Available online at www.jpit.az

13 (1)
2022

Experimental Study of Machine Learning Methods in Anomaly Detection

Makrufa Sh. Hajirahimova [a], Leyla R. Yusifova [b]

[a,b,] Institute of Information Technology, Azerbaijan National Academy of Sciences, B. Vahabzade str., 9A, AZ1141 Baku, Azerbaijan

[a]hmakrufa@gmail.com; [b]yusifova863@gmail.com

ARTICLE INFO

http://doi.org/10.25045/jpit.v13.i1.02

Article history:
Received 9 September 2021
Received in revised form 9 November 2021
Accepted 5 January 2022

Keywords:
Big data
Anomaly
DoS attacks
IDS
Machine learning
Ensemble classification

ABSTRACT

Recently, the widespread usage of computer networks has led to the increase of network threats and attacks. Existing security systems and devices are insufficient in the detection of intruders' attacks on network infrastructure, and they considered to be outdated for storing and analyzing large network traffic data in terms of size, speed, and diversity. Detection of anomalies in network traffic data is one of the most important issues in providing network security. In the paper, we investigate the possibility of using machine learning algorithms in the detection of anomalies – DoS attacks in computer network traffic data on the WEKA software platform. Ensemble model consisting of several unsupervised classification algorithms has been proposed to increase the efficiency of classification algorithms. The effectiveness of the proposed model was studied using the NSL-KDD database. The proposed approach showed a higher accuracy in the detection of anomalies compared to the results shown by the classification algorithms separately.

Anomaliyaların aşkarlanmasında maşın təlimi metodlarının eksperimental tədqiqi

Açar sözlər :
Big data
Anomaliya
DoS hücum
IDS
Maşın təlimi
Klassifikasiya ansamblı

Son zamanlar kompüter şəbəkələrindən geniş istifadə şəbəkə təhdidləri və hücumlarının artmasına səbəb olmuşdur. Mövcud təhlükəsizlik sistemləri və alətləri isə hücumçuların şəbəkə infrastrukturuna olan hücumlarını aşkarlamaqda yetərli deyildir, ölçü, sürət və müxtəliflik baxımından böyük şəbəkə trafiki verilənlərinin saxlanması və analizi üçün köhnəlmiş hesab olunur. Şəbəkə trafiki verilənlərində anomaliyaların aşkarlanması şəbəkə təhlükəsizliyinin təmin edilməsində çox vacib məsələlərdəndir, həmçinin elmi tədqiqatların əsas istiqamətlərindəndir. Şəbəkə trafikində anomaliyaların aşkarlanması sahəsində kifayət qədər tədqiqatların olmasına baxmayaraq, daha dəqiq aşkarlama modellərinin işlənməsinə ehtiyac vardır. Məqalədə anomaliyaların aşkarlanmasında istifadə olunan bəzi maşın təlimi alqoritmləri analiz olunmuşdur. Kompüter şəbəkə trafiki verilənlərində anomaliyaların - DoS hücumların aşkarlanmasında maşın təlimi alqoritmlərinin istifadəsinin mümkünlüyü WEKA proqram platformasında eksperimental olaraq tədqiq edilmişdir. Təsnifatlandırma alqoritmlərinin effektivliyini artırmaq məqsədi ilə bir neçə klassifikasiya alqoritmlərindən təşkil olunmuş ansambl modeli təklif edilmişdir. Təklif olunan modelin effektivliyi NSL-KDD verilənlər bazası istifadə edilərək tədqiq olunmuşdur. Klassifikasiya alqoritmlərinin ayrı-ayrılıqda göstərdiyi nəticələrlə müqayisədə təklif olunan yanaşma anomaliyanın aşkarlanmasında daha yüksək dəqiqlik nümayiş etdirmişdir.

Экспериментальное исследование методов машинного обучения при обнаружении аномалий

Ключевые слова:
большие данные
аномалия
DoS-атак
IDS
машинное обучение
ансамбль классификации

В последнее время широкое использование компьютерных сетей привело к увеличению сетевых угроз и атак. Существующих систем и инструментов безопасности недостаточно для обнаружения атак злоумышленников на сетевую инфраструктуру, кроме того, они считаются устаревшими для хранения и анализа больших данных сетевого трафика с точки зрения размера, скорости и разнообразия. Обнаружение аномалий в данных сетевого трафика - одна из важнейших задач обеспечения сетевой безопасности, а также одно из основных направлений научных исследований. Несмотря на то, что в области обнаружения аномалий в сетевом трафике проведено значительное количество исследований, необходимо разработать более точные модели обнаружения. В статье анализируются некоторые алгоритмы машинного обучения, используемые для обнаружения аномалий. Возможность использования алгоритмов машинного обучения при обнаружении аномалий в данных трафика компьютерной сети - DoS-атак была экспериментально исследована на программной платформе WEKA. Для повышения эффективности алгоритмов классификации предложена ансамблевая модель, состоящая из нескольких алгоритмов классификации. Эффективность предложенного метода проанализирована с использованием базы данных NSL-KDD. Предложенный подход показал более высокую точность обнаружения аномалий по сравнению с результатами, показанными алгоритмами классификации при их работе в отдельности.

1. Giriş

İnternet texnologiyalarının genişlənməsi ilə kiberhücumlar daha da intensivləşmişdir. Kiberhücumların intensivliyi ənənəvi siqnatur əsaslı təhlükəsizlik vasitələrini yeni hücum növlərinə qarşı təsirsiz etmişdir (Garofalo, 2017; Hacırəhimova, 2014). Belə ki, Big data erasında şəbəkə infrastrukturuna bədəməllər tərəfindən müxtəlif növ kiberhücumlar (məsələn, spam göndərmə, botnetlər, xidmətdən imtina (DoS - Denial of Service), paylanmış xidmətdən imtina (DDoS), fişinq, zərərli proqram–viruslar və s.) nəticəsində hökumət, enerji, səhiyyə, bank və telekommunikasiya, təhsil, nəqliyyat, tədqiqat mərkəzləri və s. kimi kritik sektorlarda əksər təşkilatlar ciddi problemlərlə üz-üzə qalmışlar (Heydari et al., 2015; Almedia, 2017; Wang et al., 2018). Bu təşkilatlar müxtəlif monitorinq vasitələrindən istifadə edərək infrastrukturlarını qorumaq üçün böyük miqdarda vəsait xərcləyirlər. Lakin hücumçuların infrastruktura nüfuz etmək üçün qabaqcıl vasitələrdən istifadə etdiklərindən onları aşkar etməkdə mövcud təhlükəsizlik və loq-faylların analizi alətləri köhnəlmiş hesab edilir (Ariyaluran et al., 2019). Şəbəkədəki potensial təhlükəni aşkar etmək üçün toplanan verilənlərin anında emalı isə vacibdir. Mövcud ənənəvi monitorinq alətlərinin böyük verilənləri emal etmək imkanının olmaması səbəbindən şəbəkə infrastrukturunu davamlı izləmək və anormal davranış və təhdidləri aşkarlamaq mümkün deyildir (Raguseo, 2018).

Anomaliya, gözlənilən davranışa uyğun olmayan verilən nümunəsi olaraq təyin edilir. Anomaliyanın aşkarlanması 1987-ci ildə D.E.Denning tərəfindən təklif edilmişdir (Denning, 1987). Bu metodun əsas konsepsiyası, əvvəlcədən təyin edilmiş davranışın normal davranışla müqayisə edildiyi şəbəkənin/sistemin davranışını təyin etməkdir. Nəticə ya onu qəbul etmək, ya da əlavə araşdırma üçün həyəcan idarəetmə sistemini işə salmaqdan ibarət olacaq. Anomaliyanın aşkarlanması verilənlərin əsas hissəsindən fərqlənən və ya kənara çıxan (outlier) verilənləri (nümunələri, şablonları) aşkar etməyi hədəfləyir (Abdulhammed, 2019). Yəni anomaliyanın aşkarlanması, verilənlər bazasında mütəxəssis tərəfindən identifikasiya edilə bilməyən, gözlənilən nümunəyə uyğun gəlməyən elementlərin və ya hadisələrin müəyyən edilməsidir.

Böyük ölçü anomaliyanın aşkarlanması üçün ciddi çətinliklər yaradır. Çünki dəyişənlərin və ya əlamətlərin sayı artdıqca verilənlərin miqdarı da artır. Cisco-nun məlumatına görə illik qlobal IP trafikin 2021-ci ildə 3.3 zettabayta çatmışdır (Global - VNI Complete Forecast Highlights, 2021). Bu hal verilənlərin təhlükəsizliyi, məlumat ötürülməsi etibarlılığı və trafik anomaliyaları və hücumlarının aşkarlanması baxımından bir çox çətinliklər yaradır. Böyük verilənlərdə anomaliyaları aşkar etmək üçün ənənəvi metod/alqoritmlər kifayət qədər dəqiqliyi qoruya bilmir (Srikanth, et al., 2020; Rehman, 2016). Son illərdə anomaliyaların aşkarlanması, maşın təlimi sahəsində əsas tədqiqat mövzusuna çevrilmiş və çoxsaylı məqalələrin predmeti olmuşdur (Dua, Du, 2011; Buczak & Guven, 2016). Kompüter şəbəkəsində anomaliyalar şəbəkə trafikində qeyri-adi və əhəmiyyətli dəyişikliklər kimi başa düşülür (Garofalo, 2017). Verilənlərin daxil edilməsi zamanı insanlar tərəfindən edilən səhvlər, ölçmə cihazlarının səhvləri, verilənlərin emalı səhvləri (verilənlərlə manipulyasiya), sistemin davranışı zamanı yaranan mexaniki səhvlər və s. anomaliyaları yaradan geniş yayılmış səbəblərdəndir.

Anomaliyaları aşkarlama yanaşmasının əsas məqsədi normal trafiki təsvir etmək üçün statistik bir model qurmaqdır. Bu modeldən hər hansı bir kənara çıxmaya anomal hadisə kimi baxıla bilər və hücum kimi qəbul edilə bilər. Qeyd olunduğu kimi anomaliya, mücərrəd səviyyədə verilənlərin əksəriyyətindən fərqlənən nümunələr kimi başa düşülür. Ədəbiyyatda anomaliyaların üç tipini qeyd edirlər: nöqtə anomaliyaları (*point anomalies*), kontekstual anomaliyalar (*contextual anomalies*), kollektiv anomaliyalar (*collective anomalies*). Bu tip anomaliyalar bir çox tədqiqatlarda geniş analiz olunmuşdur (Hodge, Austin, 2004; Chandola, Banerjee, Kumar, 2009; Nassif et al., 2021; Gogoi et al., 2011).

Uzun illər öyrənilən və elmi tədqiqatların predmetinə çevrilmiş anomaliyaların aşkarlanması müxtəlif sahələrdə tətbiq olunmaqdadır. Kredit kartlar üzrə tranzaksiya verilənlərində fırıldaqçılığın (Xuan et al., 2018; Husejinović, 2020; Phua et al., 2010; Dash, & Ng, 2010; Chaudhary, Yadav., & Mallick, 2012), kompüter şəbəkələrində anomal trafiklərin

(Aliguliyev, Hajirahimova, 2019; Shon, Moon, 2007; Alguliyev, Aliguliyev, Imamverdiyev, Sukhostat, 2017; Zhang et al., 2008), şübhəli kiber aktivliyin (Ariyaluran et al., 2019), tibbi verilənlərdə anomaliyaların aşkarlanması (məs., anomal MRT təsviri və s.) (He, Wang, Graco, Hawkins, 1997; Saneja, Rani, 2017; Antal, Hajdu, 2014; Schlegl, Seeböck et al., 2017; Varian, 2020) sensor qurğuların anomal göstəricilərinin (məs., kosmik aparat sensorunun anormal göstəriciləri kosmik gəminin bəzi komponentlərində nasazlığın olduğunu göstərə bilər) (Fujimaki et al., 2005) və s. aşkarlanması Big Data analitikanın əsas məsələlərindəndir (Chandola, Banerjee, Kumar, 2005; Wang, Jones, 2017). Bu baxımdan böyük verilənlərdə şablonlara uyğun olmayan verilənlərin (və ya əvvəllər müşahidə olunmayan verilənlərin) – anomaliyaların aşkarlanması məsələsi böyük aktuallıq kəsb edir, onun qarşısını almaq xüsusi əhəmiyyət daşıyır. Şəbəkə anomaliyası, şəbəkənin təhlükəsizliyinə təsir göstərən potensial zərərli trafikdir. Bunlar həm dövlət, həm də korporativ qurumları narahat edən əsas problemlərdir, çünki maliyyə itkilərinə, şəbəkənin işinin pozulmasına səbəb ola bilər və hətta milli təhlükəsizliyi təhlükə altına ala bilər (Ariyaluran, 2019).

Bu tədqiqatın əsas məqsədi şəbəkə trafikində anomaliyaları – DoS hücumların aşkarlanmasından ibarətdir. DoS hücumlar şəbəkə, kiber infrastruktur üçün real təhlükədir (Denning, 1987; Garofalo, 2017). DoS hücumlar şəbəkə və xidmət serverlərdən, əlaqələndirici şəbəkələrdən və şəbəkə cihazlarından (marşrutlaşdırıcılar və s.) istifadə edərək xidməti iflic edə və ya ondan tamamilə imtinaya səbəb ola bilər və bu da böyük itkilərlə nəticələnə bilər. DoS hücumun reallaşdırılmasında hədəf hostun qanuni istifadəçiyə cavabını ləngitmək və ya tamamilə itirmək üçün rabitə protokollarının zəifliklərindən istifadə edilir. Bu zərərli fəaliyyət "qurban" serverinə çox sayda sorğu göndərməklə həyata keçirilir.

Məqalədə böyük şəbəkə verilənlərində anomaliyaların – DoS hücumların aşkarlanmasında bir neçə maşın təlimi alqoritmləri NSL-KDD verilənlər bazası (NSL-KDD) üzərində test edilmiş və klassifikasiyanın effektivliyini artırmaq məqsədi ilə klassifikasiya ansamblı modeli təklif edilmişdir.

Məqalənin sonrakı hissəsi aşağıdakı kimi təşkil edilmişdir. İkinci bölmədə əlaqəli işlər icmal olunmuşdur. Üçüncü bölmədə təklif olunan yanaşmanın metodologiyası təsvir edilmişdir. Dördüncü bölmədə təklif edilmiş yanaşma eksperimental olaraq yoxlanılmış və ekperimentin nəticələri müzakirə olunmuşdur. Beşinci bölmədə tədqiqat üzrə nəticələr ümumiləşdirilmişdir.

2. Əlaqəli tədqiqatlar

Bu bölmədə anomaliyaların aşkarlanması sahəsində mövcud tədqiqatların müasir vəziyyəti icmal olunur. Qeyd etmək lazımdır ki, anomaliyaların aşkarlanması istiqamətində həm icmal xarakterli (Hodge, Austin, 2004; Agrawal, Agraüal, 2015; Chandola, Banerjee, Kumar, 2005; Srikanth, Philip, Jiong et al., 2020; Gupta, Gao, Aggarwal, Han, 2014; Nassif et al., 2021; Patcha, Park, 2007), həm də çoxsaylı yeni tədqiqatlar mövcuddur (Ariyaluran, 2019; Aliguliyev, Hajirahimova, 2019); Sukhostat et al, 2018; Akoglu et al., 2015; Wei et al., 2019; Aggarval, 2005).

Uzun illər müdaxilələrin aşkarlanması sistemləri (İntrusion Detection Systems- IDS) şəbəkə administratorlarına çox dəstək olmuşdur (Akbar, 2010). Müdaxilələrin aşkarlanması sahəsində iki yanaşma müşahidə olunmaqdadır: siqnatura əsaslanan sui-istifadənin aşkarlanması (*misuse detection*) və anomaliyaların aşkarlanması (*anomaly detection*) (Mukherjee, Heberline, Levitt, 1994). Sui-istifadənin aşkarlanmasının əsas ideyası hücumları şablonlar və ya siqnatur şəklində təqdim etməkdir.

Məlumdur ki, IDS-lərin çoxu təhlükəsizlik ekspertləri tərəfindən təyin edilən çoxsaylı qaydalara əsaslanır (Tsai et al., 2009). Məlum olmayan yeni hücumların aşkarlanmasının çətin olması bu yanaşmaların əsas çatışmayan cəhətlərindəndir. Eyni zamanda şəbəkə trafikinin həcmi böyük olduğundan, qaydaların kodlaşdırılması ləng gedir, çox vaxt sərf olunur, ekspert biliklərindən asılılıq yaranır və s. Anomaliya aşkarlama üsulları isə yüksək səviyyədə saxta həyəcan siqnalları istehsal edir (Zhang et al., 2008). IDS-lərin bu kimi məhdudiyyətlərini aşmaq və yeni müdaxilələri, normal və anomal şəbəkə trafikini daha dəqiq aşkar etmək üçün intellektual analiz metodları tətbiq edilmişdir (Lee, Stolfo, Mok, 2000; Agarwal, Mittal, 2012).

Paylanmaya, yaxınlıq ölçüsünə əsaslanan, parametrik, qeyri parametrik statistik metodlar anomaliyaların aşkarlanmasında ilkin metodlardan hesab olunur (Chandola, Banerjee, Kumar, 2005). Statistik metodlara əsaslanan anomaliya aşkarlama sistemləri iki mərhələdən təşkil olunur. Əvvəlcə sistem şəbəkə trafikinin bir və ya bir neçə statistik əlamətini müşahidə edir və toplayır, daha sonra davranış dəyişikliklərini aşkar etmək üçün stoxostik bir metoddan istifadə edərək mövcud vəziyyəti saxlanılan vəziyyətlə müqayisə edir. Burada əsas məsələ tez-tez hədəf hostlarda yerləşdirilən botlar vasitəsilə həyata keçirilən DoS, fişinq və spam kimi müxtəlif növ zərərli proqramların düzgün aşkarlanmasıdır. Bu problemlə üzləşməmək məqsədi ilə Ensemble Empirical Mode Decomposition alqoritmindən istifadə edən bir anomaliya aşkarlama metodu təklif edilmişdir (Marnerides et al., 2015).

Son zamanlar isə anomaliyaların dəqiq identifikasiya olunmasında maşın təlimi (*machine learning -ML*) və dərin təlim (*deep learning*) metodları geniş tətbiq olunmaqdadır (Goldtein, Abdulhammed, Buczak, 2016). ML, 1959-cu ildə Arthur Samuel tərəfindən "açıq şəkildə proqramlaşdırma olmadan kompüterlərin öyrənməsini təmin edən bir təlim sahəsi" olaraq müəyyən edilmişdir (Samuel, 1959). Yəni ML, açıq şəkildə proqramlaşdırma əvəzinə seçilmiş nümunə verilənlərə (və ya keçmiş təcrübəyə) əsaslanan iterativ öyrənmə yolu ilə gizli korrelyasiya qanunauyğunluqlarını aşkar etməyə imkan verir. Qeyd etmək lazımdır ki, ML metodlarının təlimə əsaslan, təlimsiz və yarımtəlimli tipləri mövcuddur (Chandola, Banerjee, Kumar, 2005; Schlegl, et al., 2017).

Zhang və digərləri IDS-lərin problemlərini həll etmək üçün, qaydalara əsaslanan (ruled-based), xüsusilə yeni müdaxilələri aşkar etmək üçün təlimsiz təsadüfi meşələr (random forests) maşın təlimi alqoritmi tətbiq etmişlər (Zhang et al., 2008).

Camacho və digərləri məhkəmə sistemləri şəbəkəsindəki anomaliyaları aşkar etmək, qeyri-adi davranışları müəyyənləşdirmək və şərh etmək üçün böyük verilənlərin dörd əlamətini (həcm, müxtəliflik, etibarlılıq və sürət) nəzərə alaraq bir yanaşma təqdim etmişlər (Camacho et al., 2014). Həm qeyri-müəyyənliyi (aşağı doğruluq), həm də yüksək ölçü problemini aradan qaldırmaq üçün Əsas Komponent Analizi (Principal Component Analysis - PCA) tətbiq olunmuşdur. Müəlliflər PCA-nın verilənlər bazalarındakı ölçüyə (müşahidələrin sayına) görə hesablama probleminin qarşısını alan və paralelliyə imkan verən nüvə hesablamasından istifadə etmişlər. Ölçü həddindən artıq olduqda iyerarxik modellər də təklif olunmuşdur. Nəhayət, verilənlər axınlarını təhlil edərkən yüksək sürəti təmin etmək üçün Exponentially Weighted Moving Average (EWMA) yanaşması tətbiq olunmuşdur.

Anomaliyaları aşkarlamaq üçün hibrid metodlar da təklif olunmuşdur (Agarwal, Mittal, 2012; Kim et al., 2014; Shon, Moon, 2007). *B.Agarwal* və həmkarı şəbəkə əlamətləri entropiyasının və dayaq vektorları metodu (Support Vector Machine - SVM) alqoritmlərinin birləşməsindən ibarət hibrid metod təklif etmişlər (Agarwal, Mittal, 2012). Nəticədə entropiya əsaslı aşkarlama metodunun şəbəkədəki anomaliyaları SVM əsaslı aşkarlama sistemindən daha yaxşı müəyyən etdiyi sübut edilmişdir.

Küylü verilənlərin təsirini minimuma endirməklə, daha böyük verilən qruplarında aşkarlama dəqiqliyini yaxşılaşdırmaq üçün MSD (Mean and Standard Deviation) statistik metodu və k-means öyrənmə metodunun kombinasiyasından ibarət yeni bir aşkarlama alqoritmi (MSD - k-means) təklif edilmişdir (Wei et al., 2019). MSD-k-means-də iki mərhələ var: 1) verilənlərdə səs-küylü verilənləri aradan qaldırmaq üçün MSD alqoritminin tətbiqi; 2) lokal optimal klasterlər əldə etmək üçün k-means alqoritminin tətbiqi. MSD-k-means-də digər metodlarla müqayisədə daha yüksək dəqiqlik əldə olunmuşdur.

Şəbəkə anomaliyalarını aşkarlaya bilən k-means klasterləşmə alqoritminə əsaslanan yanaşmalar da təklif edilmişdir (Münz, 2007), (Kumari et al., 2016). Münz şəbəkə axın yazılarını ehtiva edən təlim verilənlərini normal və anormal trafik qruplarına ayırmaq üçün k-means alqoritmini tətbiq etmişdir. Kumari və həmkarları Apache Spark analitik alətindən istifadə edərək k-means klasterləşmə əsasında kiber-hücumların qarşısını almaq üçün metodologiya təqdim etmişlər.

Məlumdur ki, hesablama sistemləri ölçüsü, sürəti və müxtəlifliyinə görə böyük verilənlərin saxlanması, emalı və analizində məhdudiyyətlərə malikdir. Bu məhdudiyyəti aradan qaldırmaq

məqsədi ilə anomaliyanın aşkarlanması üçün çəkili klasterləşməyə əsaslanan optimallaşdırma yanaşması təklif edilmişdir (Alguliyev, Aliguliyev, İmamverdiyev, Sukhostat, 2018; Alguliyev, Aliguliyev, İmamverdiyev, Sukhostat, 2017). Hər bir nöqtənin çəkisi onun bütün verilənlər toplusundakı mərkəzə nəzərən mövqeyinə görə təyin edilmişdir. Yeddi böyük ölçülü verilənlər bazasından istifadə etməklə aparılmış eksperimentlərdə təklif olunan çəkili klasterləşmə alqoritmi k-means alqoritmi ilə müqayisədə anomaliyaları daha dəqiq aşkarlamağa müvəffəq olmuşdur. Artan hesablama mürəkkəbliyi və verilənlərin müxtəlifliyi səbəbindən hər növ anomaliya üçün bir vasitə seçmək çətindir. İkinci tədqiqatda isə klasterləşmə üçün təkmilləşdirilmiş optimallaşdırma yanaşması təklif edilmişdir (Alguliyev, Aliguliyev, İmamverdiyev, Sukhostat, 2017). Üç verilənlər bazası (Avstraliya kredit kart tətbiqləri, ürək xəstəlikləri üzrə baza və NSL-KDD bazası) üzərində aparılan təcrübə nəticələri təklif olunan alqoritmin k-means alqoritmi ilə müqayisədə anomaliyaları daha dəqiq aşkarladığını göstərmişdir.

Maliyyə sektoru fırıldaqçılığın aşkarlanması (Chaudhary, 2012; Phua et al., 2010) və tranzaksiya əməliyyatlarının işlənməsi (Dash & Ng, 2010) daxil olmaqla böyük verilənlərin analitikasından aktiv şəkildə istifadə edir. Saxta sui-istifadə hallarında artan kart ödəmə problemi bank ödənişləri və ödəniş xidməti təminatçılarının diqqət mərkəzində durur. Belə ki, tez-tez kredit kartı fırıldaqçılığı hadisələri baş verir və sonu böyük maliyyə itkiləri ilə nəticələnir. Cinayətkarlar digər insanların kredit kartlarının məlumatlarını oğurlamaq üçün Troyan və ya fişinq kimi bəzi texnologiyalardan istifadə edirlər (Xuan et al., 2018; Husejinović, 2020). Normal və anormal əməliyyatların davranış əlamətlərini öyrətmək üçün Xuan və həmkarları təsadüfi meşə alqoritmindən istifadə etmişlər (Xuan et al., 2018). *Husejinović* isə öz tədqiqatında fırıldaqçılıq əməliyyatlarının nəticələrini proqnozlaşdırmaq üçün Naive Bayes, C4.5 qərar ağacı və Bagging ansambl maşın öyrənmə alqoritmlərindən istifadə etmişdir. Fırıldaqçılıq əməliyyatlarının proqnozlaşdırılmasında C4.5 qərar ağacı alqoritmi digərləri ilə müqayisədə daha dəqiq proqnoz (92.74%) sərgiləmişdir.

Başqa bir tədqiqatda DoS hücumların aşkarlanması üçün məhdud Boltsmann maşınına (restricted Boltzmann machine - RBM) əsaslanan dərin təlim metodu tətbiq olunmuşdur (Imamverdiyev, Abdullayeva, 2018). DoS hücumu aşkarlama dəqiqliyini artırmaq üçün RBM-in görünən və gizli qatları arasına yeddi qat əlavə olunmuşdur. DoS hücumun aşkarlanmasında dəqiq nəticələr təklif olunan dərin RBM modelinin hiperparametrlərinin optimallaşdırılması ilə əldə edilmişdir.

3. Təklif olunan yanaşma

Məsələnin qoyuluşu. Fərz edək ki, şəbəkə trafikini əks etdirən $D = \{x_1, x_2, ..., x_n\}$ verilənlər bazası verilmiş və $M = \{m_1, m_2, ..., m_k\}$ sayda klassifikasiya alqoritmləri seçilmişdir. Şəbəkədə zərərli trafiklərin (DoS hücumların) yüksək dəqiqliklə aşkarlanması tələb olunur.

Tədqiqatın metodologiyası. IDS-lərin işinin səmərəliliyinin artırılması üçün kompüter şəbəkəsi trafikində anomaliyaları – DoS hücumları aşkarlamağa imkan verən klassifikasiya ansamblı modeli təklif edilir. Maşın təlimi ansamblı modelini yatarmaq üçün sadəlövh bayes (NaiveBayes – NB), qərar ağacı (Decision tree – DT), təsadüfi meşə (Random Forest RF), Support Vector Machine – SVM, çoxlaylı perseptron (Multilayer Perceptron), K-ən yaxın qonşu (K-Nearest Neighbor – KNN) kimi baza klassifikatorlar seçilmişdir. Bu klassifikatorların qısaca şərhini verək:

Naive Bayes modelində obyektin hansı sinfə aid olması ehtimalı Bayes teoreminə əsasən hesablanır (Yang et al., 2016). Bayes teoreminin mahiyyətində şərti ehtimal dayanır. Şərti ehtimal – bir hadisənin baş verdiyi nəzərə alınmaqla digər bir hadisənin baş vermə ehtimalı kimi başa düşülür. Naive Bayes klassifikatoru istifadə etdiyi n sayda əlamətlərin şərti olaraq bir-birindən asılı olmadığını nəzərdə tutur. Başlanğıcda verilən hər bir əlamətin şərti ehtimalı hesablanır, sonra nümunənin sinif etiketini proqnozlaşdırmaq üçün Bayes teoremi tətbiq olunur. Bayes teoreminin riyazi ifadəsi düstur 1-də verilmişdir.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \qquad (1)$$

Burada, P (A) – A hadisəsinin baş vermə ehtimalıdır. P (A | B) – B baş verdiyi halda A-nın başvermə ehtimalıdır. P(B|A) – A baş verdiyi

halda B-nin başvermə ehtimalıdır. P(B) – B hadisəsinin başvermə ehtimalıdır.

Qərar ağacı (decision tree - DT) keçən əsrin 80-ci illərində təklif olunmuş ağac strukturlu klassifikatordur, kök, budaqlar və yarpaqlardan ibarətdir. Kök – ağacın başladığı yerdir. O, bütün verilənlər bazasını göstərir və sonradan iki və daha çox budağa ayrılır. Budaqlar verilənlər bazasındakı əlamətləri, yarpaqlar sinif nişanlarını göstərir. Qərar ağacı sual/cavab əsasında budaqlara bölünür. Qərar ağacı təsnifat və reqressiya təhlili üçün qeyri-parametrik təlim metodudur. Qərar ağacı iki mərhələdən – öyrənmə və proqnozlaşdırma mərhələlərindən ibarətdir. Model təlim prosesində verilmiş təlim verilənləri istifadə edərək öyrədilir və proqnoz mərhələsində göstərilən test verilənlərinin nəticəsini proqnozlaşdırmaq üçün istifadə olunur (Dewan et al., 2010). Qeyd etmək lazımdır ki, qərar ağacının realizasiyası zamanı əsas məsələlərdən biri də kök qovşaq və ya altqovşaqlar üçün vacib atributun seçilməsidir. Problemin həllində iki geniş yayılmış atribut seçmə metodu mövcuddur: Information Gain; Gini Index.

Təsadüfi meşələr (Random Forests – RF) 2000-ci illərin əvvəllərində L.Breyman tərəfindən təklif edilmiş, qərar ağacları ansamblı əsasında qurulur (Breyman - 2001). Təsadüfi adlanmasının səbəbi odur ki, ixtiyari ağacın qurulmasında bütün əlamətlər deyil, ancaq təsadüfi seçilmişlər iştirak edir. Yəni RF verilənlərin təsadüfi altçoxluğundan təşkil olunmuş bir neçə qərar ağacı yaradır. Məqsəd klassifikasiyanın dəqiqliyini artırmaqdır. Göründüyü kimi alqoritm iki mərhələdə aparılır: əlamətlərin seçilməsi və təsnifatlandırma. Obyekt daha çox ağacın səs verdiyi sinfə aid edilir. Yəni yekun qərar ağacların qərarlarının aqreqasiyası əsasında yaradılır, mojaritar səsvermə ilə müəyyən edilir. RF DoS hücumlarla yanaşı Probe, U2R & R2L kimi hücumların, eyni zamanda botnetlərin yüksək dəqiqliklə aşkarlanmasında da istifadə olunur (Farnaaz & Jabbar, 2016). Qeyd etmək lazımdır ki, böyük verilənlərlə işləyərkən RF böyük hesablama zamanı tələb edir.

K-ən yaxın qonşu (K-Nearest Neighbor – KNN) – sistemə daxil olan yeni obyektin sinfini təyin etmək üçün istifadə edilən maşın təlimi alqoritmidir. Daxil olan hər yeni obyekt üçün siniflərdən birinə aid olan k sayda yaxın obyekt (qonşu) müəyyən edilir. KNN alqoritmi, ümumiyyətlə təsnifat və reqressiya problemləri üçün istifadə olunan qeyri-parametrik üsuldur. (Zhang, Zhou, 2015; Wauters et al., 2017). Bu alqoritmdə K parametrinin düzgün seçilməsi və yaxınlıq metrikası probleminin həlli önəmlidir. Həndəsi olaraq obyektlər arasındakı məsafəni ölçmək üçün, ədəbiyyatlarda müxtəlif məsafə funksiyaları (Evklid məsafəsi, Minkovski, kosinus məsafəsi) istifadə edilməkdədir. Ən çox istifadə edilən Evklid məsafəsi funksiyasının (*düstur 2*) istifadə edildiyi halda yeni obyektdən bu sinfə aid olan k sayda obyektə olan məsafələrin cəmi hesablanır. Yeni obyekt ən kiçik məsafəyə malik olan sinfə aid edilir.

$$dist(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \qquad (2)$$

Burada x və y – m sayda əlamətə malik giriş vektorlarıdır. Qeyd etmək lazımdır ki, fərqli əlamətlər, ümumiyyətlə fərqli ölçülərlə ölçüldüyündən məsafə funksiyasını tətbiq etməzdən əvvəl normallaşdırılmalıdır.

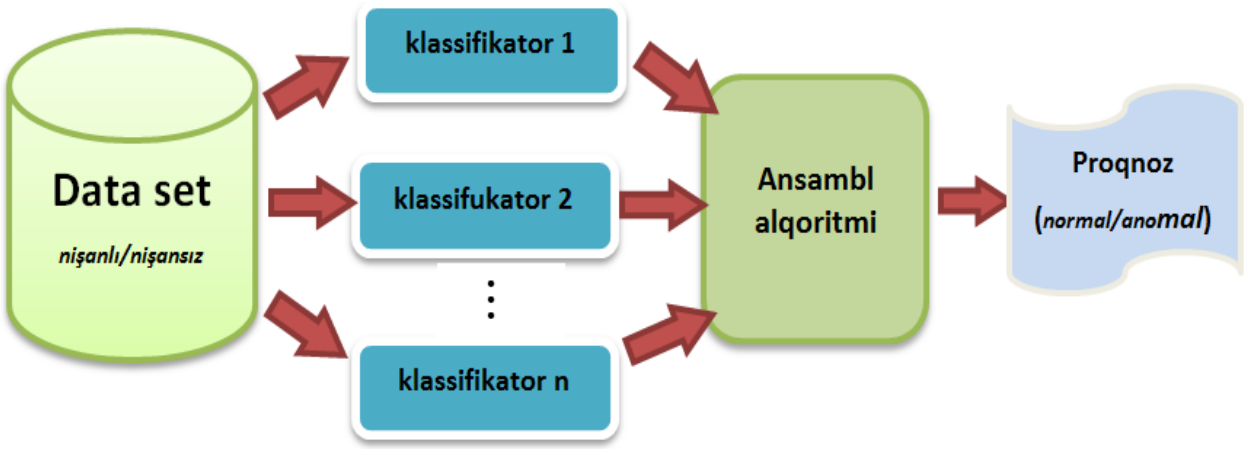
Dayaq vektorları metodu (Support Vector Machine - SVM) statistik öyrənmə nəzəriyyəsi əsasında yaradılmış ən etibarlı proqnoz metodlarındandır, V.Vapnik və A. Çervonenkis tərəfindən təklif edilmişdir. Təsnifatlandırma və reqressiya məsələlərində geniş tətbiq olunan binar xətti klassifikatordur. Fərz edilir ki, təlim toplusu (x_i, y_i) elementlərindən ibarətdir, burada x – əlamətlər vektoru, y isə ona uyğun sinif nişanıdır: $y \in \{+1,-1\}$. Elə hipermüstəvi tapmaq lazımdır ki, o, $y_i=1$ $y_i=-1$ nöqtələrini ayırsın və təlim çoxluğunun ən yaxın nöqtələrindən maksimal məsafədə keçsin. $w\cdot x-b=0$ əlamətlər fəzasını siniflərə bölən hipermüstəvini təsvir edir. Burada w, hipermüstəvinin normal vektorudur. Əgər w vektorunun x_i ilə skalyar hasili b-nin icazə verilən qiymətindən böyükdürsə, $w\cdot x_i > b \Rightarrow y_i=1$, onda nöqtə birinci kateqoriyaya, kiçikdirsə, $w\cdot x_i < b \Rightarrow y_i=-1$, ikinci kateqoriyaya aiddir (Carlos et al., 2012; Erfani et al., 2016).

Multilayer Perceptron (MLP). MLP, çoxlaylı (birdən çox) perseptron süni neyron şəbəkə sinfinə

aiddir. Məlumdur ki, süni neyron şəbəkələr bioloji neyron şəbəkələrinin simulyasiyası əsasında yaradılıb. Neyronlar, bu şəbəkələrdə əsas hesablama komponentidir. MLP ən az üç laydan ibarətdir: giriş layı, gizli lay və çıxış layı. Giriş layı istisna olmaqla, hər bir qovşaq qeyri -xətti aktivasiya funksiyasından istifadə edən bir neyrondur. Giriş layı hər bir neyronun əlamət vektorunun bir elementinə uyğun gələn giriş əlamət vektorlarından ibarətdir.

Ansambl öyrənmə alqoritmləri. Ansambl modellərinin tarixini izləmək çətin olsa da, 1990-cı illərdən bəri müxtəlif sahələrdən olan tədqiqatçılar tərəfindən araşdırılmaqdadır. Ansambl öyrənmə alqoritmləri dispersiyanı azaltmaq və ya proqnozları yaxşılaşdırmaq üçün birdən çox maşın öyrənmə alqoritmini bir proqnozlaşdırıcı modelə birləşdirən meta alqoritmlərdir (Zhou, 2015). Təlim nəzəriyyəsi kontekstində meta alqoritm ayrı-ayrı "zəif alqoritmlərin" nəticələrini birləşdirərək və hər birinə çəki verməklə onlardan yeni bir "güclü alqoritm" yaratmağı hədəfləyir. Hal-hazırda "multiplicative weights", "weighted majority", "boosting", "bagging", "ensemble averaging", "voting" və s. kimi meta alqoritm nümunələri mövcuddur. Ansambl bir qayda olaraq, ayrıca götürülmüş təlimdən daha dəqiqdir və ayrı-ayrı modellərin zəif yerlərini gizlətmək baxımından çox faydalıdır. Şəkil 1-də ansambl alqoritminin ümumi arxitekturu verilmişdir.

Sxemdən də göründüyü kimi ansambl modeli verilənlər bazası üçün daha dəqiq proqnoz vermək üçün bir neçə maşın təlimi alqoritmlərini birgə istifadə edir. Verilənlər bazası üzərində ayrı-ayrı alqoritmlər öyrədilir, hər bir alqoritm ayrılıqda proqnoz verir. Alqoritmlərin proqnozları ansambl modeldə birləşdirilir və son proqnoz səslər əsasında müəyyən edilir.



Şəkil 1. Ansambl alqoritminin ümumi arxitektur sxemi

Eksperimentlər

Bu bölmədə aparılmış eksperimentlər və nəticələr müzakirə olunur. Eksperimentlər Windows 8.1 (64bit) əməliyyat sistemi, İntel(R) Core(TM) i7-4510 prosessora, 8GB operativ yaddaşa malik kömpüterdə aparılmışdır. İşdə şəbəkə trafiki verilənlərində DoS hücumların aşkarlanması üçün təklif edilmiş klassifikasiya ansamblı alqoritminin eksperimental tədqiqi WEKA mühitində və NSL-KDD verilənlər bazasının üzərində yerinə yetirilmişdir.

NSL-KDD verilənlər bazası.

Şəbəkə trafikində anomaliyaları aşkarlamaq üçün NSL-KDD verilənlər bazasının *train.arff* və *test.arff* fayllarından istifadə olunmuşdur. NSL-KDD-nin təlim bazasında 125.973, test bazasında 22.544 yazı nümunəsi vardır. Hər bir yazı 42 atributdan ibarətdir. Sonuncu atribit hər bir yazıya "anomal", ya da "normal" vəziyyət olaraq etiketlənmişdir (NSL-KDD data set). Ümumiyyətlə, NSL-KDD atributların siyahısı və onlar haqqında geniş məlumatı (NSL-KDD data set; Dhanabal, Shantharajah, 2015; Akbar, 2010)-dən əldə etmək olar.

Maşın təlimində klassifikatorların aşkarlama performansının qiymətləndirilməsi vacib məsələdir. Aşkarlama performansının qiymətləndirilməsində dürüstlük (precision), tamlıq (recall), yanlış pozitiv hallar (false positive rate-FPR), doğru pozitiv hallar (true positive rate - TP), f-ölçü (f-measure), dəqiqlik (accuracy) metrikalarından istifadə olunmuşdur.

Xətalar matrisi (Confusion Matrix) (cədvəl 1), modelin düzgünlüyünü və dəqiqliyini qiymətləndirmək üçün istifadə edilən ən asan və ən sadə yanaşmalardan biridir (Fawcett, 2006; Holz, 2008; Gu, et al., 2006; Aliguliyev, Hajirahimova, 2019).

Cədvəl 1. Xətalar matrisi

		Actual	
		Positive	Negative
Predicted	Positive	True Positive TP	False Positive FP
	Negative	False Negative FN	True Negative TN

- doğru pozitiv hallar (TPR - true positive rate) -

$$TPR = \frac{number\ of\ TP}{number\ of\ TP + number\ of\ FN} \quad (3),$$

- doğru neqativ hallar (TNR – true negative rate) –

$$TNR = \frac{number\ of\ TN}{number\ of\ TN + number\ of\ FP} \quad (4),$$

- yanlış pozitiv hallar (FPR - false positive rate) –

$$FPR = \frac{number\ of\ FP}{(number\ of\ FP + number\ of\ TN)} \quad (5),$$

- yanlış neqativ hallar (FNR - false negative rate) –

$$FNR = \frac{number\ of\ FN}{(number\ of\ FN + number\ of\ TP)} \quad (6),$$

Xətalar matrisi və onun əsasında hesablanan klassifikasiya alqoritmlərinin aşkarlama göstəriciləri (7-10) düsturların köməyi ilə hesablanır (Huang, Charles, 2005):

- dürüstlük (precision) -

$$precision = \frac{number\ of\ TP}{(number\ of\ TP + number\ of\ FP)} \quad (7),$$

- tamlıq (recall) -

$$recall = \frac{number\ of\ TP}{(number\ of\ TP + number\ of\ FP)} \quad (8),$$

- f-ölçü (f-measure) -

$$F - measure = \frac{2*(precision*recall)}{(precision + recall)} \quad (9),$$

- dəqiqlik (accuracy) -

$$accuracy = \frac{number\ of\ TP + number\ of\ TN}{(number\ of\ TP + number\ of\ TN + number\ of\ FP + number\ of\ FN)} \quad (10).$$

Şəbəkə trafiki verilənlərində DoS hücumların aşkarlanması üçün təklif olunan maşın təlimi ansamblı modelinin yaradılmasında istifadə edilən Naive Bayes, Decision tree, Random forests, SVM, Multilayer perceptron və KNN klassifikatorları WEKA proqram mühitində test edilmişdir. Testin nəticələrinin müqayisəsi cədvəl 2-də göstərilmişdir.

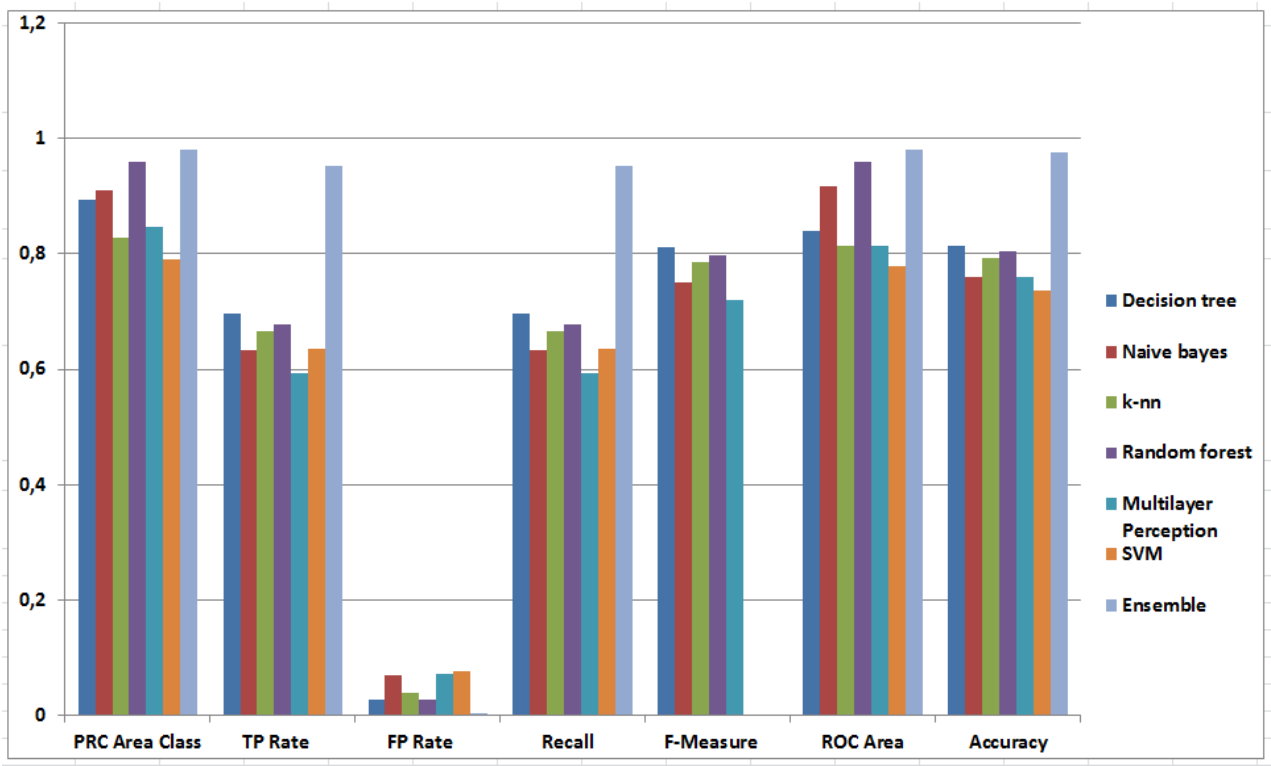
Cədvəl 2. Klassifikatorların nəticələrinin müqayisəsi

Metodlar	Accuracy	TP rate	FPRate	Prec-sion	Recall	F-meas.	ROC Area
DT	0,815	0,696	0,027	0,971	0,696	0,811	0,84
NB	0,761	0,633	0,069	0,924	0,633	0,751	0,917
KNN	0,793	0,666	0,038	0,959	0,666	0,786	0,814
RF	0,804	0,677	0,027	0,971	0,677	0,798	0,959
SVM	0,759	0,635	0,076	0,917	0,635	0,75	0,779
MLP	0,736	0,592	0,071	0,917	0,592	0,719	0,814
Ansambl(Vote)	**0,977**	**0,953**	**0,001**	**0,999**	**0,952**	**0,975**	**0,981**

Cədvəl 2-dən də göründüyü kimi DoS hücumların aşkarlanmasında bütün metrikalar üzrə ən yaxşı nəticəni təklif olunan ansambl alqoritmi nümayiş etdirmişdir. Belə ki, dəqiqlik

metrikası üzrə ən yüksək – 97.7% və yanlış təsnifatlandırma (FP) səhvi üzrə ən kiçik – 0.1% nəticə göstərmişdir. Doğruluq metrikası üzrə ən aşağı nəticəni isə MLP alqoritmi göstərmişdir.

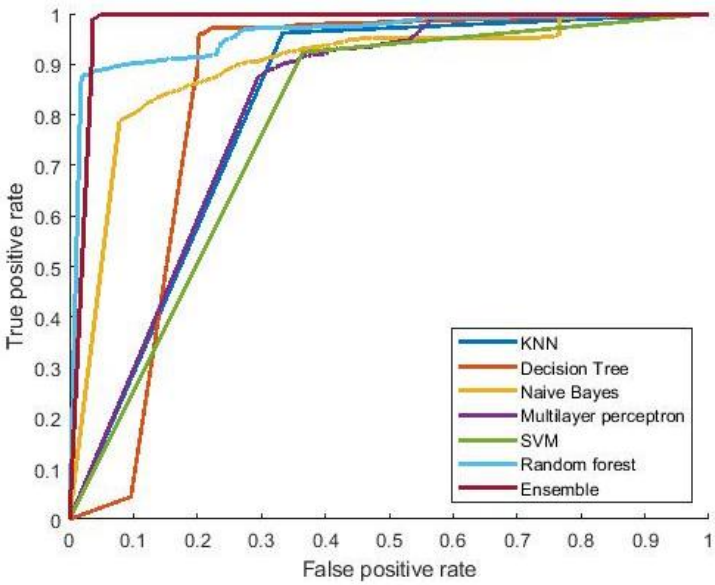
Şəkil 2-də test verilənləri üzərində tətbiq olunmuş klassifikatorların və təklif olunmuş alqoritmin müxtəlif metrikalar üzrə nəticələri vizual olaraq təqdim olunmuşdur. Qrafikdən də göründüyü kimi TP metrikası üzrə də ən yüksək nəticəni (0.953) təklif olunan ansambl alqoritmi, ən aşağı nəticələri MLP, NB və SVM alqoritmləri nümayiş etdirmişdir. FP metrikası üzrə isə təklif olunan alqoritmdən sonra ən yaxşı – 2.7% nəticəni DT və RF alqoritmləri təqdim etmişdir. Ən pis – 7.6% nəticəni isə SVM alqoritmi göstərmişdir.



Şəkil 2. Test verilənləri üzrə klassifikatorların nəticələrinin müqayisəsi

İki və daha çox sinifli təsnifatlandırma modelinin effektivliyini yoxlamaq və ya vizuallaşdırmaq üçün ən vacib qiymətləndirmə göstəricilərindən olan AUC (Area Under The Curve - əyri altındakı sahə) ROC (receiver operating characteristics- qəbuledicinin işləmə xüsusiyyətləri) əyrisindən istifadə olunur (Huang, Charles, 2005; Ling et al., 2005; Fawcett, 2006). Yəni ROC əyrisi – klassifikatorların nəticələrinin vizuallaşdırılması üsuludur. ROC qrafikləri, klassifikatorların doğru təyin etmə dərəcəsi ilə yanlış-saxta həyəcan dərəcəsi arasındakı kompromisi təsvir etmək üçün istifadə edilir (şəkil 3). Dioqramda ordinat oxu üzrə doğru müsbət hallar, absis oxu üzrə yanlış müsbət hallar göstərilir.

Şəkil 3-də ROC əyrisinin təsvirindən də aydın görmək mümkündür ki, bu əyri verilənlər bazasında tətbiq olunmuş bütün metodlar üzrə ən yüksək qiyməti (0.981) ansambl alqoritmi alaraq vahidə çox yaxınlaşmışdır. ROC Area metrikası üzrə də ən aşağı nəticəni (0.779) SVM göstərmişdir.



Şəkil 3. Test verilənləri əsasında yaradılan ROC əyrisi

Nəticə

Dövlət və özəl təşkilatlarda şəbəkə xidmətlərinin və tətbiqlərinin geniş istifadəsi şəbəkə və kompüter müdaxilələrinə qarşı adekvat təhlükəsizlik tədbirləri tələb edir Tədqiqat işində kömpüter şəbəkələrində DoS hücum olaraq bilinən şəbəkə müdaxilələrinin vaxtında aşkarlanması və qarşısının alınması üçün Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Multilayer Perceptron kimi geniş yayılmış baza təlim alqoritmlərindən ibarət ansambl modeli təşkil edilmişdir. Məqsəd anomal trafikin aşkarlanmasında ayrılıqda götürülmüş nisbətən böyük öyrənmə səhvi ehtimalı olan zəif klassifikatorlardan öyrənmə səhvi ehtimalı daha kiçik olan güclü klassifikatorun yaradılmasıdır. Ansambl modeli daha yüksək dəqiqliyi ilə cəlbedicidir. Tərəfimizdən aparılmış sınaqların nəticəsi göstərdi ki, şəbəkə trafiki anomaliyalarının - DoS hücumların aşkarlanmasında ansambl əsasında təklif olunmuş yanaşma digər baza təlim klassifikatorları ilə müqayisədə daha yüksək dəqiqlik göstərmişdir. Alınmış nəticələr 99.9% precision, 95.2% recall və 97.7% accuracy və 0.981 AUC, DoS hücumların aşkarlanmasında etibarlı bir yanaşma olduğunu deməyə əsas verir.

Təklif olunan yanaşma qənaətbəxş nəticə göstərsə də, dərin təlim və optimallaşdırma strategiyalarını və s. tətbiq etməklə daha yüksək səmərəliliyə nail olmaq olar. Eyni zamanda real-vaxt rejimində real verilənlərdən istifadə etmək də çox önəmlidir. Bu da tədqiqatımızın gələcək istiqamətini təşkil edir.

Ədəbiyyat

Abdulhammed R., Faezipour M., Abuzneid A., and AbuMallouh A. (2019). Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network trafc. IEEE Sensors Lett., Jan. 2019 3(1), pp. 1-4. https://doi.org/10.1109/LSENS.2018.2879990

Agarwal B., Mittal N. (2012). Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques. Procedia Technology, 6. pp. 996-1003. http://dx.doi.org/10.1016/j.protcy.2012.10.121

Aggarwal CC, Philip SY. (2005). An effective and efficient algorithm for high-dimensional outlier detection. VLDB J. 14(2), pp. 211–221. https://doi.org/10.1007/s00778-004-0125-5

Agrawal S, Agrawal J. (2015). Survey on anomaly detection using data mining techniques. Procedia Computer Science, 60, pp. 708–713. https://doi.org/10.1016/j.procs.2015.08.220

Akbar S., Nageswara R. K., Chandulal J. A. (2010). Intrusion detection system methodologies based on data analysis. International Journal of Computer Applications. 5(2), pp. 10–20. http://dx.doi.org/10.5120/892-1266

Akoglu L., Tong H., Koutra D. (2015). Graph based anomaly detection and description: a survey. Data Mining Knowl Discov. 29(3), pp. 626–88. https://doi.org/10.1007/s10618-014-0365-y

Alguliyev R., Aliguliyev R., İmamverdiyev Y. N., Sukhostat L. (2018). Weighted Clustering for Anomaly Detection in Big Data. Statistics, Optimization & Information Computing, 6(2), pp. 178–188. https://doi.org/10.19139/soic.v6i2.404

Alguliyev R. M., Aliguliyev R. M., Imamverdiyev Y. N. and Sukhostat L. V. (2017). An anomaly detection based on

optimization. International Journal of Intelligent Systems and Applications, 9(12), pp. 87-96. DOI: 10.5815/ijisa.2017.12.08

Aliguliyev R. M., Hajirahimova M. Sh. (2019). Classification Ensemble Based Anomaly Detection in Network Traffic. Review of Computer Engineering Research, vol. 6(1), pp. 12-23. DOI:10.18488/journal.76.2019.61.12.23

Almeida V. A., Doneda D., & de Souza Abreu J. (2017). Cyberwarfare and Digital Governance. IEEE Internet Computing. 21(2), pp. 68-71. https://doi.org/10.1109/MIC.2017.23

Antal B. and Hajdu A. (2014). An ensemble-based system for automatic screening of diabetic retinopathy. Knowl.-Based Syst., vol. 60, pp. 20–27. https://doi.org/10.1016/j.knosys.2013.12.023

Ariyaluran R. A. H., et al. (2019). Real-time big data processing for anomaly detection: A Survey. International Journal of Information Management, vol.45, pp. 289-307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

Bellman R. (2013). Dynamic programming. Chelmsford: Courier Corporation.

Breiman, L. (2001). Random forests. Machine Learning. 45(1), pp. 5–32. https://doi.org/10.1023/A:1010933404324

Buczak A. L., & Guven E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), pp. 1153-1176. https://doi.org/10.1109/comst.2015.2494502

Camacho J., Macia-Fernandez G., Diaz-Verdejo J., Garcia-Teodoro P. (2014). Tackling the big data 4 vs for anomaly detection. In: Computer communications workshops (INFOCOM WKSHPS), 2014 IEEE conference on. IEEE. pp. 500–505. https://doi.org/10.1177/1550147720921309

Carlos A. Catania, Facundo Bromberg, Carlos Garcна Garino (2012). An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection Expert Systems with Applications. 39(2), pp. 1822–1829. https://doi.org/10.1016/j.eswa.2011.08.068

Chandola V., Banerjee A., Kumar V. (2009). Anomaly detection: a survey. ACM Computing Surveys, 41(3), pp. 71-97. https://doi.org/10.1145/1541880.1541882

Chaudhary K., Yadav J., & Mallick B. (2012). A review of fraud detection techniques: Credit card. International Journal of Computers and Applications, 45(1), pp. 39–44. DOI: 10.5120/6748-8991

Dash M., & Ng W. (2010). Outlier detection in transactional data. Intelligent Data Analysis, 14(3), pp. 283–298. DOI: 10.3233/ida-2010-0422

Denning D. E. (1987). An Intrusion-Detection Model. IEEE transactions on software engineering, 13(2), pp. 222 – 232. https://doi.org/10.1109/TSE.1987.232894

Dewan Md. F., Nouria Harbi, and Mohammad Zahidur Rahman (2010). Combining naive bayes and decision tree for adaptive intrusion detection. International Journal of Network Security & Its Applications (IJNSA), 2(2), pp. 1-12. DOI: 10.5121/ijnsa.2010.2202

Dhanabal, Shantharajah S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering 9(4) (2015) pp. 446–452. http://dx.doi.org/10.4236/jcc.2016.44008

Dua S., Du X. (2011). Data mining and machine learning in cybersecurity. Boca Raton, FL, CRC Press, 256 p. https://doi.org/10.1201/b10867

Erfani S. M., Rajasegarar S., Karunasekera S., Leckie C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recogn. 58 pp. 121–34. https://doi.org/10.1016/j.patcog.2016.03.028

Farnaaz, N., & Jabbar, M. (2016). Random Forest Modeling for Network Intrusion Detection System. Procedia Computer Science, 89, pp. 213-217. https://doi.org/10.1016/j.procs.2016.06.047

Fawcett T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27 (8), pp. 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fujimaki R., Yairi T., Machida K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, pp. 401–410. https://doi.org/10.1145/1081870.1081917

Garofalo M. (2017). Big data analytics for Flow-based anomaly detection in high-speed networks. PhD Thesis. http://dx.doi.org/10.6093/UNINA/FEDOA/11617

Global - VNI Complete Forecast Highlights https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf.

Gogoi P., Bhattacharyya D. K., Borah B., and Kalita J. K. (2011). A survey of outlier detection methods in network anomaly identification. The Computer Journal, 54(4), pp. 570-588. https://doi.org/10.1093/comjnl/bxr026

Goldstein M, Uchida S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE. 11(4). https://doi.org/10.1371/journal.pone.0152173

Gu G., Fogla P., Dagon D., Lee W., Skori B. (2006). Measuring intrusion detection capability: An information-theoretic approach. Proceedings of the ACM Symposium on Information, Computer and Communications Security, pp. 90–101. https://doi.org/10.1145/1128817.1128834

Gupta M., Gao J., Aggarwal C.C., Han J. (2014). Outlier detection for temporal data: a survey. IEEE Trans Knowl Data Eng. 26(9), pp. 2250–67. https://doi.org/10.1109/TKDE.2013.184

Hacırəhimova M. Ş., (2014). Big Data texnologiyaları və informasiya təhlükəsizliyi problemləri. İnformasiya texnologiyaları problemləri, №2, pp. 49–56. http://dx.doi.org/10.25045/jpit.v07.i1.06

He H., Wang J., Graco W., and Hawkins S. (1997). Application of neural networks to detection of medical fraud. Expert Systems with Applications 13(4), pp. 329–336. https://doi.org/10.1016/S0957-4174(97)00045-6

Heydari A. et al. (2015). Detection of review spam: a survey. Expert Syst Appl, 42(7) pp. 3634–42. https://doi.org/10.1016/j.eswa.2014.12.029

Hodge V., Austin J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), pp. 85–126. https://doi.org/10.1007/s10462-004-4304-y

Holz T. (2008). Security measurements and metrics for networks. Lecture Notes in Computer Science, vol. 4909, pp. 157–165. http://dx.doi.org/10.4236/ijcns.2013.61004

Husejinović A. (2020). Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers. Periodicals of Engineering and Natural Sciences 8(1), pp. 1-5. http://pen.ius.edu.ba

Xuan S., Liu G., Li Z., Zheng L., Wang S., & Jiang C. (2018). Random forest for credit card fraud detection. IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). https://doi.org/10.1109/ICNSC.2018.8361343

Imamverdiyev Y., Abdullayeva F. (2018). Deep Learning Method for Denial of Service Attack Detection Based on Restricted Boltzmann Machine Big Data, 6(2), pp. 159-169. https://doi.org/10.1089/big.2018.0023

Jin Huang and Charles, Ling X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms IEEE transactions on knowledge and data engineering, 17(3), pp. 299-310. https://doi.org/10.1109/TKDE.2005.50

KDD data set, 1999. http://kdd.ics.uci.edu/databases/kddcup99

Kim G., Lee S., and Kim S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst. Appl., 41(4), pp. 1690-1700. https://doi.org/10.1016/j.eswa.2013.08.066

Kumari R., Sheetanshu, Singh M. K., Jha R. & Singh N. K. (2016). Anomaly detection in network traffic using K-mean clustering. 3rd International Conference on Recent Advances in Information Technology (RAIT). https://doi.org/10.1109/RAIT.2016.7507933

Lee W., Stolfo S. J., Mok K. W. (2000). Adaptive intrusion detection: A data mining approach. Artificial Intelligence Review, 14(6), pp. 533-567. https://doi.org/10.1023/A:1006624031083

Marnerides A. K., Spachos P., Chatzimisios P., and Mauthe A. U. (2015). Malware detection in the cloud under Ensemble Empirical Mode Decomposition. In 2015 Int. Conf. Comput. Netw. Commun. IEEE., pp. 82–88. https://doi.org/10.4018/IJESMA.2018070104

McHugh J. (2000). Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Transactions on Information and System Security, 3(4), pp. 262–294. https://doi.org/10.1145/382912.382923

Mukherjee B., Heberline L. T., & Levitt K. (1994). Network instruction detection. IEEE Network, 8, pp. 26–41. https://doi.org/10.1007/978-0-387-33112-6_8

Münz G., Li S., Carle G. (2007). Traffic anomaly detection using k-means clustering. In: GI/ITG Workshop MMBnet. pp. 13-14. DOI:10.1.1.323.6870

Nassif A. B. et al. (2021). Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access, vol.9, pp. 78658- 78700. https://doi.org/10.1109/access.2021.3083060

NSL-KDD data set for network-based intrusion detection systems [Electronic resource]. 2017. Access mode: http://nsl.cs.unb.ca/NSL-KDD/

Patcha A., Park J.M., (2007). An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw. 51(12), pp.3448–3470. http://dx.doi.org/10.1016%2Fj.comnet.2007.02.001

Phua C., Lee V., Smith-Miles K., & Gayler R. (2010). A comprehensive survey of data miningbased fraud detection, Research Computing Research Repository https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.1016%2Fj.chb.2012.01.002&v=67e7929e

Raguseo E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. International Journal of Information Management, 38(1), pp. 187-195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008

Rehman M. H., Liew C. S., Abbas A., Jayaraman P. P., Wah T. Y., & Khan S. U. (2016). Big data reduction methods: a survey. Data Science and Engineering, 1(4), pp. 265-284. https://doi.org/10.1007/s41019-016-0022-0

Samuel A. L. (1959). Some studies in Machine Learning using the game of checkers. IBM Journal of research and development, 3(3), pp. 210–229. https://doi.org/10.1147/rd.33.0210

Saneja B., Rani R. (2017). An efficient approach for outlier detection in big sensor data of health care. International Journal of Communication Systems, 30(17), pp. 1-10. https://doi.org/10.1002/dac.3352

Schlegl T., Seeböck P., Waldstein S. M., Schmidt-Erfurth U., and Langs G. (2017). Unsupervised Anomaly Detection With Generative Adversarial Networks to Guide Marker Discovery. International Conference on Information Processing in Medical Imaging, pp. 146-157. https://doi.org/10.1007/978-3-319-59050-9_12

Shon T., Moon J. (2007). A hybrid machine learning approach to network anomaly detection. Information Sciences, 177(18), pp. 3799-3821. https://doi.org/10.1016/j.ins.2007.03.025

Srikanth T., Philip B., Jiong J. et al. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 7(42), pp. 1-30. https://doi.org/10.1186/s40537-020-00320-x

Tsai C. F., Hsu Y. F., Lin C. Y., Lin W. Y. (2009). Intrusion detection by machine learning: A review. Expert Syst. Appl., 36(10), pp. 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029

Varian I. (2020). IMRT (Intensity Modulated Radiation Therapy). 26 June 2020. https://patient.varian.com/en/treatments/radiation-therapy/treatment-techniques

Wang C., Zhao Z., Gong L., Zhu L., Liu Z., & Cheng X. (2018). A Distributed Anomaly Detection System for In-Vehicle Network Using HTM. IEEE ACCESS, 6, pp. 9091-9098. https://doi.org/10.3390/s20143934

Wang L. & Jones R. (2017). Big data analytics for network intrusion detection: A survey. International Journal of Networks and Communications, 7(1), pp. 24-31 doi: 10.5923/j.ijnc.20170701.03

Wauters, M., & Vanhoucke, M. (2017). A Nearest Neighbour extension to project duration forecasting with Artificial Intelligence. European Journal of Operational Research, 259(3), pp. 1097-1111. https://doi.org/10.1016/j.ejor.2016.11.018

Wei Y. et al. (2019). MSD-Kmeans: A Novel Algorithm for Efficient Detection of Global and Local Outliers, pp. 1-12. https://doi.org/10.1145/3459930.3469523

Yang T. et al. (2016). Improve the Prediction Accuracy of Naive Bayes Classifier with Association Rule Mining.

IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance, and Smart Computing, IEEE International Conference on Intelligent Data and Security, pp. 129-133. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.38

Zhang J., Zulkernine M., and Haque A. (2008). Random-Forests-Based Network Intrusion Detection Systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(5), pp. 649–659. https://doi.org/10.1109/TSMCC.2008.923876

Zhang M.L., Zhou Z.H., (2005). A k-nearest neighbor based algorithm for multi-label classification / Proc. of the International Conference on Granular Computing, pp. 718–721. https://doi.org/10.1109/GRC.2005.1547385

Zhou Z., (2012). Hua Ensemble Methods. Foundations and Algorithms. CRC Press, p.234. https://doi.org/10.1201/b12207

Məkrufə Ş.Hacırəhimova [a], Leyla R. Yusifova [b]

[a,b] AMEA İnformasiya Texnologiyaları İnstitutu.
Azərbaycan, Bakı ş., AZ1141, B.Vahabzadə küç., 9A.

Макруфа Ш. Гаджирагимова [a], Лейла Р. Юсифова [b]

[a,b] Институт Информационных Технологий НАН Азербайджана.
Азербайджан, г. Баку, AZ1141, ул. Б.Вахабзаде, 9A.

[a] 0000-0003-0786-5974; [b] 0000-0002-9720-8638