Kritička analiza KDD Cup'99 skupa podataka i metodologije istraživanja mašinskog učenja u IDS sistemima

Nemanja Maček i Milan Milosavljević

***Apstrakt*—U ovom radu data je kritička analiza skupa podataka KDD Cup '99 i metodologije istraživanja mašinskog učenja u oblasti detekcije upada zasnovanih na skupu. Iako se ovaj skup najčešće koristi za obuku, validaciju i testiranje algoritama za mašinsko učenje u IDS sistemima, podaci u skupu nisu verodostojna reprezentacija saobraćaja realne računarske mreže. Skup je isuviše složen za formiranje modela zasnovanog na mašinskom učenju koji je sposoban da dovoljno tačno detektuje R2L i U2R napade, što dovodi do kontradiktornih rezultata detekcije minornih kategorija u istraživanjima. U radu je takođe analiziran uticaj postojanja duplikata i R2L instanci identičnih normalnom saobraćaju na klasifikaciju primenom metoda vektora oslonca i ispitane su performanse klasifikatora obučenog prečišćenim skupom.**

***Ključne reči*—detekcija upada; mašinsko učenje; obučavajući skupovi; validacija; metoda vektora oslonaca.**

I. Uvod

Skup podataka KDD Cup '99 formiran je na osnovu mrežnog saobraćaja prikupljenog u MIT Lincoln laboratoriji. Projekat je pokrenula agencija DARPA (*Defense Advanced Research Projects Agency*) s ciljem ispitivanja sistema za detekciju upada 1998 i 1999 godine. Ova dva skupa podataka, nazvana DARPA98 I DARPA99, respektivno, sastoje se od *tcpdump* podataka prikupljenih sa simulirane računarske mreže; mreža je formirana tako da oponaša i zadrži karakteristike tipične računarske mreže koja pripada američkoj vojnoj bazi osrednje veličine. Podaci prikupljeni sa mreže su pretvoreni u zapise o konekcijama; svaka konekcija opisana je pomoću 41 atributa i označena kao napad određenog tipa ili normalan, nemaliciozni saobraćaj.

Skup podataka KDD Cup '99 sastoji se od potpunog obučavajućeg skupa koji sadrži 4.898.431 instanci, 10% obučavajućeg skupa (verzija obučavajućeg skupa koja sadrži 10% svih instanci), i skupa za testiranje koji sadrži 311.029 instanci). Skup za testiranje sadrži 17 novih tipova napada. Napadi na računarski sistem, odnosno mrežu, mogu se klasifikovati u četiri kategorije:

- ispitivački napadi (*probing*), pomoću kojih napadač sakuplja informacije o sistemu ili mreži sa ciljem da pronađe ranjivosti koje može da iskoristi; primeri ovih napada su ipsweep, nmap, portscan, saint i satan;
- odbijanje usluge (*Denial of Service*, DoS), koji za posledicu imaju nedostupnost resursa servisa (na primer, mrežnih servisa); primeri ovih napada su apache2, land, processtable, teardrop i udpstorm;
- *User to Root* (U2R), koji eksploatišu ranjivosti operativnih sistema ili softvera radi sticanja administratorskih (root) privilegija na sistemu; primeri ovih napada su buffer_overflow, rootkit i sqlattack;
- *Remote to Local* (R2L), pomoću kojih napadač stiče pristup udaljenom sistemu na kome nema legitiman korisnički nalog; primeri ovih napada su ftp_write, httptunnel, sendmail i xlock.

Kategorije napada su u skupovima prisutne shodno razmerama navedenim u tabeli 1.

TABELA I
Broj instanci napada po kategorijama u obučavajućim skupovima i test skupu

Kateg. napada	Obuč. skup	10% skup	Test skup
normal	492.780	97.278	60.593
probe	41.102	4.107	4.166
DoS	3.883.370	391.458	229.853
U2R	52	52	70
R2L	1126	1.126	16.347

Originalni *tcpdump* podaci iz skupa su predprocesirani, tj. razvrstani u konekcije koje su opisane pomoću 41 atributa.

Atributi se mogu svrstati u četiri kategorije. Osnovni atributi se mogu pribaviti iz zaglavlja paketa bez ispitivanja samog sadržaja paketa. Primeri ovih atributa su trajanje konekcije, protokol, servis, fleg i brojevi bajtova poslatih sa izvorišta ka odredištu, kao i sa odredišta ka izvorištu. Atributi zasnovani na sadržaju se mogu pribaviti pristupom i analizom sadržaja TCP paketa. U ove atribute, na primer, spada broj neuspešnih pokušaja prijavljivanja na sistem. Vremenski atributi saobraćaja opisuju osobine koje zastarevaju nakon isteka definisanog vremenskog intervala. Primer ovakvog atributa je broj konekcija ka istom sistemu u tom vremenskom intervalu. Za razliku od vremenskih atributa, atributi sistema su zasnovani na prozoru čiji je interval zadat brojem konekcija, a ne vremenom. Samim tim, pogodni su za opisivanje napada čije je trajanje duže od intervala propisanog vremenskim atributima.

Nemanja Maček – Visoka škola elektrotehnike i računarstva strukovnih studija, Vojvode Stepe 283, 11000 Beograd, Srbija (e-mail: macek.nemanja@gmail.com).

Milan Milosavljević – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11000 Beograd, Srbija (e-mail: mmilosavljevic@singidunum.ac.rs).

II. Kritička analiza obučavajućeg skupa

Skup podataka KDD Cup '99 je predmet kritike većeg broja autora. Ključni nedostaci ovog skupa su sledeći:

- skup nije verodostojna simulacija realnog mrežnog saobraćaja;
- obučavajući skupovi i test skup su isuviše složeni za formiranje modela zasnovanog na mašinskom učenju koji je sposoban da dovoljno tačno detektuje R2L i U2R napade [1];
- postojanje duplikata u obučavajućem i test skupu.

McHugh [2] ukazuje na činjenicu da podaci obučavajućih skupova i test skupa nisu verodostojna reprezentacija saobrćaja realne računarske mreže iz sledećih razloga:

- struktura simulirane mreže ne dozvoljava izvršenje nekih napada iz skupa,
- broj instanci napada je prevelik u odnosu na broj instanci normalnog saobraćaja i
- odnos kategorija napada u skupu nije realističan.

Shodno tome, performanse sistema obučenog ovim skupom mogu biti slabe u realnom radnom okruženju. Shodno nerealističnom broju instanci napada u odnosu na normalan saobraćaj, neki od autora predlažu filtriranje podataka kako bi prilagodili obučavajući skup realnom okruženju. Primeri filtriranja dati su u istraživanjima vezanim za detekciju anomalija [3], [4], [5] zato što je za detekciju anomalija potreban znatno manji broj anomalija (napada) kako bi one bile uspešno detektovane.

Veoma mali broj instanci U2R napada se nalazi u skupovima: 52 u potpunom i 10% obučavajućim skupovima i 70 u test skupovima. Ukoliko se formiraju podskupovi za obučavanje i test slučajnim odabirom, instance u test skupu mogu biti značajno različite od instanci u obučavajućem skupu, što njihovu detekciju otežava ili čini nemogućim;

Neke instance R2L napada su identične ili veoma slične normalnom saobraćaju [6] što otežava njihovu detekciju i dovodi do pogrešne klasifikacije [7], [8], [9]. 8,054 instalnci normalnog saobraćaja identične su instancama snmpgetattack napada, pri čemu se 7,273 instanci nalazi u test skupu, a ostale u obučavajućem skupu (10% verzija). Instance spy i warezclient R2L kategorije napada ne postoje u test skupu. U skupovima postoje samo dve instance napada tipa spy i 1,020 instanci napada tipa warezclient, što čini 90,59% svih R2L instanci u obučavajućem skupu. Osim toga, u test skupu se nalazi znatno veći broj instanci R2L napada u odnosu na obučavajući skup (1.126 i 16.347 instanci, respektivno) i skoro dvostruko više tipova R2L napada (8 i 14 tipova, respektivno). Ove činjenice ukazuju na slabu mogućnost detekcije R2L napada u slučaju korišćenja nemodifikovanog obučavajućeg skupa.

Duplikati instanci u skupu negativno utiču na proces obuke algoritama za mašinsko učenje. Skupa za obuku veštačkih neuronskih mreža metodom propagacije greške unazad treba da sadrži raznovrsne instance [10], [11]. Duplikati smanjuju raznovrsnost, ali ni Haykin ni LeCun nisu analizirali njihov uticaj na obuku klasifikatora. Duplikati u obučavajućem skupu analizirani u [12] tako što je eksperimentalno ispitan njihov uticaj na obuku Naive Bayes algoritama i neuronskih mreža namenjenih za detekciju neželjene elektronske pošte. Autori su zaključili da duplikati negativno utiču na tačnost klasifikatora i da ih je neophodno ukloniti. Kategorija DoS napada sadrži najveći broj duplikata, zbog same prirode napada..

III. Analiza izbora alternativnih obučavajućih skupova i metoda validacije na detekciju U2R i R2L kategorija napada

Osim nedostataka u originalnom obučavajućem skupu, odabir alternativnih skupova podataka koji se koriste za obučavanje algoritama za mašinsko učenje takođe dovodi do nepreciznih i kontradiktornih rezultata istraživanja.

Obučavajući skup se može formirati na nekoliko načina:

- formiranjem manjeg podskupa od obučavajućeg skupa,
- korišćenjem samo originalnog obučavajućeg skupa,
- korišćenjem originalnih skupova za obuku i testiranje,
- formiranjem novog obučavajućeg skupa kao unije skupova za obuku i testiranje i
- filtriranjem instanci s ciljem postizanja određene proporcije instanci napada i normalnog saobraćaja.

Poslednji pristup je korišćen za značajno ili potpuno smanjenje broja napada u odnosu na normalan sadržaj u istraživanjima vezanim za nenadgledanu detekciju anomalija [3], [4], [5].

Tačnost detekcije normalnog saobraćaja i ispitivačkih napada je vrlo slična nevezano od ostalih korišćenih podskupova. Tačnost detekcije R2L i U2R napada i učestalosti lažnih alarma zavise od odabranog podskupa.

Eksperimenti izvršeni na manjim podskupovima [13], [14], [15], [16] ukazuju na visoku stopu stvarnih alarma (TPR) u odnosu na eksperimente izvršene nad ostalim podskupovima: 48–68% za U2R i 84,19–95% za R2L kategorije napada. Proces odabira podskupova nad kojima su izvršeni eksperimenti nije detaljno opisan u radovima, što sprečava dalju analizu.

U radu [17] nisu objavljeni rezultati klasifikacije za stabla odlučivanja, već rezultati hibridnog sistema koji je eksperimentalno testiran samo na obučavajućem skupu. Treba naglasiti da se u istraživanjima zasnovanim samo na obučavajućem skupu koriste metode holdout i unakrsne validacije.

Rezultati eksperimenata izvšenim na test skupu [18], [19], [20], [21] su slični za U2R kategoriju (0–13.20%), ali značajno niži za R2L kategoriju napada (0.52–27%). Razlozi smanjenja performansi klasifikatora opisani su u delu 4.1.3 i potiču od velike sličnosti nekih instanci R2L napada normalnom saobraćaju kao i od neadekvatne raspodele instanci R2L kategorije.

Formiranje novog obučavajućeg skupa kao unije skupova za obuku i testiranje [1] dovodi do najveće stope detekcije: 87,50–89,28% za U2R i 99,18–99,44% za R2L kategoriju napada. Verodostojnost rezultata zasnovanih na ovakvom skupu je predmet diskusije, jer u test skupu ne postoje novi napadi.

IV. KONTRADIKTORNI REZULTATI DETEKCIJE U2R I R2L KATEGORIJA NAPADA

Uzevši u obzir činjenicu da originalni test set sadrži 17 novih napada, očekivano je da se rezultati istraživanja drugih autora razlikuju. Međutim rezultati nisu samo različiti, već su u nekim slučajevima i kontradiktorni. Na primer, u radu [22] autori tvrde da veštačke neuronske mreže ne mogu da detektuju U2R i R2L kategorije napada, dok su u [15] prijavljene visoke stope detekcije: 48% za U2R i 95% za R2L kategorije. Slično, u [1] naznačeno je da stabla odlučivanja dostižu visoke stope detekcije (99,18% za U2R i R2L kategorije), dok u [19] autori iznose veoma niske stope (10,09% i 0,56% za U2R i R2L kategorije, respektivno)

V. UTICAJ DUPLIKATA I R2L INSTANCI IDENTIČNIH NORMALNOM SAOBRAĆAJU NA KLASIFIKACIJU

.Duplikati u obučavajućem skupu nastaju kao posledica same prirode određenih napada (uglavnom ispitivačkih i DoS napada). Duplikati negativno utiču na obučavanje klasifikatora i zbog toga ih treba ukloniti iz obučavajućih skupova. Međutim, postoje razlozi za i protiv uklanjanja duplikata iz test skupa. Postojanje duplikata verodostojnije oslikava realni mrežni saobraćaj, ali istovremeno i umanjuje verodostojnost rezultata testiranja klasifikatora. Na primer, ako je većina napada koji je prestavljeni jednom instancom sa velikim brojem duplikata ispravno detektovani, tačnost detekcije klasifikatora značajno raste. Ukoliko su takvi napadi pogrešno klasifikovani kao normalan saobraćaj, tačnost detekcije opada. Klasifikacija obavljena pomoću takvih skupova može dovesti do neispravnih rezultata eksperimenata. Ukoliko duplikati postoje u test skupu, potrebno je da realno oslikavaju mrežno okruženje u kome se nalazi IDS sistem.

Analiza uticaja duplikata na klasifikaciju obavljena je uklanjanjem duplikata iz 10% obučavajućeg skupa i test skupa. Broj duplikata i broj instanci nakon uklanjanja dat je u tabelama 2 i 3.

TABELA II
BROJ DUPLIKATA U 10% OBUČAVAJUĆEM I TEST SKUPU

Kategorija	10% skup	Test skup
normal	4.896	12.680
probe	1.976	1.484
DoS	336.886	206.267
U2R	0	0
R2L	127	13.289

TABELA III
BROJ INSTANCI NAPAD U 10% OBUČAVAJUĆEM I TEST SKUPU NAKON UKLANJANJA DUPLIKATA

Kategorija	10% skup	Test skup
normal	87.832	47.413
probe	2.131	2.682
DoS	54.572	23.568
U2R	52	70
R2L	999	3.058

Slični eksperimenti obavljeni na spojenim obučavajućim i test skupovima upotrebom Naïve Bayes metode i stabla odlučivanja opisani su u [23].

Testiranje je izvršeno upotrebom osnovnog modela SVM klasifikatora i modela proširenog težinskim koeficijentima atributa, tj. njihovim značajem za klasifikaciju. Slično višeslojnim perceptronima, metode vektora oslonaca [24], [25] su algoritmi nadgledanog mašinskog učenja koji se u skorije vreme sve češće primenjuju u sistemima za detekciju upada. Metode vektora oslonaca su pogodne za primenu u IDS sistemima zbog visoke efikasnosti učenja iz višedimenzionih skupova podataka [26] bez opasnosti od dovođenja u stanje prenaučenosti, kao i zbog velike brzine i tačnosti klasifikacije. Optimalni hiperparametri (granica meke margine i parametar RBF jezgra) određeni su unakrsnom validacijom i grid-search pretragom. Eksperiment je izvršen upotrebom softverskih paketa LibSVM (verzija 3.16) i MATLAB (verzija 7.12, R2011a) i operativnog sistema Windows 7 Professional na računaru sa Intel(R) Core(TM)2 Duo CPU @ 1.66GHz procesorom i 8GB radne memorije. Rezultati testiranja su dati u tabeli 4.

TABELA IV
UTICAJ DUPLIKATA INSTANCI NA KLASIFIKACIJU

model	skup	tačnost	FNR
osnovni	originalan	97,90%	1,37%
osnovni	prečišćen	95,02%	2,21%
prošireni	originalan	99,07%	0,51%
prošireni	prečišćen	98,86%	0,72%

Umanjenje tačnosti primetnije je u slučaju osnovnog modela, što ukazuje na postojanje velikog broja duplikata instanci u test skupu koje su ispravno detektovane pre uklanjanja. Umanjenje tačnosti proširenog modela nakon uklanjanja duplikata nije toliko značajno. Na osnovu toga se može zaključiti da uticaj duplikata zavisi i od klasifikatora koji se koristi.

Neke instance R2L napada su identične ili veoma slične normalnom saobraćaju što otežava njihovu detekciju i dovodi do pogrešne klasifikacije. Konkretno 8,054 instalnci normalnog saobraćaja identične su instancama snmpgetattack napada, pri čemu se 7,273 instanci nalazi u test skupu. Postoji nekoliko uzroka pojave identičnih instanci: ograničenja u tcpdump podacima (nedovoljno informacija na osnovu kojih bi se maliciozno ponašanje razlikovalo od normalnog) i pogrešna dodela oznake atributa instanci. U [6] naznačeno je da je pojava identičnih instanci posledica i ograničenja pri transformaciji DARPA tcpdump podataka u KDD Cup '99 skupove; autori ističu značaj uklanjanja ovih instanci ali ne predlažu alternativnu transformaciju podataka koja bi rešila ovaj problem. Uklanjanjem ovih instanci empirijski je utvrđen njihov uticaj na tačnost klasifikacije.

Analiza uticaja idetntičnih instanci na klasifikaciju obavljena je uklanjanjem tih instanci iz 10% obučavajućeg skupa i test skupa. Testiranje je izvršeno upotrebom osnovnog

modela klasifikatora i proširenog modela SVM klasifikatora. Slični eksperimenti opisani su u [6], [23]. Rezultati testiranja su dati u tabeli 5.

TABELA V
UTICAJ INSTANCI R2L NAPADA IDENTIČNIH NORMALNOM SAOBRAĆAJU NA KLASIFIKACIJU

model	skup	tačnost	R2L%
osnovni	originalan	97,90%	96,02%
osnovni	prečišćen	98,01%	97,81%
prošireni	originalan	99,07%	98,72%
prošireni	prečišćen	99,16%	99,24%

Na osnovu rezultata zaključuje se da tačnost detekcije i tačnost detekcije R2L kategorije napada uvećava i za osnovni i za prošireni model klasifikatora nakon uklanjanja instanci R2L napada identičnih normalnom saobraćaju..

VI. ZAKLJUČAK

Iako je skup podataka KDD Cup '99 često korišćen u istraživanjima vezanim za primenu metoda mašinskog učenja u sistemima za detekciju upada, skup ima nekoliko fundamentalnih nedostataka. Nesrazmeran odnos broja napada i instanci normalnog saobraćaja, odnos broja instanci različith kategorija i složenost za formiranje modela zasnovanog na mašinskom učenju koji je sposoban da dovoljno tačno detektuje R2L i U2R napade, ukazuju na nereprezentativnost realnog mrežnog saobraćaja. Na osnovu izloženog može se zaključiti da je skup upotrebljiv, ali zahteva dodatnu obradu pre obuke algoritama za mašinsko učenje.

LITERATURA

[1] M. Sabhnani, G. Serpen, "Why machine learning algorithms fail in misuse detection on kdd intrusion detection data set," Intelligent Data Analysis, pp. 403-415, 2004.

[2] J. McHugh, "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, pp. 262-295, 2000, ISSN 1094-9224.

[3] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data," Data Mining for Security Applications, Kluwer, 2002.

[4] K. Leung, C. Leckie, "Unsupervised anomaly detection in network intrusion detection using clusters," In ACSC '05, pages 333–342, Darlinghurst, Australia, 2005, Australian Computer Society, Inc.

[5] L. Portnoy, E. Eskin, S. Stolfo, "Intrusion Detection With Unlabeled Data Using Clustering," In Proceedings of the ACM Workshop on Data Mining Applied to Security, 2001.

[6] Y. Bouzida, F. Cuppens, "Neural networks vs. decision trees for intrusion detection," In IEEE / IST MonAM, 2006.

[7] I. Gadaras, L. Mikhailov, S. Lekkas, "Generation of Fuzzy Classification Rules Directly from Overlapping Input Partitioning," In IEEE Int. Fuzzy Systems Conf., pp. 1–6, 2007.

[8] C. L. Liu, "Partial discriminative training for classification of overlapping classes in document analysis," Int. Journal on Document Analysis and Recognition, pp. 53–65, 2008, ISSN 1433-2833.

[9] S. Raudys, "Multiple Classification Systems in the Context of Feature Extraction and Selection," In MCS '02, pp. 27–41, London, UK, 2002, Springer-Verlag, ISBN 3-540-43818-1.

[10] S. Haykin, "Neural Networks: A Comprehensive Foundation, 2nd edition," Prentice Hall, 1998.

[11] Y. LeCun, L. Bottou, G. B. Orr, K. R. Mueller, "Efficient BackProp, In Neural Networks: Tricks of the Trade," pages 9–50, London, UK, 1998. Springer-Verlag. ISBN 3-540-65311-2.

[12] A. Kolcz, A. Chowdhury, J. Alspector, "Data duplication: an imbalance problem?, Workshop on Learning from Imbalanced Data Sets II, 2003", In ICML'2003.

[13] A. Bosin, N. Dessì, B. Pes, "Intelligent Bayesian Classifiers in Network Intrusion Detection," In Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence, pp. 445–447, London, UK, 2005, Springer-Verlag, ISBN 3-540-26551-1.

[14] S. Chebrolu, A. Abraham, J. P. Thomas, "Feature Deduction and Ensemble Design of Intrusion Detection Systems," Computers and Security, June 2005, pp. 295–307.

[15] S. Mukkamala, A. H. Sung, "Artificial Intelligent Techniques for Intrusion Detection," In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2003.

[16] S. Peddabachigari, A. Abraham, C. Grosan, J. Thomas, "Modeling Intrusion Detection Systems Using Hybrid Intelligent Systems," Journal of Network and Computer Applications, January 2007, pp. 114-132.

[17] O. Depren, M. Topallar, E. Anarim, M. K. Ciliz, "An Intelligent Intrusion Detection System for Anomaly and Misuse Detection in Computer Networks," Expert systems with Applications, pp. 713–722, 2005.

[18] N. Ben Amor, S. Benferhat, Z. Elouedi, Naive Bayes vs Decision Trees in Intrusion Detection Systems, In ACM SAC '04, pp. 420–424, New York, NY, USA, 2004, ISBN 1-58113-812-1.

[19] S. Benferhat, K. Tabia, "On the combination of naive bayes and decision trees for intrusion detection," In CIMCA '05, Vol-1 (CIMCA-IAWTIC'06), pp. 211–216, Washington, DC, USA, 2005, IEEE Computer Society, ISBN 0-7695-2504-0-01.

[20] M. Sabhnani, G. Serpen, "Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context," In Proceedings of MLMTA 2003, vol. 1, pp. 209–215, 2003.

[21] K. Shafi, T. Kovacs, H. A. Abbass, W. Zhu, "Intrusion detection with evolutionary learning classifier systems," Natural Computing: an international journal, pp. 3–27, 2009, ISSN 1567-7818.

[22] Z. S. Pan, S. C. Chen, G. B Hu, D. Q. Zhang, "Hybrid Neural Network and C4.5 for Misuse Detection," In Machine Learning and Cybernetics, pp. 2463–2467, Xi'an, 2003.

[23] V. Engen, "Machine Learning for Network Based Intrusion Detection," PhD thesis, Bournemouth University, 2010.

[24] C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, 2, 1998, pp. 121-167.

[25] C. Cortes, V. Vapnik, "Support-vector networks," Machine Learning, 1995, pp. 273–297.

[26] B. E. Bosner, I. M. Guyon, V. N. Vapnik, "A training algorithm for optimal margin classifier," Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144-152, Pittsburg, Pensylvania, US, 1992.

ABSTRACT

This paper criticizes KDD Cup '99 data set and machine learning IDS research methodology. Although this data set is commonly used to train, validate and test IDS machine learning algorithms, the data in the set is not a realistic representation of real network traffic. The set is too complex to form a machine learning model capable of detecting R2L and U2R categories with high detection rate, which leads to contradictory minor category detection results in research papers. This paper also presents an analysis of change in support vector machine classification abilities if duplicate instances and R2L instances identical to normal traffic are removed from the data set..

Critical analysis of KDD Cup '99 data set and machine learning IDS research methodology

Nemanja Maček, Milan Milosavljević