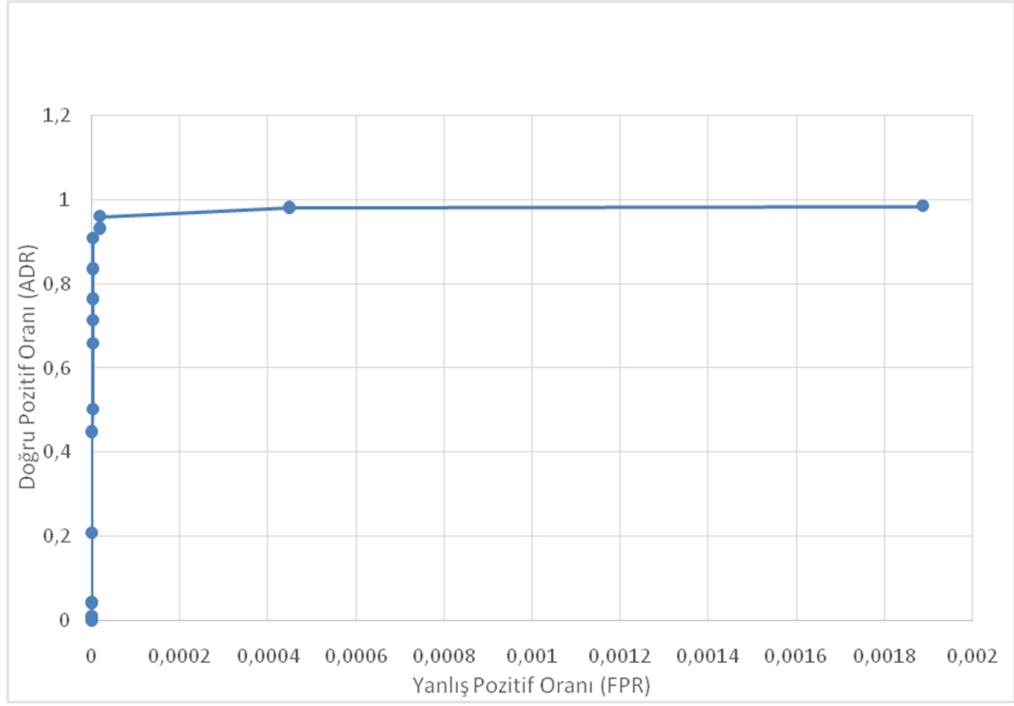
Eşik değerlerinin tespiti çalışmasına örnek olarak Normal akışlar ve Neptune atak akışları için çizilen ROC eğrileri ileriki bölümlerde verilmiştir.

4.3.1. Normal

Normal akışların analizinde kullanılan eşik değerler ve elde edilen ADR ve FPR değerleri Tablo 4.3’de verilmiştir. Bu parametreler kullanılarak Şekil 4.2’deki ROC eğrisi oluşturulmuştur. Oluşan ROC eğrisinde sol üst köşeye en yakın noktayı veren eşik değeri 0,0001 olarak belirlenmiştir.

Tablo 4.3: Normal akışlar için analiz parametreleri.

Eşik Değer	ADR	FPR
0,0001	0,983920953	0,001887867
0,001	0,980163868	0,001887867
0,005	0,960476322	0,000448652
0,01	0,93052578	1,76436E-05
0,02	0,907728556	1,76436E-05
0,03	0,835473669	2,52052E-06
0,04	0,7633143	2,52052E-06
0,05	0,713018191	2,52052E-06
0,06	0,657617117	2,52052E-06
0,07	0,502112033	2,52052E-06
0,1	0,447920868	2,52052E-06
0,2	0,207892	0
0,3	0,04212392	0
0,4	0,043376282	0
0,5	0,010921016	0
0,6	0,008278321	0
0,7	0,008278321	0
0,8	0	0
0,9	0	0
1	0	0



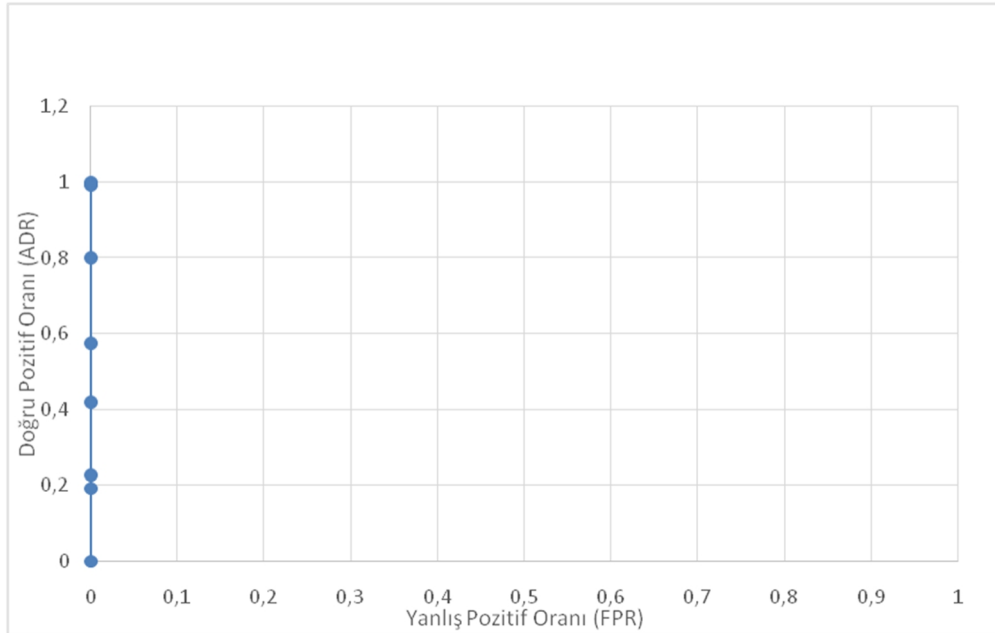
Şekil 4.2: Normal akışlar için ROC eğrisi.

4.3.2. Neptune

Neptune atak akışlarının analizinde kullanılan eşik değerler ve elde edilen ADR ve FPR değerleri Tablo 4.4’de verilmiştir. Bu parametreler kullanılarak Şekil 4.3 ROC eğrisi oluşturulmuştur. Oluşan ROC eğrisinde sol üst köşeye en yakın noktayı veren eşik değeri 0,0001 olarak belirlenmiştir.

Tablo 4.4: Neptune atak akışları için analiz parametreleri.

Eşik Değer	ADR	FPR
0,0001	1	0
0,001	0,999794778	0
0,005	0,999794778	0
0,01	0,999794778	0
0,02	0,994076548	0
0,03	0,993330286	0
0,04	0,993330286	0
0,05	0,993330286	0
0,06	0,993330286	0
0,07	0,993330286	0
0,1	0,992584024	0
0,2	0,992584024	0
0,3	0,992584024	0
0,4	0,800794769	0
0,5	0,574267031	0
0,6	0,419063255	0
0,7	0,191649332	0
0,8	0,227273999	0
0,9	0	0
1	0	0



Şekil 4.3: Neptune atak akışları için ROC eğrisi.

4.4. Özet Analiz Sonuçları

Analizler yapılırken Nüfuz/Sızma Tahmin Sistemi her atak tipi ve normal kayıtları için ayrı ayrı çalıştırılmıştır. Bir akış analiz edilirken alarm üretilirse o atak akışı tahmin edilmiş sayılmaktadır. Tüm veri kümesi analiz edildikten sonra tahmin edilen ve tahmin edilemeyen atak akışları değerlendirilerek aşağıdaki deneysel sonuçlar çıkartılmıştır.

Tablo 4.5: Akış çizgesi tabanlı nüfuz tahmin sistemi özet analiz sonuçları.

Atak Tipi	Saldırı Tespit Oranı	Yanlış Pozitif Oranı	Yanlış Negatif Oranı	Kayıt Sayısı (Test)
Back	1	0	0	2203
Land	1	0,0002	0	21
Neptune	0,994	0,01	0,006	1072002
Pod	0,697	0	0,303	264
Smurf	1	0	0	2807886
Teardrop	0,897	0	0,103	979
Ftp write	0,50	0,391	0,50	8
Guesspassword	1	0,125	0	53
Imap	1	0,047	0	12
Multihop	1	0,323	0	7
Phf	1	0	0	4
Spy	-	-	-	2
Warezclient	0,613	0,158	0,387	1020
Warezmaster	-	-	-	20
Bufferoverflow	0,867	0,129	0,133	30
Loadmodule	0,333	0,126	0,667	9
Perl	1	0,002	0	3
Rootkit	1	0,779	0	10
Ipsweep	0,926	0,114	0,074	1247
Nmap	0,892	0,001	0,108	2316
PortswEEP	0,708	0	0,292	10413
Satan	0,996	0	0,004	1589
Normal	0,988	0,001	0,012	972780

4.5. Detaylı Analiz Sonuçları

4.5.1. DoS (Denial of Services)

Tablo 4.6’da DoS grubu atak akışları için detaylı analiz sonuçları verilmiştir.

Tablo 4.6: DoS grubu atak tipleri detaylı analiz sonuçları.

	Back	Land	Neptune	Pod	Smurf	Teardrop
True Positive	2203	21	1065863	184	2807886	879
True Negative	4896228	4897399	3787067	4898167	2090545	4897452
False Positive	0	1011	39362	0	0	0
False Negative	0	0	6139	80	0	100
Doğruluk	1	0,99979	0,99071	0,99998	1	0.99998
Hata	0	0,00021	0,00929	0,00001	0	0,00002
Duyarlılık	1	1	0,99427	0,69697	1	0.89785
Kesinlik	1	0,02035	0,96439	1	1	1
ADR	1	1	0,99427	0,69697	1	0.89785
FPR	0	0,00021	0,01029	0	0	0
FPR/ADR	0	0,00021	0,01035	0	0	0

4.5.2. R2L (Remote to Local)

Tablo 4.7 ve Tablo 4.8’de R2L grubu atak akışları için detaylı analiz sonuçları verilmiştir.

Tablo 4.7: R2L grubu atak akışları birinci kısım detaylı analiz sonuçları.

	Ftp write	Guess password	Imap	Multihop
True Positive	4	53	12	7
True Negative	2980669	4284437	4668690	3314065
False Positive	1917754	613941	229729	1584359
False Negative	4	0	0	0
Doğruluk	0.60850	0.87467	0.95310	0.67656
Hata	0.39150	0.12533	0.04690	0.32344
Duyarlılık	0.50000	1	1	1
Kesinlik	0	0.00009	0.00005	0
ADR	0.50000	1	1	1
FPR	0.39150	0.12534	0.04690	0.32344
FPR/ADR	0.78301	0.12534	0.04690	0.32344

Tablo 4.8: R2L grubu atak akışları ikinci kısım detaylı analiz sonuçları.

	Phf	Spy	Warezcclient	Warezmater
True Positive	4		626	
True Negative	4487633		4122430	
False Positive	0		774981	
False Negative	0		394	
Doğruluk	1		0.84171	
Hata	0		0.15829	
Duyarlılık	1		0.61373	
Kesinlik	1		0.00081	
ADR	1		0.61373	
FPR	0		0.15824	
FPR/ADR	0		0.25784	

4.5.3. U2R (User to Root)

Tablo 4.9’de U2R grubu atak akışları için detaylı analiz sonuçları verilmiştir.

Tablo 4.9: U2R grubu atak akışları detaylı analiz sonuçları.

	Bufferoverflow	Loadmodule	Perl	Rootkit
True Positive	26	3	3	10
True Negative	4265402	4280474	4821363	1081543
False Positive	632999	617948	9877	3816878
False Negative	4	6	0	0
Doğruluk	0.87077	0.87385	0.99796	0.22080
Hata	0.12923	0.12615	0.00204	0.77920
Duyarlılık	0.86667	0.33333	1	1
Kesinlik	0.00004	0	0.0003	0
ADR	0.86667	0.33333	1	1
FPR	0.12923	0.12615	0.00204	0.77921
FPR/ADR	0.14911	0.37846	0.00204	0.77921

4.5.4. Probe

Tablo 4.10’de Probe grubu atak akışları için detaylı analiz sonuçları verilmiştir.

Tablo 4.10: Probe grubu atak akışları detaylı analiz sonuçları.

	Ipsweep	Nmap	Portssweep	Satan
True Positive	11498	2066	7373	15821
True Negative	4327780	4887935	4888018	1448739
False Positive	558241	8180	0	0
False Negative	912	250	3040	57
Doğruluk	0.88585	0.99828	0.99938	0.99996
Hata	0.11415	0.00172	0.00062	0.00004
Duyarlılık	0.92651	0.89206	0.70806	0.99641
Kesinlik	0.02018	0.20164	1	1
ADR	0.92651	0.89206	0.70806	0.99641
FPR	0.11425	0.00167	0	0
FPR/ADR	0.12331	0.00187	0	0

4.5.5. Normal

Tablo 4.11’de Normal akışlar için detaylı analiz sonuçları verilmiştir.

Tablo 4.11: Normal akışlar detaylı analiz sonuçları.

	Normal
True Positive	961295
True Negative	3919967
False Positive	5597
False Negative	11572
Doğruluk	0,99650
Hata	0,00350
Duyarlılık	0,98811
Kesinlik	0,99421
ADR	0,98811
FPR	0,00143
FPR/ADR	0,00144

4.6. Karşılaştırma

Atak gruplarının saldırı tespit oranları hesaplanırken ilgili gruptaki atak tiplerinin saldırı tespit oranları atak tipindeki kayıt sayılarına göre ağırlıklı ortalaması alınır.

$$\frac{\sum_{i=Ataktipi} (SaldiriTespitOrani_i * KayitSayisi_i)}{KayitSayisi_{AtakGrubu}} \quad (4.8)$$

Yapılan analizler sonucu elde edilen ve detayları yukarıdaki bölümlerde verilen sonuçlardan aşağıdaki tablo elde edilmiştir.

Tablo 4.12: Akış çizgesi tabanlı nüfuz tahmin sistemi analiz sonuçları.

	Saldırı Tespit Oranı	Yanlış Pozitif Oranı	Yanlış Negatif Oranı
DoS	0,998	~0,000	0,002
R2L	0,626	0,154	0,374
U2R	0,808	0,246	0,192
Probe	0,782	0,009	0,218
Normal	0,988	0,012	0,001

Elde edilen veriler; nüfuz tahmin sistemi önerisinde bulunan ve yine KDD Cup 99 veriseti ile test edilmiş olan iki farklı çalışmanın sonuçları ile karşılaştırmada kullanılmıştır.

Karşılaştırma yapılan çalışmalardan ilki Sindhu ve arkadaşları tarafından yapılmıştır [34]. Bu çalışmada sunulan sistemde Neuro-Genetic algoritmaları kullanılarak ataklar tespit edilmekte ve yeni atak tipleri tahmin edilerek sınıflandırılmaktadır. IDS'lerin dinamik olarak değişen ortamlarda çalışmasından dolayı tespit modelinin sürekli güncellenmesi gerekmektedir. Elle yapılan model güncellemesindeki operatör bağımlılığını azaltma motivasyonu ile geliştirilen sistemi deneysel sonuçlarına bakıldığında, sistemin diğer bilinen IDS sistemlerine göre yanlış sınıflandırma maliyeti bakımından ilerleme sağladığı görülmüştür. Önerilen sistem, yanlış tahminler yapıldığında sistem operatörü tarafından sağlanan geribildirim ile tespit modelini sürekli otomatik olarak güncellemektedir.

Geliştirilen sistem KDDCup'99 veriseti ile değerlendirilmiştir ve nüfuzları belli eden baskın öz nitelikler seçilirken genetik paradigması kullanılmıştır. Her bir öz niteliğin ağırlık değeri hesaplanıp, Neuro-Genetic algoritması tarafından işlenebilir hale getirilmiştir.

Sindhu ve arkadaşlarının yaptığı çalışmanın sonuçları Tablo 4.13'de verilmiştir.

Tablo 4.13: A neuro-genetic based short-term forecasting framework for network intrusion prediction system analiz sonuçları.

	Saldırı Tespit Oranı	Yanlış Pozitif Oranı	Yanlış Negatif Oranı
DoS	0,968	-	0,032
R2L	0,97	-	0,003
U2R	0,885	-	0,115
Probe	0,6873	-	0,3127
Normal	0,998	0,002	-

Karşılaştırma yapılan ikinci çalışma ise Yu ve arkadaşları tarafından yapılmıştır [35]. Bu çalışmada da diğer çalışmada olduğu gibi yanlış tahminler yapıldığında sistem operatörü tarafından sağlanan geribildirim ile tespit modelini sürekli otomatik olarak güncelleyen bir sistem sunulmaktadır. Bu sistem de KDDCup'99 veriseti ile değerlendirilmiştir. Deneysel sonuçlarına bakıldığında sistem, güncelleme yapılmayan sistemlere göre yanlış sınıflandırma maliyetinde %35 iyileşme sağlamaktadır. Eğer yanlış tahminlerin sadece %10'u bile modeli güncellemek için kullanılırsa sistem %30 iyileşme sağlamaktadır. Eğer yanlış tahminlerin sadece %1,3'ü modeli güncellemek için kullanılırsa sistem %20 iyileşme sağlamaktadır. Sistemin modelinin güncellenmesi için operatörlerin düşük güvenilirlik değeri ile yapılan tahminleri değerlendirmesi yeterlidir.

Yu ve arkadaşlarının yaptığı çalışmanın sonuçları da Tablo 4.14’de verilmiştir.

Tablo 4.14: An adaptive automatically tuning intrusion detection system analiz sonuçları.

	Saldırı Tespit Oranı	Yanlış Pozitif Oranı	Yanlış Negatif Oranı
DoS	0,992	0,023	0,008
R2L	0,244	0,255	0,756
U2R	0,131	0,891	0,869
Probe	0,883	0,134	0,117
Normal	0,953	0,047	0,147

Karşılaştırılmalarda önerilen yöntemin Remote to Local (R2L) ve User to Root (U2R) saldırı grubu atakların tahmininde Tablo 4.14’deki yöntemden daha iyi sonuç verirken Probe saldırı grubu atakların tahmininde ise Tablo 4.13’deki yöntemden daha iyi sonuç verdiği gözlemlenmiştir. Denial of Service (DoS) grubu atakların tahmininde ise iki yöntemden de daha başarılı olmuştur.

5. SONUÇLAR ve ÖNERİLER

Bu çalışmada Nüfuz Tahmin Sistemleri konusunda yapılan çalışmalar konusunda özet bilgi verilmiş ve farklı bir yöntem önerisinde bulunulmuştur.

Nüfuz Tespit Sistemlerinin kullanımının ve buna bağlı olarak da öneminin artması nedeniyle bu konuda oldukça fazla çalışma yapılmaktadır. Gelişen teknoloji kullanılarak bu sistemlere saldırıları tespit etmenin yanında kullanıcının amacını önceden tahmin etmek ve uyarı vermek gibi yeteneklerin kazandırılması zorunlu hale gelmiştir.

Bu çalışmada, bilgisayar ağları saldırılarını önceden tahmin etmek için akış çizgesi tabanlı bir yöntem önerisinde bulunulmuştur. Bu yöntemde daha önceden gerçekleşen saldırı kayıtları analiz edilerek çıkarılan hibrit imzalar kullanılarak saldırılar önceden tahmin edilir. Yapılan deneysel sonuçlar karşılaştırılmalı olarak verilmiş ve yöntemin güvenilirliği gösterilmiştir.

Önerilen yöntemin ilk kez gerçekleşecek olan saldırı tiplerini tahmin edememektedir. Bu yaklaşımda, saldırıları önceden tahmin etmek üzere kullanılacak imzaları çıkarmak için, daha önce gerçekleşmiş olan saldırı bilgilerine ihtiyaç duyulmaktadır. İleriki çalışmalarda; sistemin gerçekleşen ataklarla öğrenebilir ve iyileşebilir olması konusunda çalışmalarda bulunulabilir.

Bununla birlikte ağ güvenliğinde sızma tahmin ve harekete geçme süresi oldukça önem taşımaktadır. Dolayısı ile yapılacak çalışmalarda önerilen yöntemin zaman açısından performansı analiz edilebilir ve sonuçları değerlendirilebilir.

KAYNAKLAR

- [1] Lee W., Stolfo S., (1998), "Datamining approaches for intrusion detection", Proceedings of the 7th USENIX Security Symposium, 79-94, San Antonio, Texas, USA, 26-29 January.
- [2] Han J., Kamber M., (2006), "Data Mining: Concepts and Techniques", 2nd Edition, Morgan Kaufmann Publishers.
- [3] Silahtaroglu G., (2008), "Kavram ve Algoritmalarıyla Temel Veri Madenciliği", 1st Edition, Papatya Yayıncılık.
- [4] Warrender C., Forrest S., Pearlmutter B., (1999), "Detecting intrusions using system calls: alternative data models", In Proceedings of the 1999 IEEE Symposium on Security and Privacy, 133-145, Oakland, California, USA, 9-12 May.
- [5] Yeung D., Ding Y., (2003), "Host-based intrusion detection using dynamic and static behavioral models", Pattern recognition, 36 (1), 229-243.
- [6] Lee W., Xiang D., (2001), "Information-theoretic measures for anomaly detection", In Proceedings of the 2001 IEEE Symposium on Security and Privacy, 130-143, Oakland, California, USA, 14-16 May.
- [7] Liao Y., Vemuri V., (2002), "Use of k-nearest neighbor classifier for intrusion detection", Computers & Security, 21 (5), 439-448.
- [8] Hu W., Liao Y., Vemuri V., (2003), "Robust support vector machines for anomaly detection in computer security", Proceeding of the 2003 International Conference on Machine Learning and Applications, 168-174, Los Angeles, California, USA, 23-24 June.
- [9] Sharma A., Pujari A., Paliwal K., (2007), "Intrusion detection using text processing techniques with a kernel based similarity measure", Computers & Security, 26 (7), 488-495.
- [10] Wang W., Guan X., Zhang X., (2008), "Processing of massive audit data streams for real-time anomaly intrusion detection", Computer Communications, 31 (1), 58-72.
- [11] Wang W., Guan X., Zhang X., Yang L., (2006), "Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data", Computers & Security, 25 (7), 539-550.

- [12] Wang W., Guan X., Zhang X., (2004), "Profiling program and user behaviors for anomaly intrusion detection based on non-negative matrix factorization", Proceedings of 43rd IEEE Conference on Control and Decision, 99-104, Atlantis, Paradise Island, Bahamas, 14-17 December.
- [13] Forrest S., Hofmeyr S.A., Somayaji A., Longstaff T.A., (1996), "A sense of self for Unix processes", In Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy, 120-128, Oakland, CA, USA, 6-8 May.
- [14] Mihajlovic, V., Petkovic, M., (2001), "Dynamic Bayesian Networks: A State of the Art", CTIT technical reports series, TR-CTIT-34.
- [15] Gimpel K., Rudoy D., (2007), "Statistical Inference in Graphical Models", TR-1115, Lexington Lincoln Laboratory, Massachusetts Institute Of Technology, USA.
- [16] Web 1, (2013), http://www.cs.bilkent.edu.tr/~saksoy/courses/cs551/slides/cs551_parametric4.pdf, (Erişim Tarihi: 24/06/2013).
- [17] Ghahramani, Z., (1998), "Learning dynamic Bayesian networks. In: Adaptive processing of sequences and data", 1st Edition, Springer Berlin Heidelberg.
- [18] Ghahramani, Z., (2001), "An Introduction to Hidden Markov Models and Bayesian Networks", International Journal of Pattern Recognition and Artificial Intelligence, 15 (1), 9-42.
- [19] Zheng Yu S., (2009), "Hidden semi-Markov models", Artificial Intelligence, 174 (2), 215-243.
- [20] Tweed D., Fisher R., Bins J., List T., (2005), "Efficient Hidden Semi-Markov Model Inference for Structured Video Sequences", In Proceedings of the 14th International Conference on Computer Communications and Networks (ICCCN '05), 247-254, Washington, DC, USA, 16 October.
- [21] Feng L., Wang W., Zhu L., Zhang Y., (2004), "Predicting the intrusion intentions by observing system call sequences", Computers & Security, 23 (3), 241-252.
- [22] Feng L., Wang W., Zhu L., Zhang Y., (2009), "Predicting intrusion goal using dynamic Bayesian network with transfer probability estimation", Journal of Network and Computer Applications, 32 (3), 721-732.
- [23] Haslum K., Abraham A., Knapskog S., (2007), "DIPS: A Framework for Distributed Intrusion Prediction and Prevention Using Hidden Markov Models and Online Fuzzy Risk Assessment", IAS '07 Proceedings of the Third International Symposium on Information Assurance and Security, 183-190, Manchester, United Kingdom, 29-31 August.

- [24] Haslum K., Abraham A., Knapskog S., (2008), "Fuzzy Online Risk Assessment for Distributed Intrusion Prediction and Prevention Systems", Tenth International Conference on Computer Modeling and Simulation, 216-223, Cambridge, United Kingdom, 1-3 April.
- [25] Haslum K., Moe M.E.G., Knapskog S., (2008), "Real-time Intrusion Prevention and Security Analysis of Networks using HMMs", The 33rd IEEE Conference on Local Computer Networks, 927-934, Montreal, Quebec, Canada, 14-17 October.
- [26] Zhengdao Z., Zhumiao P., Zhiping Z., (2008), "The study of intrusion prediction based on HsMM", Proceedings of the 3rd IEEE Asia-Pacific Services Computing Conference, 1358-1363, Yilan, Taiwan, 9-12 December.
- [27] Cheng-Bin L., (2009), "A New Intrusion Prediction Method Based on Feature Extraction", Second International Workshop on Computer Science and Engineering, 7-10, Qingdao, China, 28-30 October.
- [28] Li Z., Lei J., Wang L., Li D., (2007), "A Data Mining Approach to Generating Network Attack Graph for Intrusion Prediction", Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 307-311, Haikou, Hainan, China, 24-27 August.
- [29] Li Z., Lei J., Wang L., Li D., (2007), "Assessing Attack Threat by the Probability of Following Attacks", International Conference on Networking, Architecture, and Storage, 91-100, Guilin, China, 29-31 July.
- [30] Bayoglu B., Sogukpınar I., (2012), "Graph based signature classes for detecting polymorphic worms via content analysis", Elsevier, Computer Networks, 56 (2), 832-844.
- [31] Kayacik, H.G., Zincir-Heywood, A.N., Heywood, M.L., (2005), "Selecting Features for Intrusion Detection: A Feature Analysis on KDD 99 Intrusion Detection Datasets", In Proceedings of the third annual conference on privacy, security and trust, St. Andrews, New Brunswick, Canada, 12-14 October.
- [32] Olusola, A.A., Oladele, A.S., Abosede, D.O., (2010), "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features", In Proceedings of the World Congress on Engineering and Computer Science, 20-22, San Francisco, USA, 20-22 October.
- [33] Kim, B. J., Kim, I. K., (2005), "Robust Real-time Intrusion Detection System", Robust Real-time Intrusion Detection System, JIPS, 1 (1), 9-13.
- [34] Sindhu, S. S. S., Geetha, S., Marikannan, M., Kannan, A., (2009), "A neuro-genetic based short-term forecasting framework for network intrusion prediction system", International Journal of Automation and Computing, 6 (4), 406-414.

- [35] Yu, Z., Tsai, J. J., Weigert, T., (2008), “An adaptive automatically tuning intrusion detection system”, ACM Transactions on Autonomous and Adaptive Systems (TAAS), 3 (3), 10.
- [36] Qin, Z., (2005), “ROC Analysis for Predictions Made by Probabilistic Classifiers”, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 3119 – 3124, Guangzhou, China, 18-21 August.
- [37] Zweig, Mark H., Campbell G., (1993) “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine”, Clinical chemistry, 39 (4), 561-577.
- [38] Web 2, (2013), <http://gridsec.usc.edu/images/alert.jpg>, (Eriřim Tarihi: 12/04/2013).
- [39] Web 3, (2013), <http://www.ctf.istanbul.edu.tr/dergi/online/2001v32/s1/a4s1.jpg>, (Eriřim Tarihi: 19/08/2013).

ÖZGEÇMİŞ

Öznur Erdur Sever, 1982 yılında Manisa / Salihli’de doğdu. İlköğretim ve lise öğrenimlerini sırasıyla Altınordu İlkokulu ve Sekine Evren Anadolu Lisesi’nde tamamladı. 1999 yılında başladığı Dokuz Eylül Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü’ndeki lisans eğitimini 2003 yılında tamamladı. 2011 yılında Gebze Yüksek Teknoloji Enstitüsü (GYTE), Mühendislik ve Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalında yüksek lisans eğitimine başladı. 2004-2012 yılları arasında Yaltes A.Ş. bünyesinde Yazılım Mühendisi olarak çeşitli savunma sanayii projelerinde görev aldı. 2012 yılından bu yana Ak Yatırım A.Ş.’de Yazılım Mühendisi olarak çalışmaktadır.