Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

Big Data Stream Mining ile Bilgisayar Ağlarında Anomali Tespiti

HAZIRLAYANLAR

Hasan Bilal AYYILDIZ	131180011
Mesut AKTAŞ	131180501
Muhammet YURT	121180072

2017-2018 Güz Dönemi
Ankara

BIG DATA STREAM MINING İLE BİLGİSAYAR AĞLARINDA ANOMALİ TESPİTİ

(Bitirme Projesi Raporu)

GAZİ ÜNİVERSİTESİ

Bilgisayar Mühendisliği Bölümü

Ocak 2018

ÖZET

Bu çalışmada büyük veri madenciliği tekniklerinden biri olan kümeleme tekniği ile bilgisayar ağlarında anomali tespit analizleri gerçekleştirilmiştir. KDD Cup'99 verisinin yüzde onluk kısmı ile sistem eğitilmiş ve kalan kısmı ile test edilmiştir.

Sayfa Adedi : 12

Danışman : Prof. Dr. M. Ali AKCAYOL

TEŞEKKÜR

Çalışmalarımız boyunca değerli yardım ve katkılarıyla bizi yönlendiren Danışman Hocamız Sayın Prof. Dr. M. Ali AKCAYOL'a, Arş. Gör. Anıl UTKU' ya ve ayrıca Lisans eğitimimiz süresince yardımlarını esirgemeyen tüm hocalarımıza teşekkür ederiz.

İçindekiler Tablosu

1. GİRİŞ **1**

2. KULLANILAN ARAÇ VE YÖNTEMLER **2**

2.1 Kullanılan Veriseti 2

2.1.1 KDDCup99 Veri Setinin Detaylı Anlatımı 2

2.1.2 Saldırı Tipleri 4

2.1.2.1 Hizmet aksattırma saldırıları 5

2.1.2.2 U2R Saldırıları 5

2.1.2.3 R2L saldırıları 5

2.1.2.4 Probing saldırıları 5

2.1.3 Veri Setinin Hazırlanması 7

2.2 Kullanılan Ortamlar 8

2.2.1 Apache Spark 8

2.1.2 Databricks 8

2.3 Kullanılan Diller 8

2.4 Kullanılan Algoritmalar 8

2.4.1 K-Means Kümeleme Algoritması 8

2.5 Kullanılan Yöntemler 9

2.5.1 PCA (Principal Component Analysis) 9

3. SONUÇ VE DEĞERLENDİRME **10**

REFERANSLAR 1**2**

Çizelgelerin Listesi Ve Kısaltmalar

Çizelge 1.1 İçerik Özellikleri 3
Çizelge 1.2 Sunucu Tabanlı Trafik Özellikleri 3
Çizelge 1.3 Zamana bağlı trafik özellikleri 4
Çizelge 1.4 KDD'99 veri kümesinin %10'luk kısmındaki saldırı örneklerinin sayıları 6
Çizelge 1.5 Eğitim kümesinde yer almayan test veri kümesinde bulunan saldırılar 7

ICMP	Internet Control Message Protocol
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
KDD	Knowledge Discovery and Data Mining

1. GİRİŞ

Yaşadığımız çağda şimdiye kadar hiç olmadığı kadar bilgi üretilmekte işlenmekte ve bunlara erişilmektedir. Bilginin bu denli hızlı üretilip yayılmasında hiç kuşkusuz bilgisayar teknolojileri en büyük teknik faydayı sağlamaktadır. Bilgiye erişim ve paylaşım için ise en fazla verimliliği bilgisayar ağları sağladığından, ağ işleyişinin düzgün olması hayati önem taşımaktadır. Ağ trafiğindeki anormallikler ise ağın gerektiği gibi kullanımını engelleyen unsurların başında gelmektedir. Bu anormallikler altyapı sorunlarından kaynaklanabileceği gibi ağın kötüye kullanılması veya ağa yapılan saldırılardan da kaynaklanabilmektedir [1].

Birçok kaynakta saldırı tespit sistemleri olarak da anılan anomali tespit sistemleri ağda oluşan düzensizlikleri tespit edip ilgili kişileri veya yazılımları uyarmayı sağlayan sistemlerdir.

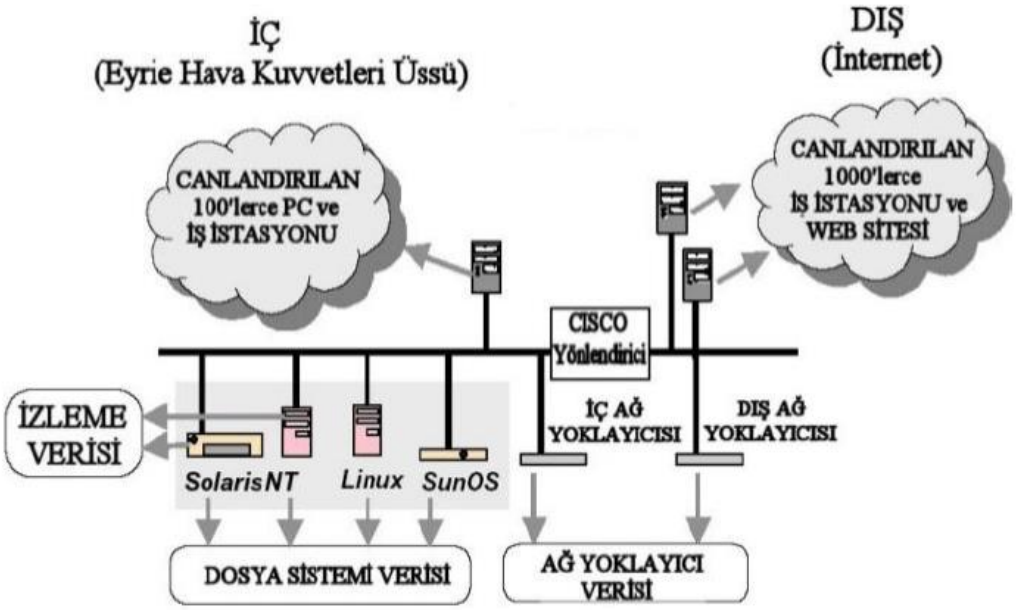
Bu çalışmada biz de büyük veri madenciliği tekniklerinden biri olan kümeleme tekniği algoritmalarından k-ortalamalar algoritması ile KDDCup99 veri seti üzerinde analizler yaparak olası anomalilerin tespitini gerçekleştirdik.

2. KULLANILAN ARAÇ VE YÖNTEMLER

2.1 Kullanılan Veri Seti

2.1.1 KDDCup99 Veri Setinin Detaylı Anlatımı

Saldırı Tespit Sistemleriyle ilgili çalışmalarda en sık kullanılan veri seti DARPA 1998 ve 1999 veri setleridir. Biz de model oluşturma çalışmamızda yine bu verilerden türetilen KDD Cup'99 veri setini kullanacağız. Veri setini oluşturan kaynak aşağıdaki Şekil6'da da görüldüğü gibi saldırının hedefi olan bir iç ağ ve saldırıyı gerçekleştiren bir dış ağ olmak üzere iki farklı ağdan oluşmaktadır:



Şekil 2.1 Darpa Ağ Yapısı [3]

KDD'99 veri kümesi DARPA veri kümesinin bazı önişlemlerden geçirilmesiyle elde edilen 41 özelliğe sahip bir veri kümesidir. Bu veri kümesinin amacı diğer veri kümelerinde olduğu gibi farklı anomali tespit sistemlerinin eğitim ve test işlemlerini kolaylıkla gerçekleştirmek ve bu sistemlerin değerlendirilmesini sağlamaktır. DARPA ile her ne kadar anomali tespit sistemleri için veri kümesi probleminin çözülmesi adına önemli adımlar atılmışsa da KDD'99 ile eğitim ve test sonuçları çok daha hızlı alınabilmesi ve buna bağlı olarak anomali tespit sistemlerinin değerlendirilmesinin daha kısa sürede yapılabilmesi adına bu alanda atılmış önemli bir adımdır [2].

KDD'99 veri kümesi, 9 temel ve 32 adet türetilmiş olmak üzere toplam 41 özellikten oluşmaktadır. Bu 41 özellik aşağıdaki gibi 3 ana gruba ayrılmıştır:

- İçerik özellikleri (content features)
- Sunucu tabanlı trafik özellikleri (host-based traffic features)
- Zamana bağlı trafik özellikleri (time-based traffic features)

Çizelge 1.1, Çizelge 1.2 ve Çizelge 1.3'de, sırasıyla bu 3 grup ve gruplara ait veri özellikleri gösterilmiştir.

Çizelge 1.1 İçerik Özellikleri

Özellik Adı	**Tanım**	**Tip**
Duration	Bağlantı Uzunluğu	Sürekli
Protocol_type	Protokol Tipi	Ayrık
Service	Servis tipi	Ayrık
Src_bytes	Kaynaktan hedefe veri	Sürekli
Dst_bytes	Veri bayt sayısı	Sürekli
Flag	Bayrak	Ayrık
Land	Kaynak ve Hedef ip aynı ise 1 değil ise 0	Ayrık
Wrong_fragment	Yanlış parçalama	Sürekli
Urgent	Acil Paket sayısı	Sürekli

İçerik özellikleri, sadece TCP bağlantılarıyla ilgili olan özelliklerdir. Diğer gruplarda olduğu gibi ağdaki veriler üzerinde herhangi bir önişlem yapılmasına gerek yoktur [2]. Sunucu tabanlı trafik özellikleri, etki alanı (domain) bilgisi ile ilgili içerik özellikleridir.

Çizelge 1.2 Sunucu tabanlı trafik özellikleri

Özellik Adı	**Tanım**	**Tip**
Hot	"hot" göstergesi	Sürekli
Num_failed_logins	Hatalı giriş sayısı	Sürekli
Logged_in	Giriş başarılı ise 1 değil ise 0	Ayrık
Num_compromised	Gizliliğin ihlal edilme sayısı	Sürekli
Root_shell	"Root Shell" elde edildiyse 1 değilse 0	Ayrık
Su_attempted	"Su root" komutu girildiyse 1 değilse 0	Ayrık

Num_root	“root” erişim sayısı	Sürekli
Num_file_creations	Dosya oluşturma işlemleri sayısı	Sürekli
Num_shells	Shell promptlarının sayısı	Sürekli
Num_access_files	Kontrol dosyalarına erişim işlemleri sayısı	Sürekli
Num_outbound_cmds	ftp oturumunda giden komut sayısı	Sürekli
İs_hot_login	Giriş “hot” listesindeyse 1 değilse 0	Ayrık
İs_guest_login	Giriş “guest” ise 1 değilse 0	Ayrık

Zamana bağlı trafik özellikleri, “aynı sunucu” ve “aynı servis” özelliklerine göre kullanılan özelliklerdir. “Aynı sunucu” özellikleri, son iki saniye içerisinde aynı sunucuya yapılan bağlantılardır, “aynı servis” özellikleri ise son iki saniye içerisinde aynı servise yapılan bağlantılardır [2].

Çizelge 1.3 Zamana bağlı trafik özellikleri

Özellik Adı	**Tanım**	**Tip**
Count	Aynı sunucuya önceki iki bağlantıyla aynı bağlantıların sayısı	Sürekli
Serror_rate	“SYN” hata bağlantılarının yüzdesi	Sürekli
Rerror_rate	“Rej” hata bağlantılarının yüzde	Sürekli
Same_srv_rate	Aynı servise bağlantıların yüzdesi	Sürekli
Diff_srv_rate	Farklı servislere bağlantıların yüzdesi	Sürekli
Srv_count	Aynı servise önceki iki bağlantıyla aynı bağlantıların sayısı	Sürekli
Srv_serror_rate	“SYN” hata bağlantılarının yüzdesi	Sürekli
Srv_rerror_rate	“Rej” hata bağlantılarının yüzdesi	Sürekli
Srv_diff_host_rate	Farklı servislere bağlantıların yüzdesi	Sürekli

2.1.2 Saldırı Tipleri

Bilgisayar sistemlerine yapılan saldırılar, birçok araştırmacı tarafından farklı gruplandırmalar yapılmıştır [4,5]. Saldırganların sürekli kendilerini yenilemeleri ve bilgisayar sistemlerindeki açıkları tespit etmeleri nedeniyle saldırı tiplerindeki

çeşitlilik çok fazla artmıştır. MIT Lincoln Laboratuvarlarında yapılan bu çalışmalar sonucunda bilgisayar sistemlerine yapılan saldırılar kullandıkları yöntemlere göre dört ana gruba ayrılmış ve **DoS, U2R, R2L** ve **Probing** olarak adlandırılmıştır [6]. Bu saldırıların detaylı tanımları aşağıdaki başlıklarda sunulmuştur.

2.1.2.1 Hizmet aksattırma saldırıları

Hizmet aksattırma saldırıları olarak da bilinen DoS saldırıları sistemin hizmetlerini engellemek amacıyla yapılırlar. Bunu yapmak için sisteme cevap verebileceğinden çok daha fazla istek gönderilerek verilen hizmet aksattırılır. DoS saldırılarına örnek olarak SYN flood, Smurf, UDPstorm, Pingflood, Neptune, Mailbomb saldırıları verilebilir [7].

2.1.2.2 U2R Saldırıları

U2R saldırıları sayesinde kullanıcılar normal yetkilere sahip olan kendi hesaplarından oturum açtıktan sonra yönetici yetkisine ulaşarak sistem üzerinde istedikleri bilgilere erişebilirler. En çok bilinen U2R saldırı tipleri Eject, Ffbconfig, Fdformat, Loadmodule, Perl gibi saldırılardır [7].

2.1.2.3 R2L Saldırıları

Bu saldırı tipinde saldırgan saldırdığı makineye ağ üzerinden paketler yollayarak açıkları tespit etmeye çalışır. Bu konuda birçok araç olması ve bu araçlara kolay erişilebilir olması nedeniyle, sistemde var olan açıklar kapatılmamışsa etkili ve kolay bir saldırı yöntemidir. En çok bilinen R2L saldırı tiplerine Dictionary, Guest, Imap, Named, Sendmail gibi saldırılar örnek olarak verilebilir [7].

2.1.2.4 Probing Saldırıları

Probing saldırıları, ağı veya bilgisayarı tarayarak zayıflıkları tespit etmek ve sistem yapısıyla ilgili genel bir bilgiye ulaĢmak için yapılmaktadır. Bu tip saldırılarda önce sistem hakkında detaylı bilgi edinildikten sonra saldırının nasıl olacağı belirlenir. Bu saldırılar için kullanılan araçlar aynı zamanda güvenlik uzmanları tarafından sistemin

güvenliğinin test edilmesi için de kullanılır. En çok bilinen Probe saldırı tipleri, Ipsweep, Mscan, Nmap, Saint, Satan gibi saldırılar örnek olarak verilebilir [7].

KDD'99 veri kümesi 38 farklı atak içermektedir. KDD'99 eğitim veri kümesinde bulunan 24 atak ve bu atakların ait oldukları saldırı tipleri ve saldırıları veri kümesinde bulunduran örnek sayıları Çizelge 1.4'de verilmiştir.

Çizelge 1.4 KDD'99 veri kümesinin %10'luk kısmındaki saldırı örneklerinin sayıları

Saldırı	**Örnek sayısı**	**Kategori**
Smurf.	280790	DoS
Neptune.	107201	DoS
Back.	2203	DoS
Teardrop.	979	DoS
Pod.	264	DoS
Land.	21	Dos
Normal.	97277	Normal
Satan.	1589	Probe
Ipsweep.	1247	Probe
Portsweep.	1040	Probe
Nmap.	231	Probe
Warezclient.	1020	R2L
Guess_passwd	53	R2L
Warezmaster	20	R2L
İmap.	12	R2L
ftp_write	8	R2L
Multihop.	7	R2L
Phf.	4	R2L
Spy.	2	R2L
Buffer_overflow	30	U2R
Rootkit.	10	U2R
Loadmodule	9	U2R
Perl	3	U2R

Sadece test veri kümesinde yer alan, eğitim kümesinde yer almayan ve etiketlenmiş KDD'99 dosyasından alınan 14 farklı atağa ait saldırı tipi ve örnek sayıları Çizelge 1.5'de sunulmuştur.

Çizelge 1.5 Eğitim kümesinde yer almayan test veri kümesinde bulunan saldırılar

Saldırı	Örnek Sayısı	Kategori
Apache	794	DoS
Mailbomb	5000	DoS
Processtable	759	DoS
Udpstorm	2	DoS
Mscan	1053	Probe
Saint	736	Probe
Httptunnel	138	R2L
Named	17	R2L
Sendmail	17	R2L
Snmpgetattack	1040	R2L
Xlock	9	R2L
Xsnoop	4	R2L
Ps	16	U2R

2.1.3 Veri Setinin Hazırlanması

Biz çalışmamızda "kddcup.data_10_ percent_corrected" dosya ismi ile internetten indirilebilen yaklaşık 75Mb büyüklüğünde ve içinde yaklaşık 500bin kayıt bulunan gerçek kddCupp-99 veri seti'nin 10%'una karşılık gelen veri setini kullandık. Veri setinde toplam 41 adet değişken bulunmaktadır. Biz aşağıdaki prensiplere uyarak bu sayıyı 9'a indirdik:

- Parametrelerin birbirlerinden bağımsız olanları seçilmiştir. Örneğin root_shell, su_attempted, num_root alanlarının tümü birden alınmak yerine su_attempted alanı alınmıştır.
- Parametrelerin bağımlı değişkeni etkilemeyecek olanları seçilmemiştir.
- Veriler daha kolay işlenebilmek için binary hale getirilmiştir.

2.2 Kullanılan Ortamlar

2.2.1 Apache Spark

Apache Spark Scala programlama dili ile yazılmış, JVM üzerinde çalışan, büyük ölçekli veri işleyebilen açık kaynaklı bir dağıtık hesaplama sistemidir. Hızlı ve in-memory çalışabilen bir veri işleme motorudur. İlk olarak 2009 da UC Berkeley'in AMPLab'ında geliştirilmiştir. Daha sonra 2010 da açık kaynaklı bir Apache projesi haline gelmiştir.

2.2.2 Databricks

Azure Databricks, Microsoft Azure bulut hizmetleri platformu için iyileştirilen Apache Spark tabanlı bir analiz platformudur. Apache Spark'ın kurucuları ile birlikte tasarlanan Databricks, tek tıklama ile kurulum olanağı ve kolaylaştırılmış iş akışlarının yanı sıra veri uzmanları, veri mühendisleri ve iş analistleri arasında işbirliği sağlayan etkileşimli bir çalışma alanı sunmak amacıyla Azure ile tümleştirilmiştir.

2.3 Kullanılan Diller

Bu çalışmada **Python** programlama dili kullanılmıştır. Python, nesne yönelimli, yorumlamalı, birimsel ve etikileşimli yüksek seviyeli bir programlama dilidir.

2.4 Kullanılan Algoritmalar

2.4.1 K-Means Kümeleme Algoritması

Kümeleme: Bir popülasyondaki bireylerin çoklu özelliklerinin nicel karşılaştırmalarını yaparak farklı gruplara girip girmediğini keşfetmek için yapılan istatistiksel bir sınıflandırma tekniğidir.

İlk olarak James MacQueen tarafından önerilen kMeans, bilinen ve yaygın olarak kullanılan kümeleme algoritmasıdır. k-Means, veri eğitiminde benzer örnek – nesne, nesne - nokta grubunu otomatik olarak tanımak için kullanılabilen makine öğrenmesindeki en basit kümeleme algoritmalarından biridir. Algoritma, örnekleri kullanıcı tarafından belirtilen önceden tanımlanmış sayıda küme sınıflandırır (örneğin, küme sayısını K olarak alır). İlk önemli adım k örneklerinin kümesini centroid (kümelenme merkezleri) olarak rasgele seçmektir, genellikle her küme için mümkün

olan bir merkez seçilir. Ardından algoritma, her örneği veri kümesinden okumaya ve en yakın kümeye atamaya devam eder. Örnek ile merkez arasındaki mesafeyi ölçmek için bazı yöntemler vardır, ancak en popüler olan Öklid uzaklığıdır. Küme merkezleri her örnek yerleştirildikten sonra daima yeniden hesaplanır. Bu işlem, daha fazla değişiklik yapılıncaya kadar yinelenir. K-Means algoritması aşağıdaki aşağıdaki sözde açıklanabilir:

1. Toplam küme sayısını belirle (K)
2. Rasgele bir küme seç ve kütle merkezi olarak ata
3. Öklid yöntemini kullanarak her bir örneğin tüm merkezlere olan uzaklığını hesapla
4. Her örneği en yakın kümeleme merkezine ata
5. Kütle merkezi konumlarını yeniden hesapla
6. 3-5. Arasındaki adımları kütle merkezi değişmeyene kadar tekrarla

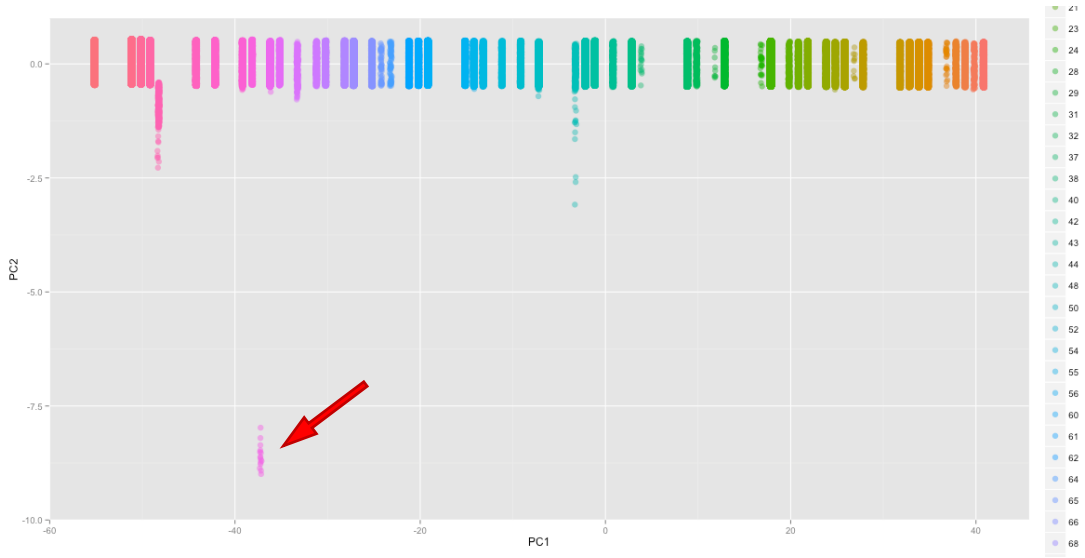
2.5 Kullanılan Yöntemler

2.5.1 PCA (Principal Component Analysis)

Bilgisayar bilimlerinde boyut indirmeye yarayan bir yöntemdir. Kısaca iki bilgi arasında bir bağlantı varsa bu bağlantı sayesinde iki veriden birisini ve bağlantıyı tutarak iki bilginin de geri bulunabilmesini sağlar. Kısaca PCA olarak da ifade edilen bu terime göre bir veri kümesinin (veri matrisinin, data matrix) kovaryans matrisinin (covariance matrix) veya tekil değer çıkarımı (singular value decomposition) yöntemi ile elde edilen basitleştirilmiş halidir.

3. SONUÇ VE DEĞERLENDİRME

Bu çalışmada, KDD Cup'99 veri kümeleri kullanılarak büyük veri madenciliğinin kümeleme tekniklerinden biri olan k-ortalamalar tekniği ile bilgisayar ağlarında alışılmışın dışında bir durum olup olmadığı hakkında analizler yapılmış ve bu analizlerin nasıl gerçekleştirildiği üzerinde durulmuştur. Şekil 3.1 de görülen ekran çıktısında KDD Cup'99 verilerinin işlendikten ve kümelendikten sonraki görülmektedir. Burada okla gösterilen kısım küme merkezinden çok daha uzakta olduğu için anomali olarak değerlendirilir.



Şekil 3.1 Kümeleme Algoritması Ekran Görüntüsü

Bu çalışmada da başvurulan KDD Cup'99 veri kümesi yaygın olarak şimdiki saldırı tespit sistemlerinde kullanılmakta ama bu 1999'un verisi olduğu için ve ağ teknolojileri ve saldırı metotları sürekli değiştiğinden bu, şimdiki gerçek ağ durumunu yansıtamayabilir. Bu yüzden gelecekte yapılacak çalışmalarda, günümüzdeki ağ durumunu daha doğru bir şekilde yansıtabilmek için daha yeni bilgiye sahip olunup, test edilip ve kıyaslama yapılması önerilmektedir.

Gerçekleştirilen çalışma sonucunda, büyük veri madenciliğinin kümeleme yöntemlerinden biri olan k-ortalamalar hakkında geniş bir bilgi birikimi elde edilmiştir.

Bu çalışmada karşılaşılan güçlükler;

- Toplanan verilerin işlenmesinin uzun sürmesi

- Güncel saldırı verilerinin ücretsiz olmaması
- Algoritmalar öğrenen algoritmalar olduğu için büyük verilere ihtiyaç duyulması

şeklinde sıralanabilir.

REFERANSLAR

1. Budak, İ, Şen, B ve Yıldırım, M.Z., "Lojistik regresyon ile bilgisayar ağlarında anomali tespiti", XIII. Akademik Bilişim Konferansı, Antalya (2013).
2. Güven, E. N., "Zeki saldırı sistemlerinin incelenmesi, tasarımı ve gerçekleĢtirilmesi", Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, 1-91 (2007).
3. M. A. Aydın, "Bilgisayar Ağlarında Saldırı Tespiti için İstatistiksel Yöntem Kullanılması", İTÜ Yüksek Lisans Tezi, 2005.
4. Cabrera, B.D., Cabrera, L. Lewis and R.K. Mehra, "Detection and classification of intrusions and faults using sequence of system calls", ACM SIGMOD record, 30(4): 25-34 (2001).
5. Bace, R., Mell, P., "Intrusion detection systems", Technical Report, National Institute of Standards and Technology, NIST SP300-31, Scotts Valley, CA, 5-46 (2001).
6. İnternet: Massachusetts Teknoloji Enstitüsü Lincoln Laboratuarları "Off-line intrusion detection evaluation data" http://www.ll.mit.edu/IST/ideval/ (2009).
7. Mukkamala, S., Janoski, G., Sung, A., "Intrusion detection using neural networks and support vector machines", IEEE International Joint Conference on Neural Networks, IEEE Computer Society Press, 1702-1707 (2002).