განაწილებულ დიდ მონაცემთა ნაკადებზე თავდასხმის აღმოჩენა მანქანური სწავლებით

გულნარა ჯანელიძე
საინჟინრო მეცნიერებათა დოქტორი, პროფესორი, საქართველოს ტექნიკური უნივერსიტეტი
janelidzegulnara08@gtu.ge

დათა დათაშვილი
დოქტორანტი, სამცხე-ჯავახეთის სახელმწიფო უნივერსიტეტი
datadatashvili99@gmail.com

აბსტრაქტი
თანამედროვე ციფრულ სამყაროში მონაცემების რაოდენობა დღითიდღე ექსპონენციურად იზრდება, რამაც უსაფრთხოების ახლებური მიდგომების შემუშავების აუცილებლობა წარმოშვა. დღესდღეობით განსაკუთრებით აქტუალური გახდა განაწილებული საინფორმაციო სისტემების გამოყენება, რომელთა მეშვეობით მონაცემთა ნაკადები მოძრაობს სხვადასხვა პლატფორმებზე. თუმცა, სწორედ ეს მაღალი განაწილებულობა და ღიაობა ქმნის ხელსაყრელ პირობებს კიბერ თავდასხმების განსახორციელებლად. აღსანიშნავია, რომ შემოჭრის აღმოჩენის ტრადიციული საშუალებები საუკეთესოდ მუშაობს შედარებით მცირე სიჩქარის მონაცემებზე. ისინი არაეფექტურია დიდ მონაცემებზე და არ შეუძლიათ მაღალსიჩქარიანი მონაცემების დამუშავება, ამიტომ ახალი მეთოდების ადაპტირებაა საჭირო დიდ მონაცემებზე სამუშაოდ, რათა აღმოაჩინონ შემოჭრისთვის დამახასიათებელი ნებისმიერი ნიშანი. ამ თვალსაზრისით განსაკუთრებულ მნიშვნელობას იძენს მანქანური სწავლების (Machine Learning) მეთოდების გამოყენება, რომლებიც შეიძლება გამოყენებულ იქნას როგორც ანომალიის დეტექტირებისთვის, ასევე ცნობილ თავდასხმებთან დაკავშირებული ნიშნების იდენტი-ფიცირებისთვის.
ნაშრომში გაანალიზებულია განაწილებულ დიდ მონაცემებზე DdoS (Distributed Denial of Service) შეტევების რეალურ დროში აღმოჩენის მნიშვნელობა და მასთან დაკავშირებული სირთულეები. აღწერილია DoS შეტევები, რომლებიც უშუალოდ დაკავშირებულია დიდ მონაცემთა სისტემებთან. განსაკუთრებული ყურადღება ეთმობა განაწილებულ საინფორმაციო ნაკადებზე სხვადასხვა სახის შეტევების დროულ აღმოჩენას მანქანური სწავლების გამოყენებით. აღნიშნული პრობლემის გადასაწყვეტად შემოთავაზებულია შემთხვევითი ტყის (Random Forest) მოდელი. დამუშავებულია სისტემაში არასანქცირებული შეღწევის აღმოჩენის ალგორითმი შემთხვევითი ტყის კლასიფიკატორის გამოყენებით. წარმოდგენილია ღრმა შემთხვევით ტყეზე დაფუძნებულ ქსელში შეღწევის გამოვლენის მოდელი.
ნაშრომში გაანალიზებულია შემთხვევითი ტყის გამოყენებასთან დაკავშირებული სირთულეები, რაც ძირითადად ეხება მონაცემთა ბალანსს და გამოთვლით სირთულეებს, მაგრამ შემთხვევითი ტყის მეთოდი ინარჩუნებს თავის უპირატესობას ქსელური ანომალიების აღმოჩენის ამოცანების გადასაწყვეტად, განსაკუთრებით რეალურ დროში განაწილებული მონაცემებისთვის.

საკვანძო სიტყვები: DDoS შეტევები, Random Forest მეთოდი, ღრმა პარალელური შემთხვევითი ტყე.

JEL: C55; C45; D83

DOI: 10.52244/c2025.27

შესავალი

თანამედროვე ქსელური გარემო წარმოადგენს კომპლექსურ და დინამიკურ სტრუქტურას, რაც ართულებს ტრადიციული თავდაცვითი სისტემების — მაგალითად, firewall-ებისა და სიგნატურაზე დაფუძნებული შეღწევადობის დეტექციის სისტემების (IDS) ეფექტურ გამოყენებას. **ახალი ტიპის, ინტელექტუალური თავდასხმები** ხშირად იმალება კანონზომიერ ქსელურ ტრაფიკში და არ ტოვებს აშკარა ნიშნებს, რის გამოც მათი დროული აღმოჩენა საკმაოდ პრობლემურია. რეალურ დროში მაღალი სიზუსტით შეტევების ამოცნობა დინამიკურ და მოცულობით მონაცემთა ნაკადებში კვლავ **აუხსნელი გამოწვევაა.**

მონაცემთა უსაფრთხოება და კონფიდენციალურობა ასევე დიდი გამოწვევაა დიდი მონაცემების არსებობისას, განსაკუთრებით კი ქსელური შეტევების დროს. ერთ-ერთი მთავარი შეტევაა DDoS (Distributed Denial of Service) შეტევა. DDoS შეტევები არის კიბერშეტევები კონკრეტულ სერვერებზე ან ქსელზე, რომელთა მიზანია ამ ქსელის ან სერვერის ნორმალური ფუნქციონირების დარღვევა. DDoS შეტევების რეალურ დროში აღმოჩენა და შემცირება მარტივი არ არის, მაგრამ ამის გადაწყვეტას დიდი ღირებულება აქვს, რადგან შეტევებმა შეიძლება დიდი პრობლემები გამოიწვიოს. დიდი მონაცემები (Big Data), უსაფრთხოების კუთხით რამდენიმე მნიშვნელოვან პრობლემას ქმნის. ეს პრობლემები გამომდინარეობს მონაცემების მოცულობის, მრავალფეროვნებისა და დამუშავების სირთულეებიდან [1]. მათ შორის:

– მონაცემთა დაცვა და კონფიდენციალურობა: როდესაც დიდი რაოდენობით ინფორმაცია გროვდება, იზრდება კონფიდენციალური მონაცემების გაჟონვის რისკი. მაგალითად, პერსონალური, ფინანსური ან სამედიცინო მონაცემების უსაფრთხოების დარღვევამ შესაძლოა სერიოზული ზიანი მიაყენოს როგორც ორგანიზაციას, ისე ცალკეულ პირებს;

– მონაცემების მართვისა და მონიტორინგის სირთულე: დიდი მოცულობის მონაცემების მუდმივი მონიტორინგი და მართვა რთულია. ამასთან, მონაცემები ხშირად ინახება სხვადასხვა ადგილას და ფორმატში, რაც კიდევ უფრო ართულებს მათ დაცვას;

– უსაფრთხოების სისტემების მოწყვლადობა: ტრადიციული უსაფრთხოების ინსტრუმენტები და მეთოდები ხშირად ვერ უმკლავდება დიდი მონაცემების ნაკადს. აღნიშნულის გამო საჭიროა ახალი, უფრო მოწინავე ტექნოლოგიების გამოყენება, როგორიცაა ხელოვნური ინტელექტი და მანქანური სწავლება;

– კანონმდებლობასთან შესაბამისობა: ბევრ ქვეყანაში არსებობს მკაცრი რეგულაციები პერსონალური მონაცემების დაცვის შესახებ (მაგალითად, GDPR ევროკავშირში). Big Data-ს შემთხვევაში ამ წესებთან შესაბამისობის უზრუნველყოფა ხშირად რთულდება.

ძირითადი ნაწილი

დიდი მონაცემების (Big Data) სისტემებთან დაკავშირებული DoS (Denial-of-Service) შეტევები განსხვავდება ტრადიციული შეტევებისგან, რადგან ისინი არა მხოლოდ სერვისის შეფერხებას, არამედ მონაცემების მთლიანობის, ხელმისაწვდომობისა და დამუშავების პროცესების დარღვევას ისახავენ მიზნად. რამდენიმე ტიპის DoS შეტევა, რომელიც სპეციფიკურია დიდი მონაცემების სისტემებისთვის შესაძლოა შემდეგნაირად გამოვყოთ [2]:

– მონაცემთა წყაროების გადატვირთვა: Big Data სისტემები ხშირად იღებენ მონაცემებს მრავალი სხვადასხვა წყაროდან. შესაბამისად, თავდამსხმელებს შეუძლიათ შექმნან მონაცემთა ყალბი წყაროები, რომლებიც სისტემას უზარმაზარი, არასაჭირო ინფორმაციით დატვირთავენ, რაც გამოიწვევს სისტემის რესურსების (CPU, მეხსიერება) სრულ ამოწურვას;

– რესურსების ამოწურვა: Big Data დამუშავების სისტემები (როგორიცაა Hadoop ან Spark) იყენებენ დიდი რაოდენობით გამოთვლით რესურსებს. DoS შეტევის დროს შესაძლოა გამოიკვეთოს მონაცემთა დამუშავების რთული და უსარგებლო ამოცანები, რომლებიც რესურსებს სრულად დაიკავებს;

– მონაცემების „წებოვნება“: თავდამსხმელებს შეუძლიათ სისტემაში შეაპარონ მცირე, მაგრამ მავნე მონაცემები, რომლებიც გამოიწვევს შეცდომებს მონაცემთა დამუშავების პროცესში. მაგალითად, მათ შეიძლება შეცვალონ მონაცემთა სტრუქტურა ან ჩაწერონ არასწორი მონაცემები, რის შედეგადაც მონაცემთა ნაკადი გახდება უვარგისი;

– მონაცემთა ბაზის ბოროტად გამოყენება: მონაცემთა ბაზის გადატვირთვა, რაც გამოიწვევს მონაცემებზე წვდომის შეფერხებას.

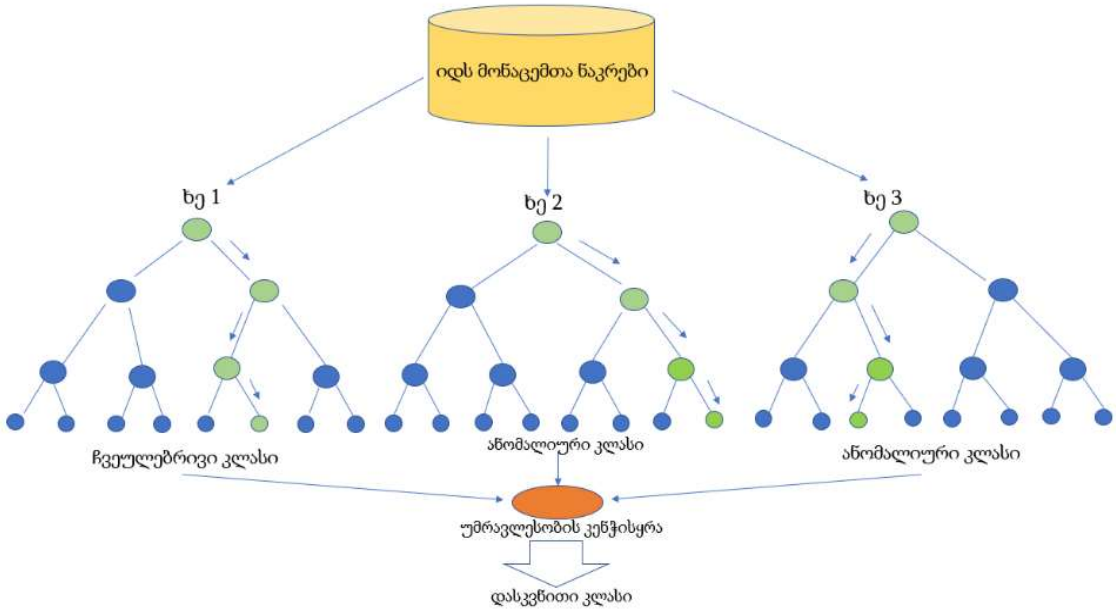
ყოველივე განხილულის საფუძველზე იკვეთება ნაშრომში დამუშავებული ძირითადი პრობლება, თუ როგორ შეიძლება განაწილებულ საინფორმაციო ნაკადებზე განხორციელებული სხვადასხვა ტიპის კიბერშეტევების ეფექტური და დროული დეტექტირება მანქანური სწავლების მეთოდების გამოყენებით.

ანომალიაზე ორიენტირებული ინფორმაციის დაცვის სისტემა ეფუძნება მოვლენათა კორელაციის მოდელს, სადაც გამოიყენება ხისებრი სტრუქტურის წესების კრებული. ნაშრომში წარმოდგენილია შემთხვევითი ტყის (Random Forest) მოდელის გამოყენების მიდგომა, რაც განპირობებულია იმით, რომ ფაქტობრივად ეს არის ანსამბლის მეთოდი, იგი აერთიანებს მრავალი ინდივიდუალური მოდელის პროგნოზებს, რათა მიიღოს უკეთესი და სტაბილური შედეგი. ამდენად, ეს ადასტურებს ინფორმაციის დაცვის სისტემისთვის (IDS) შემთხვევითი ტყის მოდელის გარკვეულ უპირატესობას სხვა ცალკეულ კლასიფიკატორებთან შედარებით [3].

ზოგადად, შემთხვევითი ტყის კლასიფიკატორის ეფექტურობა შეიძლება დამტკიცდეს ტესტირებით: სხვადასხვა რაოდენობის ხეების ვარიანტებითა და შემდეგ კლასიფიკატორებს შორის მუშაობის განსხვავებების შეფასებით სტატისტიკური ანალიზის გამოყენებით.

შემთხვევითი ტყე წარმოადგენს კლასიფიკატორის ანსამბლს, რომელიც გამოიყენება კლასიფიკაციის ან რეგრესიის ამოცანებისთვის. ორიგინალური ვერსიის მიხედვით შემთხვევითი ტყის მიდგომა შეიძლება განვმარტოთ როგორც კრებულის, ე.წ. ჩანთების (bagging) ვერსია, სადაც საბაზისო კლასიფიკატორს წარმოადგენს შემთხვევითი ხე. აღნიშნული მიდგომა განიხილება, როგორც ანსამბლის სწავლების პროცესი, რომელშიც გადაწყვეტილების ხე (decision tree) მიიღება, როგორც საბაზისო კლასიფიკატორი.

გარდა ამისა, შემთხვევითი ტყე არის კლასიფიკატორი, რომელიც შედგება ხეებისგან შემდგარ კლასიფიკატორთა ჯგუფისგან. თითოეული ხე იზრდება შემთხვევითი ვექტორის შესაბამისად, რომლებიც დამოუკიდებელი და იდენტურად განაწილებულია. ამასთან, „ხმის მიცემა“ ხდება თითოეული ხიდან ანსამბლში შემავალი ვექტორის ყველაზე პოპულარული კლასისთვის.



ნახ. 1. შემთხვევითი ტყიდან კლასების ფორმირება პროგნოზირებისთვის

შემთხვევითი ტყის მრავალფეროვნება მიიღება ატრიბუტთა სიმრავლიდან ნიმუშის აღებით, ან მონაცემთა კრებულიდან, ან უბრალოდ გადაწყვეტილების ხის ზოგიერთი პარამეტრის შემთხვევითი შეცვლით . განირჩევა შემთხვევითი ტყის ორი პარამეტრი, რომელიც შეიძლება დაზუსტდეს: თითოეულ კვანძში შერჩეული ცვლადების რაოდენობა, რომელიც ჩვეულებრივ ფიქსირდება ყველა კვანძში და ხეების რაოდენობა, რომლებიც აშენებენ ტყეს.

ნახ.1.-ზე ნაჩვენებია შემთხვევითი ტყის მაგალითი კლასიფიკაციისთვის, სადაც შემთხვევითი ტყე აყალიბებს ცალკეულ ხეებს (ტყის შესაქმნელად), რომელიც შემდგომ გამოიყენება ნორმალური ან ანომალიის კლასის პროგნოზირებისთვის. კლასის საბოლოო პროგნოზი იწარმოება თითოეული ცალკეული ხის კლასის პროგნოზისთვის უმრავლესობით ხმის მიცემით. შემთხვევითი ტყის ეფექტური განხორციელების ექსპერიმენტული შემოწმებისთვის უნდა შეირჩეს ხეების სხვადასხვა რაოდენობა, აგრეთვე სხვა პარამეტრები, როგორიცაა მინიმალური სიღრმე, მაქსიმალური სიღრმე, დანიმუშების სიხშირე (Sampling Rate), ჰისტოგრამის ტიპი და ა.შ. როგორიცაა კლასტერიზაცია, ანომალიებისა და არასწორი გამოყენების ძებნა. Random Forest თეორიული ნაწილი საკმაოდ მარტივია და იგი წარმატებით გამოიყენება ანომალიების ძებნისას. მხოლოდ a(x) საბოლოო კლასიფიკატორის ფორმულა არის საჭირო:

$$a(x) = \frac{1}{N} \sum_{i=1}^{N} b_i(x)$$

სადაც: N – ხეების რაოდენობა; i – მთვლელი ხეებისათვის; b – გადაწყვეტილების ხე; x – მონაცემთა საფუძველზე ჩვენს მიერ გენერირებული ამორჩევა.

ალგორითმის მნიშვნელოვანი პარამეტრია თვისებათა მაქსიმალური რაოდენობა დაშლის არჩევისთვის – max_features, რომლის ზრდის შემთხვევაში ტყის აგების დრო იზრდება, ხოლო ხეები ერთმანეთის მსგავსნი ხდებიან. კიდევ სხვა მნიშვნელოვანი პარამეტრებია თვისებათა მინიმალური რაოდენობა დაშლის არჩევისთვის – min_samples_split და ობიექტების შეზღუდული რაოდენობა min_samples_leaf, ასევე ხეების მაქსიმალური სიღრმე – max_depth. რაც უფრო

ნაკლებია მაქსიმალური სიღრმე, მით უფრო სწრაფად შენდება და მუშაობს შემთხვევითი ხის ალგორითმი. სიღრმის გაზრდის დროს მკვეთრად იზრდება მოდელის სწავლებისა და ტესტირების ხარისხი.

შემთხვევითი ტყის ალგორითმი. დავიწყოთ გადაწყვეტილების ხიდან, რომელიც წარმოადგენს შემთხვევითი ტყის ალგორითმის ძირითად სტრუქტურულ ელემენტს. სწორედ იმაზე, თუ როგორ არის აგებული თითოეული ხე, იქნება დამოკიდებული ალგორითმის მუშაობა და მდგრადობა.

განვიხილოთ ობიექტების შეფასებული ამორჩევა:

$$\{(x_i, y_i)\}_{i=1}^{N}$$

სადაც $x_i \in \mathrm{R}^2$ — ობიექტის ნიშან-თვისებრივი აღწერაა ორგანზომილებიან სივრცეში, ხოლო $y_i \in \{0, 1\}$ - კლასის ჭდე.

ალგორითმის ყოველ ბიჯზე აუცილებელია ნიშან-თვისებისა და ზღურბლის მნიშვნელობის არჩევა რომლის მიხედვითაც მოხდება ოპტიმალური დანაწევრება.

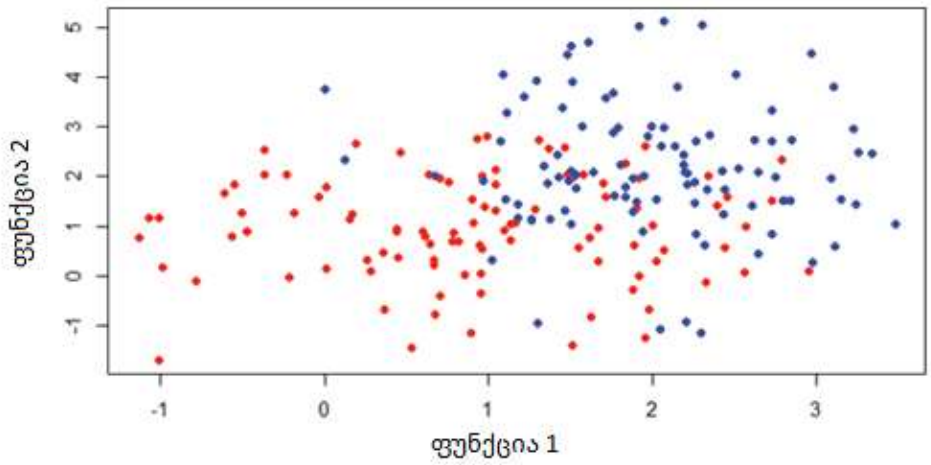
ა) კლასიფიკაციის ამოცანის შემთხვევაში გამოიყენება კრიტერიუმი iGain:

$$\mathrm{iGain}(S) = H(S) - \sum_{v \in \{L,R\}} \frac{|S_v|}{|S|} H(S_v),$$

$$H(S) = -\sum_{c \in C} p_c \log_2(p_c),$$

სადაც C - არის კლასების სიმრავლე,

p_c - არის c კლასის ალბათობა ობიექტთა S სიმრავლისთვის;



ნახ.2. კლასიფიკაციის ამოცანის საწყისი მონაცემები

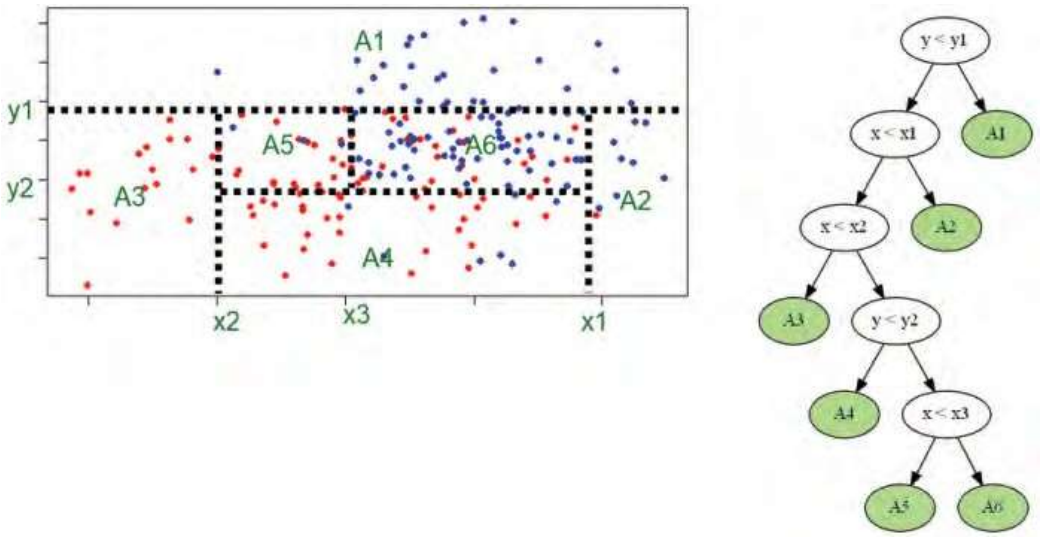
ბ) რეგრესიის ამოცანისთვის iGain კრიტერიუმი დისპერსიის გამოყენებით:

$$\mathrm{iGain}(S) = |S|\mathrm{Var}(S) - \sum_{v \in \{L,R\}} |S_v|\mathrm{Var}(S_v),$$

სადაც: Var(S) - არის S სიმრავლიდან ობიექტების გამოძახილის დისპერსია.

თითოეული დანაწევრების დროს ყველა ობიექტი იყოფა ორ უფრო მცირე ჯგუფად ანუ თითოეულ კვანძში ამოცანა იყოფა ორ მცირე ქვეამოცანად. ხის მწვერვალ-კენწეროში ობიექტების მაქსიმალური რაოდენობის მიცემით ხდება ალგორითმის გაჩერების კრიტერიუმის დადგენა. ალგორითმის მუშაობის შედეგად ხე შეიძლება

აღიწეროს სხვადასხვა ხერხით (ნახ.3.)



ნახ.3. კლასიფიკაცია ერთი გადაწყვეტილების ხის სახით

ამგვარად, შეიძლება ობიექტთა განხილული ამორჩევა ხარისხიანად კლასიფიცირდეს თუნდაც მხოლოდ ერთი გადაწყვეტილების ხის მეშვეობით, თუ ტესტური ობიექტისთვის A_i უჯრედში მოხვედრილი პასუხის ნომერი ამ კლასის უჯრედში ყველაზე ხშირად შეგვხვდება. თუმცა რეალურ ამოცანებში ხშირად გვხვდება გაზომვათა ცდომილებები და ამოვარდნილი ობიექტები, რომლებიც სერიოზულად აზიანებენ ერთი გადაწყვეტილების ხისგან მიღებული კლასიფიკაციის ხარისხს [4].

ამიტომ, თითოეული ახალი ხის აგების წინ, ხდება ახალი ნიმუშების შერჩევა ამორჩევათა გამეორებებით $\{(x^k, y^k)\}^N$, რომელზეც ხორციელდება k ნომრის ხის სწავლება. ყველა ხის აგების შემდეგ ყოველი z_i ტესტური ობიექტი, შუალედური პასუხის სახით, მიიღებს თითოეული ხისგან მინიჭებულ ჭდეთა ვექტორს, რომელიც უბრალო კენჭისყრის მეთოდით გარდაიქმნება საფინალო ჭდედ.

შემთხვევითი ტყის მოდელირება ქსელის IDS-სთვის. აღვწეროთ სისტემაში შეღწევის აღმოჩენის მოდელი შემთხვევითი ტყის კლასიფიკატორის გამოყენებით. აღნიშნული მოდელი შეიძლება განვიხილოთ ალგორითმის სახით შემდეგი ბიჯების თანმიმდერობით (ნახ.4.)

ბიჯი 1: მონაცემთა ნაკრების მიწოდება შესასვლელზე.

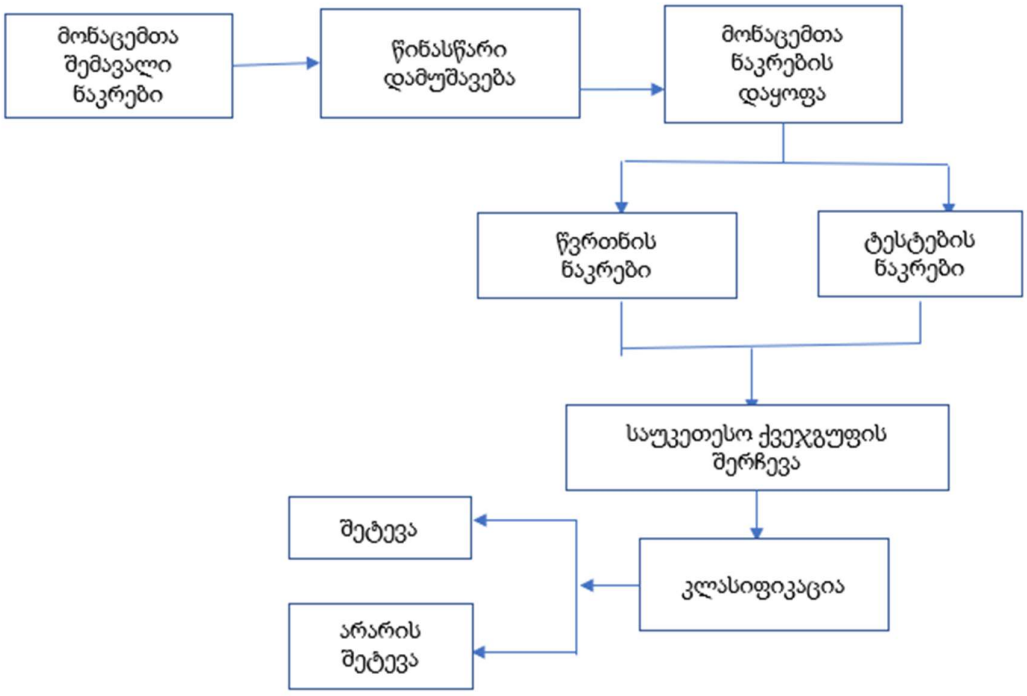
ბიჯი 2: წინასწარი დამუშავება ნიშან-თვისებების შესაბამისად.

ბიჯი 3: მონაცემთა ნაკრების დაყოფა წვრთნისა და ტესტირების ნაკრებებად.

ბიჯი 4: ხდება ნიშან-თვისებების საუკეთესო ქვეჯგუფის შერჩევა. შემთხვევითი ტყის წვრთნა. შემდეგ ტესტის მონაცემთა მასივი მიეწოდება რანდომიზებულ ტყეს კლასიფიკაციისთვის.

ბიჯი 5: განისაზღვრება არის თუ არა შეღწევის მცდელობა.

კლასიფიკაციის მოდელის ამონახსნის სისწორის შეფასების თვალსაზრისით, მიზანშეწონილია ორობითი კლასიფიკაციის მოცემულობის დაშვება. კლასიფიკაციისა და აღმოჩენისთვის შემთხვევითი ტყის გამოყენებასთან ერთად ორობითი კლასიფიკაციის ასპექტში ეფექტურად არის მიჩნეული აგრეთვე ბინარული ნაწილაკთა გროვის ოპტიმიზაციის (BPSO) ალგორითმი.



ნახ.4. შემთხვევითი ტყის ალგორითმი

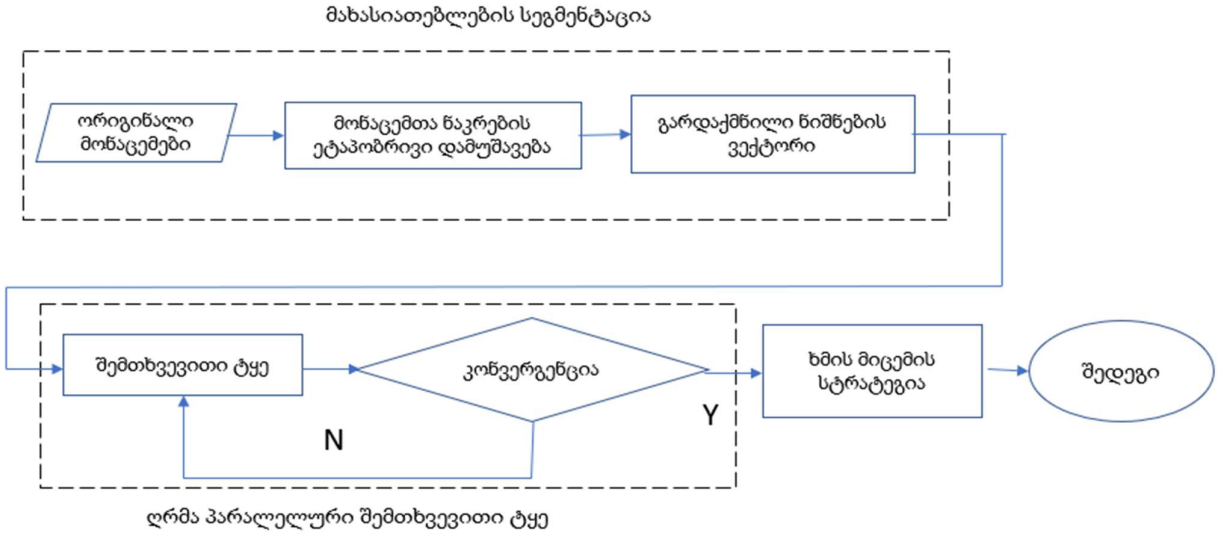
ღრმა შემთხვევით ტყეზე დაფუძნებული IDS მოდელი. დიდ მონაცემთა ნაკრების შემთხვევითი ტყის მოდელირების თვალსაზრისით, უნდა აღინიშნოს, რომ შეღწევის აღმოჩენის პრობლემების გადასაჭრელად მიზანშეწონილად მიგვაჩნია დიდ მონაცემთა ნაკრებების პარალელური დამუშავების პარადიგმაზე დაფუძნებული ახალი მიდგომის შემუშავება. კერძოდ, ღრმა შემთხვევით ტყეზე დაფუძნებულ ქსელში შეღწევის გამოვლენის მოდელი ანუ IoT ქსელში შეღწევის დეტექტირება შემთხვევითი ტყის მოდელის გამოყენებით პარალელური დამუშავების მიდგომით.

განვიხილოთ ღრმა შემთხვევით ტყეზე დაფუძნებულ ქსელში შეღწევის გამოვლენის მოდელი. პირველ ეტაპზე ხდება ორიგინალურ ნიშან-თვისებათა ფრაგმენტებად დაყოფა და შემთხვევითი ტყის წვრთნა. შედეგად მიიღება კონკატენირებული კლასის ვექტორი, რომელიც გამოიყენება მეორე ეტაპზე მრავალდონიანი კასკადური პარალელური რანდომიზებული ტყის მოსამზადებლად და ბოლოს, ორიგინალი მონაცემების კლასიფიკაცია ხორციელდება ხმის მიცემის სტრატეგიით კონვოლუციური კასკადის ბოლო შრის შემდეგ (ნახ. 5).

ანომალიების კლასიფიკაციის და ამოცნობის ამოცანა მნიშვნელოვნად მარტივდება ხმის მიცემის სტრატეგიის გამოყენებით. პარალელური შემთხვევითი ტყის (Parallelized Random Forest - PRF) კასკადში ბოლო ფენაში მიიღება საბოლოო შედეგი. ყველა გამომავალი კლასი ანუ გადაწყვეტილების ხეები ბოლო ფენაში ითვლება და შემდეგ, გადაწყვეტილება მიიღება ალბათობის განაწილებით ხმის მიცემის სტრატეგიის გამოყენებით [5].

უმრავლესობის ხმის მიცემა გამოიყენება ანომალიის გამოვლენის ამოცანებში მაღალი საიმედობის დაცვით. თუ ნიმუში იღებს ხმების ნახევარზე მეტს, პროგნოზირებულია როგორც გამონაკლისი, წინააღმდეგ შემთხვევაში იგი უარყოფილია. ამ დროს, უმრავლესობის ხმის მიცემის მეთოდი გარდაიქმნება მრავლობითი ხმის მიცემის მეთოდად. ამ მდგომარეობაში, თუ პროგნოზის შედეგად მრავალი მიიღებს ერთნაირი რაოდენობის ხმას, მაშინ ერთ-ერთი იქნება

შერჩეული.



ნახ. 5. ღრმა პარალელური შემთხვევითი ტყე

დასკვნა

როგორც ნაშრომიდან ჩანს შემთხვევითი ტყის მოდელი ძალიან ეფექტურია ქსელური ანომალიების აღმოსაჩენად. მისი ეფექტურობა რამდენიმე ძირითადი უპირატესობით არის განპირობებული: მაღალი სიზუსტე: შემთხვევითი ტყე კლასიფიკატორების ანსამბლია და, როგორც წესი, იშვიათად უშვებს შეცდომებს. ის აერთიანებს მრავალი გადაწყვეტილების ხის შედეგებს, რაც მას უფრო საიმედოს ხდის, ვიდრე ერთი ხე. გადაჭარბებული მორგების (Overfitting) მიმართ მდგრადობა: ეს მეთოდი შედარებით მდგრადია მონაცემების გადაჭარბებული მორგების მიმართ, რაც ხშირად გვხვდება სხვა მოდელებში. მნიშვნელოვანი მახასიათებლების შერჩევა: შემთხვევით ტყეს შეუძლია დაადგინოს, რომელი მახასიათებლები (მაგალითად, პორტის ნომერი, ტრაფიკის ზომა) არის ყველაზე მნიშვნელოვანი ანომალიების აღმოსაჩენად. პარალელური დამუშავების შესაძლებლობა: ამ მოდელის შექმნა შესაძლებელია პარალელურ რეჟიმში, რადგან თითოეული ხე დამოუკიდებლად იქმნება. ეს კი ამცირებს წვრთნის დროს დიდ მონაცემებთან მუშაობისას.

მიუხედავად უპირატესობებისა, შემთხვევით ტყეს აქვს გარკვეული შეზღუდვები, კერძოდ:

მონაცემების ბალანსი: თუ ანომალიური მონაცემები ძალიან მცირეა (რაც ხშირად ხდება), მოდელმა შეიძლება ვერ ისწავლოს მათი ამოცნობა. ამ შემთხვევაში საჭიროა მონაცემების ბალანსირება (მაგალითად, SMOTE ტექნიკის გამოყენებით). გამოთვლითი სირთულე: დიდი მონაცემთა ნაკრებებისთვის შემთხვევითი ტყე შეიძლება მოითხოვდეს დიდ გამოთვლით რესურსებს.

საბოლოოდ, შეგვიძლია გავაკეთოთ დასკვნა, რომ განხილული და დამუშავებული მოდელები, რაც დაკავშირებულია დიდ მოცულობის მონაცემთა ნაკრებების უსაფრთხოების ამოცანებში შემთხვევითი ტყის გამოყენებასთან, არის ძლიერი და საიმედო არჩევანი ქსელური ანომალიების აღმოსაჩენად, განსაკუთრებით მაშინ, როდესაც საქმე გვაქვს კომპლექსურ და არაწრფივ მონაცემებთან.

გამოყენებული ლიტერატურა

Mark Talabis, Jason Martin, Robert McPherson, Inez Miyamoto, Information Security Analytics: Finding Security Insights, Patterns, and Anomalies in Big Data, ISBN- 978-0128002070, 2014.

Pradip Kumar Das, Privacy and Security Issues in Big Data, ISBN 978-981-16-1006-6, 2021.

Clarence Chio, David Freeman, Machine Learning and Security: Protecting Systems with Data and Algorithms, ISBN - 978-1491979907, 2018.

Bojan Kolosnjaji, Huang Xiao, Peng Xu, Apostolis Zarras, Artificial Intelligence for Cybersecurity, ISBN - 978-1805124962, 2024.

Ronny Hänsch, Handbook of Random Forests, ISBN – 978–981-322-405-6, 2025.