**MOKSLAS – LIETUVOS ATEITIS**
**SCIENCE – FUTURE OF LITHUANIA**
ISSN 2029-2341 / eISSN 2029-2252
*http://www.mla.vgtu.lt*

ELEKTRONIKA IR ELEKTROTECHNIKA
ELECTRONICS AND ELECTRICAL ENGINEERING

2015 7(3): 340–344

Telekomunikacijų inžinerija T 180
Telecommunications Engineering T 180

http://dx.doi.org/10.3846/mla.2015.796

# TINKLO SRAUTO ANOMALIJŲ IDENTIFIKAVIMAS, TAIKANT KLASIFIKAVIMO METODUS

**Donatas RAČYS**[1], **Dalius MAŽEIKA**[2]
*Vilniaus Gedimino technikos universitetas*
*El. paštas: [1]Donatas.Racys@bite.lt; [2]Dalius.Mazeika@vgtu.lt*

**Santrauka.** Straipsnyje nagrinėjama kompiuterių tinklo srauto anomalijų atpažinimo problema. Apžvelgiami kompiuterių tinklų anomalijų aptikimo metodai bei aptariami jų privalumai ir trūkumai. Naudojant *IBM SPSS Modeler* programų paketą sudarytas nagrinėjamo tinklo srauto anomalijų atpažinimo modelis, pritaikytas SNMP protokolu pagrįstiems maršruto parinktuvo duomenims apdoroti. Pagal tris klasifikavimo metodus ir skirtingus mokymui skirtus duomenų rinkinius atlikti skaičiavimai tinklo anomalijoms identifikuoti. Palyginant gautus rezultatus nustatyta, kad C5.1 sprendimo medžio algoritmas yra tiksliausias ir sparčiausias, todėl ir tinkamiausias tinklo srauto anomalijoms atpažinti.

**Reikšminiai žodžiai:** anomalijų atpažinimas, klasifikavimo metodai, kompiuterių tinklai.

## Įvadas

Didėjant tinklo paslaugų svarbai, keliami vis aukštesni reikalavimai paslaugų kokybei garantuoti, todėl itin svarbu, kad duomenų perdavimo tinklas veiktų patikimai, saugiai ir nuolat. Duomenų perdavimo srautas gali sutrikti, t. y. nukrypti nuo įprasto pasiskirstymo laike ir sukelti duomenų perdavimo anomalijų (Anukool *et al*. 2004; Katzela *et al.* 2005). Anomalijų gali kilti dėl įvairių priežasčių. Jos klasifikuojamos į tris apibendrinamąsias kategorijas: anomalijas, įvykstančias dėl antplūdžio, anomalijas, susijusias su piktavališka veikla ir kenksmingu programiniu kodu, bei anomalijas, atsirandančias dėl fizinių ar programinių tinklo infrastruktūros problemų (Barford, Plonka 2001).

Tinklo stebėjimo bei įsilaužimų aptikimo sistemos privalo aptikti ir identifikuoti galimus nuokrypius nuo įprastinių tinklo veikimo normų ir užkirsti kelią potencialioms grėsmėms. Per vėlai lokalizuotas ir pašalintas incidentas gali neigiamai paveikti pavienius tinklo elementus, o neretais atvejais – ir didelę dalį tinklo. Siekiant pagerinti tinklo patikimumą ir valdymą didelės spartos perdavimo tinkluose, siūloma naudoti išmaniąsias tinklo stebėjimo sistemas, taikančias adaptyviuosius metodus (Miluocheva, Muller 2003).

Dauguma tinklo stebėjimo ir valdymo sistemų yra pakankamai tikslios ir patikimos, tačiau dažniausiai jų veikimas pagrįstas tuo, kad pranešimai apie vykstančius gedimus generuojami viršijus nustatytus veiklos rodiklių slenksčius (Lazar *et al.* 2012). Tokie metodai nėra lankstūs. Be to, neaptinkamos nežinomos anomalijos. Darbe nagrinėjama klasifikavimo metodų taikymo galimybė tinklo anomalijoms aptikti, atliekamas tyrimas taikant sprendimų medžio, neuronų tinklų ir Bajeso tinklo metodus, gauti rezultatai lyginami, nustatomas tiksliausias metodas.

## Tinklo duomenų šaltiniai ir protokolai

Kompiuterių tinklo anomalijoms atpažinti reikiami duomenys surenkami iš tinklo įrenginių. Duomenys gali būti išgaunami panaudojant tinklo zondus, maršruto parinktuvus, „medaus puodynės“ tipo sistemas ir ugniasienes (Markopoulou *et al.* 2004). Kiekvienas šaltinis generuoja duomenis apie įvykius ir juos kaupia.

Egzistuoja du pagrindiniai tinklo srauto stebėjimo metodai: aktyvusis ir pasyvusis. Aktyvieji metodai dažniausiai taikomi kokybinėms tinklo charakteristikoms, pvz., vėlinimui ar pralaidumui, nustatyti. Tačiau jie gali būti parankūs ir reikiamiems duomenims surinkti tiriant anomalijas. Aktyviojo matavimo metodų privalumas yra tai, kad tam nereikia specializuotosios techninės įrangos, o programinė įranga nesudėtinga. Tačiau reikia atkreipti dėmesį, kad taikant aktyviuosius metodus siunčiami *ICMP*, *TC*P ar *UDP* paketai trikdo tinklo darbą, todėl gaunami duomenys netikslūs, o anomalijų atpažinimo rezultatų

2015 © Straipsnio autoriai. Leidėjas VGTU leidykla „Technika“.
Šis straipsnis yra atvirosios prieigos straipsnis, turintis Kūrybinių bendrijų (*Creative Commons*) licenciją (CC BY-NC 4.0), kuri leidžia neribotą straipsnio ar jo dalių panaudą su privaloma sąlyga nurodyti autorių ir pirminį šaltinį. Straipsnis ar jo dalys negali būti naudojami komerciniams tikslams.

paklaida didesnė (Landfeldt *et al.* 2000). Pasyviaisiais metodais tinklo srautai analizuojami naudojant specializuotą techninę įrangą. Kaip pavyzdį galima paminėti maršruto parinktuvus arba tinklo pasiklausymo įrenginius, kurie duomenų srautą nukreipia į tinklo stebėjimo stotį. Pasyviųjų metodų privalumai:

- tinklo srautų stebėjimo metu nėra siunčiami papildomi duomenų paketai, todėl nesutrikdomas stebimo tinklo veikimas;
- surenkama išsami informacija apie tinklo protokolo parametrus ir perduodamos informacijos turinį, kurią galima efektyviai panaudoti anomalijoms atpažinti.

Įvertinus aktyviųjų ir pasyviųjų metodų privalumus ir trūkumus, buvo nuspręsta duomenims surinkti taikyti pasyvųjį metodą.

Tinklo įrenginiai renka ir kaupia duomenis apie tinklo srautą, taikydami *SNMP*, *CMIP*, *NetFlow* protokolus. *SNMP* protokolas yra vienas iš pagrindinių tinklo valdymo protokolų, reglamentuotas *RFC 6353* standartu. Tai taikomojo lygmens protokolas, skirtas tinklo įrenginiams valdyti ir stebėti. Protokolas veikia klientas–serveris principu, jungtims naudoja 161 ir 162 prievadus. *SNMP* protokolo branduolį sudaro nesudėtingų operacijų rinkinys ir taisyklės, aprašančios, kaip šios operacijos turi būti vykdomos. Šis protokolas naudojamas maršruto parinktuvams, *UNIX* bei *Windows* serveriams valdyti bei duomenims apie tinklo įrenginį gauti. Pagal *SNMP* protokolą veikiantis tinklas susideda iš trijų pagrindinių komponentų: agento, valdytojo ir valdymo informacinės bazės, kurioje tinklo įrenginys kaupia duomenis apie įvykius. Šiuos duomenis galima naudoti tinklo anomalijoms atpažinti. Bendrasis valdymo informacinis protokolas *CMIP* buvo pasiūlytas kaip alternatyva *SNMP* protokolui. Jis sudėtingesnis ir visapusiškesnis nei *SNMP* protokolas, tačiau sunkiau įsisavinamas ir valdomas. *CMIP* protokolas skirtas informacijai perduoti naudojant protokolinių duomenų (*PDU*) kintamuosius. *CMIP* protokolas reglamentuotas *RFC 1189* ir *RFC 1095* dokumentuose.

**Tinklo srauto anomalijų atpažinimo metodai**

Anomalijų, arba kompiuterių tinklo srautų nukrypimų, aptikimas bei prognozavimas yra aktuali tyrimų sritis (Anukool *et al.* 2004; Roughan *et al.* 2004). Išanalizavus tinklo srauto anomalijų atpažinimo metodus buvo išskirtos tokios metodų grupės:

- taisyklių metodai;
- klasifikavimo metodai;
- klasterizavimo metodai;
- statistiniai metodai;
- kiti metodai.

Taisyklėmis grįsti metodai taikomi žinių sistemose. Remiantis eksperto sudarytomis taisyklėmis pateikiamos išvados apie nekorektišką tinklo veikimą. Tačiau taisyklėmis grįstos sistemos yra lėtos ir priklauso nuo žinių bazės dydžio bei korektiškumo. Jei įvykęs konkretus tinklo veikimo nuokrypis nuo įprastinių normų nėra įrašytas į žinių duomenų bazę, jis tiesiog nebus identifikuotas. Šiems trūkumams eliminuoti naudojami neraiškiosios aibės kognityvieji žemėlapiai.

Klasifikavimo metodai grįsti tuo, kad jie suformuojamas klasifikavimo modelis panaudojant specialias mokymui skirtas einamųjų ir anomaliją apibūdinančias duomenų klases. Toliau tikslingai suformuotas, „apmokytas“, modelis taikomas nežinomiems duomenims ar įvykiams klasifikuoti. Klasifikavimo metodai skirstomi į kontroliuojamus ir pusiau kontroliuojamus. Taikant pirmuosius būtina turėti žinių apie einamąsias ir anomalias duomenų klases, o taikant pusiau kontroliuojamus – žinių tik apie einamąsias duomenų klases. Klasifikavimo metodų grupei priklauso neuronų tinklai, Bajeso metodas, sprendimų medžiai, maksimaliõs entropijos pagrindu grįsti metodai. Klasifikavimo metodų privalumai: gebėjimas aptikti dar nežinomas anomalijas, didelis tikslumas, greita klasifikavimo fazė.

Klasterizavimo metodų tikslas – sugrupuoti panašius duomenis į klasterius, o nepriklausantys klasteriams duomenys arba labai maži klasteriai laikomi anomalijomis. Klasterizavimo metodai skirstomi į pusiau kontroliuojamus – tai iš anksto sudaromi duomenų klasteriai, apibūdinantys normalią sistemos veiklą, ir nekontroliuojamus metodus, t. y. tokius, kai atlikus klasterizavimą būtini papildomi žingsniai klasterių dydžiams bei atstumams tarp jų įvertinti bei anomalijoms priklausantiems taškams nustatyti. Pagrindiniai klasterizavimo metodų trūkumai – tai skaičiavimo išteklių imlumas ir šio tipo metodų neveiksmingumas, jei einamieji duomenys nesitelkia į klasterius.

Statistiniai metodai remiasi prielaida, kad einamųjų įvykių duomenys patenka į stochastinio modelio didelės tikimybės zoną. Siekiant nustatyti duomenų anomalijas, tikrinama, ar jie priklauso sudarytam modeliui. Duomenys, turintys mažą tikimybės vertę, laikomi anomaliais. Statistiniai metodai yra dviejų tipų: parametriniai ir neparametriniai. Parametriniuose metoduose einamieji duomenys ir galimos duomenų anomalijos sugeneruojamos iš pagrindinių parametrinių skirstinių, o parametrai apskaičiuojami modelio formavimo, „apmokymo“, metu. Pagal statistinius neparametrinius metodus daroma prielaida, kad modelio struktūra nėra žinoma iš anksto ir nustatoma pagal

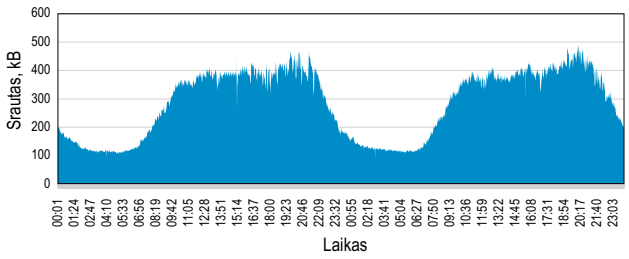
turimus duomenis. Tokio tipo metodai yra tikslesni nei parametriniai metodai. Pagrindinis statistinių modelių trūkumas yra tai, kad parametriniai įverčiai dažniausiai neatitinka realių duomenų pasiskirstymo, todėl šių modelių tikslumas labai priklauso nuo pradinio duomenų pasiskirstymo.

Išanalizavus anomalijų aptikimo metodus buvo nuspręsta tinklo srauto anomalijoms nustatyti taikyti šiuos klasifikavimo metodus: neuronų tinklą, Bajeso tinklą ir sprendimų medžius. Šie metodai yra gana lankstūs, anomalijos atpažįstamos esant įvairiems tinklo srauto nukrypimams. Atliktas metodų tikslumo tyrimas naudojant įvairius mokymo duomenų rinkinius.

**Duomenų surinkimas tirti**

Ieškant patikimiausių anomalijų nustatymo būdų buvo analizuojama maršruto parinktuvo fizinių sąsajų apkrova, t. y. įeinantis ir išeinantis duomenų srautas. Pagal *SNMP* protokolą srauto duomenys perduodami į *PRTG* (angl. *Paessler Router Traffic Grapher*) tinklo stebėjimo serverį, ten apdorojami ir grafiniu pavidalu atvaizduojami. Naudojami du pagrindiniai identifikatoriai, aprašantys maršruto parinktuvo fizinių sąsajų apkrovą, t. y. bendras per fizinę maršruto parinktuvo sąsają gautas ir perduotas baitų kiekis. 1 pav. pateiktas įprastinis dviejų dienų trukmės maršruto parinktuvo įeinančių duomenų srautas. Matyti įvairūs srauto apkrovos svyravimai tinkle, priklausantys nuo paros meto ir kitų veiksnių. Remiantis tyrime taikomos sistemos duomenimis, srautas laikomas normaliu, jei kinta nuo 100 Mbps iki 500 Mbps.

Atliekant anomalijų aptikimo analizę svarbus etapas yra parengti duomenis, nes nuo to priklauso tyrimų rezultatų tikslumas. Tyrimui buvo suformuoti mokymui skirti duomenys ir aktualaus tiriamo srauto duomenys. Rengiant duomenis buvo atlikti šie etapai: duomenų atranka, duomenų struktūrizavimas ir duomenų formato keitimas. Analizei reikiami duomenys buvo imami *XML* formato iš tinklo stebėjimo serverio ir filtruojami, nes byloje yra didelis kiekis perteklinių, šiam tyrimui nerikalingų, duomenų.




1 pav. Tiriamo maršruto parinktuvo įeinančių duomenų srautas

Fig. 1. Incoming network traffic of analyzed router

Tiriant klasifikavimo metodų efektyvumą buvo taikomas „mokymo su mokytoju“ metodas, t. y. apibrėžtos einamųjų ir anomaliųjų duomenų klasės, naudotos modeliui formuoti. Einamųjų duomenų rinkinius sudarė maršruto parinktuvo, įeinančio, išeinančio srauto ir abiejų šių srautų sumos duomenys. Mokymui buvo naudota:

1) dviejų dienų istoriniai duomenys,
2) vienos savaitės istoriniai duomenys,
3) dviejų savaičių istoriniai duomenys,
4) mėnesio istoriniai duomenys.

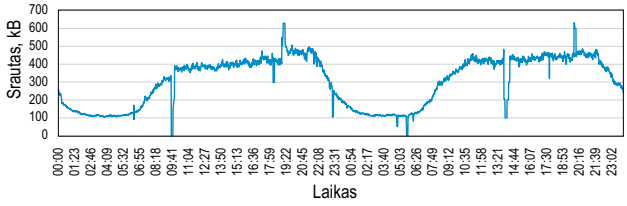
Dviejų dienų trukmės einamųjų duomenų srautas pavaizduodas 2 pav.

Anomaliems duomenims identifikuoti buvo skaičiuojamas vidutins kvadratinis nuokrypis ir, remiantis ekspertinėmis žiniomis, nustatytas leidžiamasis nuokrypis. Vidutinis kvadratinis nuokrypis apskaičiuojamas pagal formulę

$$\sigma = \sqrt{E\left[(X-\mu)^2\right]}, \qquad (1)$$

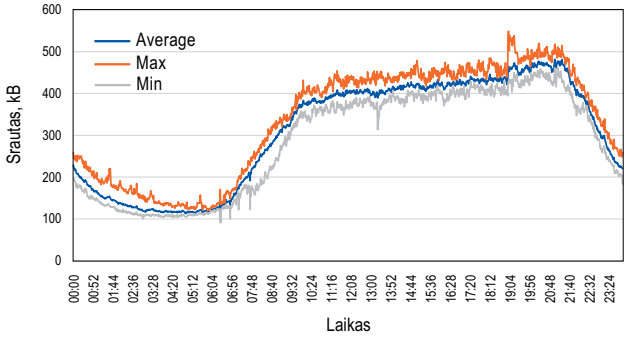
čia $E$ – vidurkio funkcija; $X$ – tinklo srauto tiriamoji vertė; $\mu$ – tinklo srauto vidurkis.

Remiantis tyrime naudojamos sistemos našumo parametrais, leistinu nuokrypiu buvo laikoma 50 Mbps riba. Taip pat buvo laikoma, kad duomenų srautas turi būti didesnis nei 75 Mbps, bet neturi viršyti 900 Mbps. Šios ribos nustatytos remiantis mėnesio trukmės istoriniais duomenimis. Grafinis toleruojamo nuokrypio atvaizdas pateiktas 3 pav.




2 pav. Dviejų dienų trukmės einamieji duomenys

Fig. 2. Two days long data used for the learning



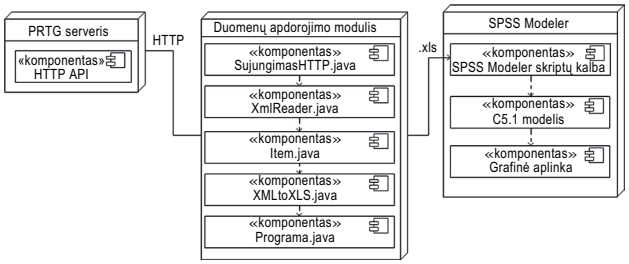

3 pav. Modeliui apmokyti skirti duomenys

Fig. 3. Model training data

### Tinklo srauto anomalijų tyrimo modelis

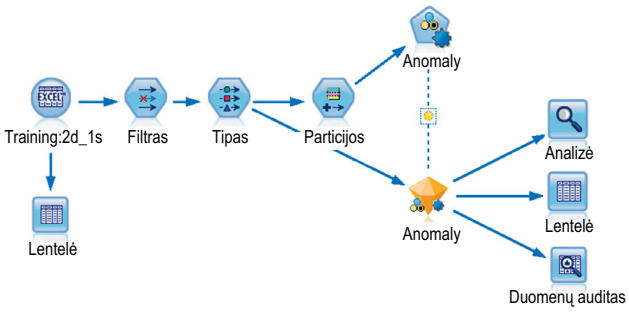
Anomalijų aptikimo galimybėms analizuoti buvo sudarytas tyrimų modelis. Jam sukurti buvo naudojama *IBM SPSS Modeler* programinė įranga bei straipsnio autorių sukurtas programinis komponentas. Modelio komponentų diagrama pateikta 4 pav.

Modelį sudaro tinklo srauto duomenų surinkimo serveris *PRTG*, duomenų apdorojimo modulis ir skaičiavimų modulis, įdiegtas *IBM SPSS Modeler* aplinkoje.



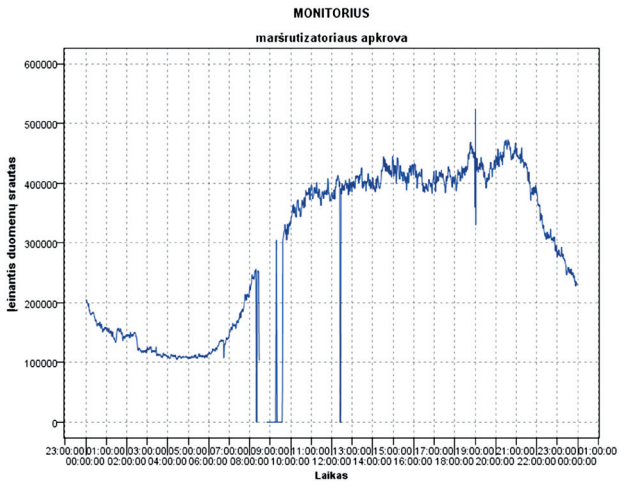

4 pav. Anomalijų atpažinimo programinio įrankio komponentai

Fig. 4. Components of software tool for anomalies detection




5 pav. Tinklo anomalijų tyrimo modelio skaičiavimų modulis

Fig. 5. Computational module of the model used for investigation of network anomalies




6 pav. Anomalijų atpažinimo programinės priemonės ekrano kopija

Fig. 6. Screenshot of software tool for anomalies detection

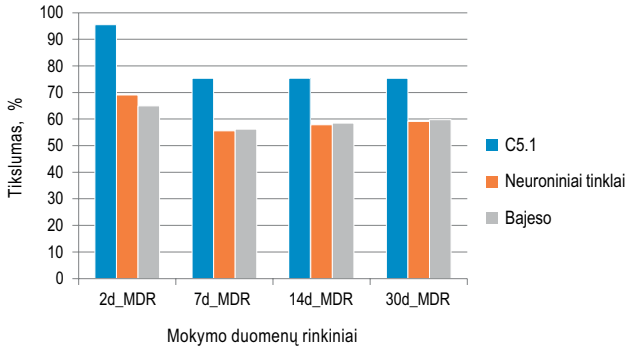
Skaičiavimų modulį sudaro analizuojamų duomenų filtras, duomenų tipą nusakantis mazgas, funkcinis mazgas, duomenis skirstantis į mokymo ir testavimo, klasifikavimo metodą įgyvendinantis mazgas bei duomenų audito ir analizės mazgai (5 pav.). Skaičiavimų modelyje buvo įdiegti Bajeso tinklo, neuronų tinklo ir C5.1 metodai. Duomenų nuskaitymas iš *PRTG* serverio vyksta pagal *HTTP* protokolą, o suformuoti duomenys perduodami į skaičiavimų modulį *xls* failu. Tinklo srautas grafiškai atvaizduotas 6 pav.

### Eksperimentinis tyrimas

Eksperimentiniam tyrimui buvo pasirinkti trys klasifikavimo metodai: neuronų tinklas, Bajeso tinklas bei C5.1 metodas, įgyvendinantis sprendimų medį. Pirmu etapu buvo formuojamas modelis t. y. skaičiuojami ryšių svoriniai koeficientai. Siekiant nustatyti, kokią įtaką modelio tikslumui daro mokymo duomenų rinkinio dydis, buvo atlikti bandymai su skirtingo dydžio duomenų kiekiais (jų gavybos procesas pateiktas ankstesniame skyriuje). Bandymų rezultatai matyti 7 paveiksle.

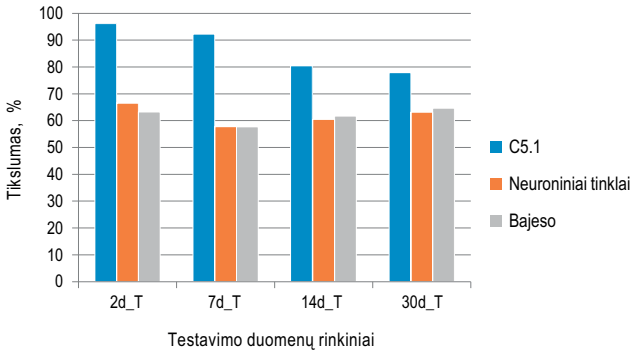
Analizuojant gautus rezultatus akivaizdu, kad visais atvejais tiksliausiai nuokrypiai identifikuojami C5.1 metodu – tikslumas siekė net 95,5 %. Iš gautų rezultatų taip pat matyti, kad didžiausias tikslumas naudojant dviejų dienų duomenis (7 pav.).

Suformuotas modelis toliau buvo naudojamas tiriant tinklo srauto anomalijų atpažinimo galimybes. Eksperimentui atlikti buvo naudojami kito laikotarpio tinklo srauto, kurio anomalijos žinomos, duomenys. Buvo tiriami dviejų, septynių, keturiolikos ir trisdešimties dienų trukmės maršruto parinktuvo duomenys. Anomalijų atpažinimo rezultatai pavaizduoti 8 pav. Analizuojant gautus rezultatus matyti, kad C5.1 metodu anomalijos atpažįstamos geriausiai. Keturių bandymų vidutinis tikslumas yra




7 pav. Modelio tikslumo tyrimo rezultatai, naudojant mokymo duomenis

Fig. 7. Results of model accuracy investigation with training data




8 pav. Anomalijų atpažinimo tyrimo rezultatai
Fig. 8. Results of anomalies detection

86,8 %, o tikimybė, kad atlikti skaičiavimai yra klaidingi, lygi tik 4,03 % (6 pav.). Antrasis pagal tikslumą yra neuronų tinklų modelis, jo tikslumas 62,2 %. Taikant Bajeso tinklo modelį anomalijos aptiktos 61,9 % tikslumu. Skaičiavimų paklaida visais atvejais neviršijo 0,1 %. Analizuojant rezultatus galima teigti, kad tikslumas taip pat priklauso nuo analizuojamų duomenų kiekio (8 pav.). Didinant duomenų kiekį, Bajeso tinklo ir neuronų tinklo metodų tikslumas didėja, o C5.1 metodo mažėja.

Apskaičiuojant tinklo anomalijas buvo matuojama kiekvieno metodo skaičiavimo trukmė. Palyginus rezultatus nustatyta, kad C5.1 metodas yra sparčiausias.

## Išvados

1. Apžvelgus anomalijų aptikimo metodus galima teigti, kad jiems būdinga skirtingas tikslumas bei duomenų apdorojimo sparta.
2. Atliktas trijų klasifikavimo metodų testavimas parodė, kad C5.1 sprendimų medžio metodu tinklo srauto anomalijos atpažįstamos tiksliausiai ir sparčiausiai. Vidutinis tikslumas siekia 86,8 %.
3. Remiantis tyrimo rezultatais nustatyta, kad tinklo srauto anomalijų atpažinimo tikslumas priklauso nuo analizuojamų duomenų kiekio. Didėjant duomenų kiekiui, Bajeso tinklo ir neuronų tinklo metodų tikslumas didėja, o C5.1 metodo mažėja.

## Literatūra

Anukool, L.; Crovella, M.; Diot, C. 2004. Diagnosing network-wide traffic anomalies, in SIGCOMM '04: *Proceedings of the 2004 conference on applications, technologies, architectures, and protocols for computer communications*, 30 August – 3 September, 2004, Portland, Oregon, USA, 219–230 http://dx.doi.org/10.1145/1015467.1015492

Barford, P.; Plonka, D. 2001. Characteristics of network traffic flow anomalies, in *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, 1–2 November, 2001, Burlingame 69–73.

Katzela, I.; Schwarz, M. 2005. Schemes for fault identification in communication networks. *IEEE/ACM Transactions on Networking*, 3: 753–764.

Landfeldt, B.; Sookavatana, P.; Seneviratne, 2000. A. The case for a hybrid passive/active network monitoring scheme in the wirel, in *8th IEEE International Conference on Networks*: 5–8 September, 2000, 139–147. http://dx.doi.org/10.1109/ICON.2000.875781

Lazar, A.; Wang, W.; Deng, R. 2012. Models and algorithms for network fault detection and identification: a review, in *Proceedings IEEE INFOCOM* 25–30 March, Orlando, Florida, USA, 121–125.

Markopoulou, A.; Iannaccone, G.; Bhattacharyya, S.; Chuah, C.; Diot, C. 2004. Characterization of failures in an IP backbone, in *23 Annual Joint Conference of the IEEE Computer and Communications Societies: INFOCOM* 2004, 7–11 March, 2004, Hohg Kong, China, 4: 2307–2317. http://dx.doi.org/10.1109/INFCOM.2004.1354653

Miluocheva, I.; Muller, E. 2003. A practical approach to forecast quality of service parameters considering outliers, in *1st Int. Workshop on Inter-Domain Performance and Simulation*, 21–21 February, 2003, Salzburg, Austria, 163–172.

Roughan, M; Griffiny, T; Mao, M.; Greenbergx, A.; Freeman B. 2004. IP forwarding anomalies and improving their detection using multiple data sources, in *ACM SIGCOMM workshop on Network troubleshooting*, 30 August – 03 September, 2004, Portland, USA, 307–312.

## NETWORK TRAFFIC ANOMALIES IDENTIFICATION BASED ON CLASSIFICATION METHODS

**D. Račys, D. Mažeika**

Abstract

A problem of network traffic anomalies detection in the computer networks is analyzed. Overview of anomalies detection methods is given then advantages and disadvantages of the different methods are analyzed. Model for the traffic anomalies detection was developed based on IBM SPSS Modeler and is used to analyze SNMP data of the router. Investigation of the traffic anomalies was done using three classification methods and different sets of the learning data. Based on the results of investigation it was determined that C5.1 decision tree method has the largest accuracy and performance and can be successfully used for identification of the network traffic anomalies.

**Keywords:** anomalies detection, classification methods, computer network