Методологија детекције аномалија у *BGP*-у

Маријана Ћосовић
Универзитет у Источном Сарајеву, Електротехнички факултет
Источно Сарајево, Босна и Херцеговина
marijana.cosovic@etf.unssa.rs.ba

Слободан Обрадовић
Висока школа струковних студија за информационе технологије
Београд, Србија
slobodan.obradovic@its.edu.rs

Садржај— **У раду је описана методологија детекције аномалија у BGP-у (енг. *Border Gateway Protocol*). Проблем детекције аномалија присутан је у различитим доменима. Важност методолошког приступа је од великог значаја у домену детекције аномалија у *BGP*-у како за истраживаче тако и за мрежне оператере. Приказан је дијаграм тока детекције аномалија у *BGP*-у и детаљно су разрађене три фазе процеса. Прву фазу чине одабир података, обрада података и екстракције својстава. Затим слиједе анализа прикупљених и обрађених података из прве фазе примјеном метода селекције својстава и метода машинског учења. У посљедњој фази се интерпретирају добијени резултате у складу са мјерама евалуације модела учења.**

Кључне ријечи – BGP; детекција аномалија; методологија

I. Увод

Процес откривања аномалија је актуелан у различитим подручјима па се истражује унутар различитих научних области. Са порастом ресурса за похрањивање података јавља се и потреба да се у различитим гранама науке постојећи проблеми решавају путем аутоматске обраде података, која би за циљ имала олакшање у самом процесу откривања аномалија. У овом раду користе се технике машинског учења за откривање аномалија у комуникационим мрежама. *BGP* је међудоменски протокол за рутирање и основа тренутне инфраструктуре Интернета. У систему попут Интернета аномалије могу имати драматичне посљедице, као што су прекид везе и губитак стабилности на локалном и глобалном нивоу. Откривање аномалија у *BGP* протоколу је врло занимљиво поље истраживања, како са економског тако и са академског аспекта. За истраживаче и администраторе, исправна идентификација аномалија је од суштинског значаја за очување података и услуга.

BGP је протокол рутирања, који се користи за усмјеравање на цијелом Интернету [1]. Интернет се састоји од великог броја одвојених мрежа, званих аутономни системи (енг. *Autonomous System, AS*). *AS* је повезана група једног или више *IP* (енг. *Internet Protocol, IP*) префикса (блокова *IP* адреса), којим управљају један или више мрежних оператера са јединственом и јасно дефинисаном политиком рутирања на Интернету [2]. *IANA* (енг. *Internet Assigned Numbers Authority, IANA)* је међународна организација која се бави расподјелом *IP* адресног простора [3]. Сваки *AS* има јединствени регистрацијски број (енг. *Autonomous System Numbers*, *ASN*) који додијељује *IANA*. Број *AS*-ова се повећао за петнаест пута од 1999. године и тренутно износи преко 75.000 [4].

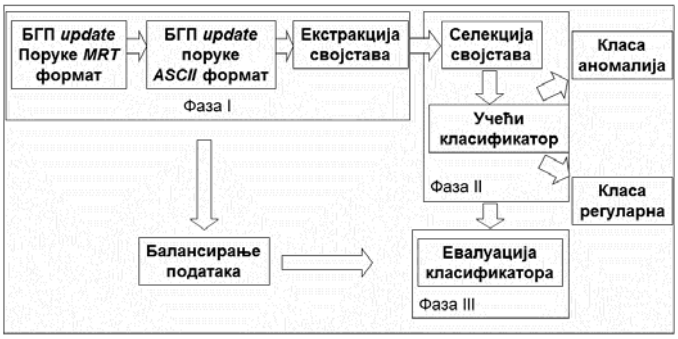
Интернет се може представити као граф у ком су аутономни системи чворови графа, а гране се формирају на основу информација које размјењују *BGP* рутери. *BGP* рутери у свом раду могу размјењивати четири врсте порука: *Open*, *Update*, *KeepAlive* и *Notification*. Сваки рутер садржи информације о путањама (енг. Routing Information Base, *RIB*), на основу којих одређује путеве усмјеравања пакета. С обзиром да *BGP* рутери одржавају табеле рутирање цијелог Интернета, неопходно је да су то уређаји са великим меморијским капацитетом. За одлуке рутирања и одабир путање, *BGP* се ослања на шири скуп информација познатих као атрибути путање. Атрибуте путање рутери размјењују у *update* порукама. Поред атрибута путање, велику улогу у одабиру путање, која се извршава путем алгоритма за одабир путање, имају мрежни администратори. Они филтрирају информације о усмјеравању тако што конфигуришу рутере у складу са политиком рутирања одређеног *AS*-а. Ове смјернице за усмјеравање саобраћаја доносе организације које поседују *AS*-ове. Може се закључити да је избор најбољег пута диктиран алгоритмом за одабир руте, али да се последња одлука у том избору доноси на бази политике. Рутирање на Интернету је базирано на моделу повјерења у коме оператери не имплементирају увијек све доступне *BGP* безбједоносне механизме провјере легитимног власништва над мрежним префиксом [5, 6]. *BGP* је у основи ефикасан протокол али имплицитно вјерује свим аутономним системима, да су информације које се преносе између сусједа тачне и да сви аутономни системи одржавају добре праксе рутирања *BGP* протоколом и добре праксе политика рутирања [7]. Добра пракса се дефинише као претпоставке понашања које нису неопходне, нити су протоколом или законом регулисане, али о њима зависи оптимално функционисање рутирања и веза на Интернету.

II. Детекција Аномалија У *BGP* Протоколу

На слици 1. приказан је дијаграм тока детекције аномалија у *BGP*-у развијен у склопу истраживања. Методологија детекције аномалија у *BGP*-у је изведена у три фазе. Прва фаза се може посматрати као процес одабирања података, обраде података и екстракције својстава. У другој фази се врши анализа прикупљених и обрађених података из прве фазе примјеном метода

селекције својстава и методама машинског учења. У посљедњој фази детекције аномалија у *BGP*-у интерпретирамо добијене резултате у складу са мјерама за евалуацију модела учења.

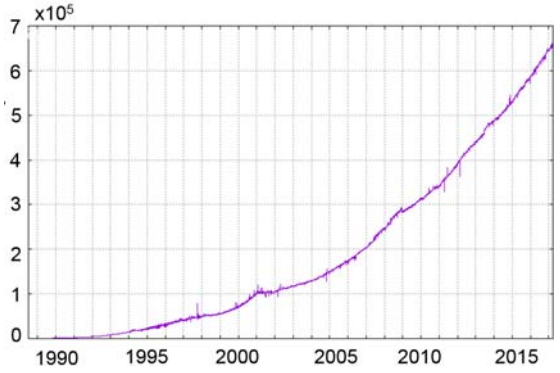
Процес одабирања података се састоји од прикупљања података о догађајима који су препознати као атипични за функционисање *BGP*-а. У припремном периоду ове фазе консултоване су специјализоване радне групе у склопу Техничке радне групе за Интернет (енг. *Internet Engineering Task Force, IETF*) [8]. Основна функција радне групе за међудоменско рутирање (енг. *Inter-Domain Routing, IDR*) [9] јесте да подржи употребу *BGP*-а за *IPv4* и *IPv6*. Поред поменуте, радна група за глобално рутирање (енг. *Global Routing Operations, GROW*) [10], такође при *IETF*-у разматра оперативне проблеме који су повезани са *IPv4* и *IPv6* глобалним системом рутирања; раст табела рутирања; ефекте интеракције унутрашњег и ванјског *BGP* протокола; утицај политике и праксе алокације адресног простора на глобани систем рутирања. Обје радне групе су биле од велике користи у првој фази за прикупљање података везаних за атипичне догађаје у *BGP*-у.



Слика 1. Дијаграм тока детекције аномалија у *BGP*-у

IETF је развио *RFC* документе (енг. *A Request for Comments, RFC*) који су, такође, били од значаја у процесу припреме одабирања података. *RFC* су формални документи који настају као резултат заједничког рада свих заинтересованих страна у одређеном пољу. Неки од *RFC* докумената су информативног карактера. Радна група за рутирање при *RIPE* (*Réseaux IP Européens, RIPE*) организацији има архиве маилинг листа од августа 2003. до данас. Такође, група мрежних оператера Сјеверне Америке (енг. *North American Network Operators' Group, NANOG*) [11] посједује архивиране маилинг листе од маја 1994. [12] и убједљиво је најбољи ресурс за анализу историјских догађаја у *BGP*-у.

Припрема података за анализу изводи се у три корака у првој фази. Први корак је процес у коме се подаци припремају на основу метода студије појединих случајева и компилације података из неколико извора. Према критеријуму квалитета података и техничких ограничења којима су подаци подложни бира се подскуп података. Историјски догађаји одређених профила, са старијим датумом, одабрани су искључиво да би се илустровала независност између тачности класификације догађаја и количине података који се обрађују. У случају аномалија новијег датума, обрађују се велике количине података, што додатно продужава процес обраде података. На слици 2 приказан је тренд раста глобалних табела рутирања, па се на основу тога може закључити да је и број путања повећан, а самим тим и садржај update порука које се анализирају. Методологија рада првог корака (прве фазе) детаљно је приказана у дијелу А.



Слика 2. Раст *BGP* табела рутирања од 1990. до 2017. године

Сљедећи корак прве фазе је обрада података. У овој фази се подаци преводе у читљив формат помоћу различитих алата. Након превођења је констатовано да је квалитет података који се користе за ово истраживање добар у највећем дијелу. У незнатном дијелу подаци су корумпирани или недостају. Проблем недостајућих података надомјешта се алтернативама које ће бити детаљније објашњене касније.

У посљедњем кораку прве фазе изабрана су својства података која су неопходна да се из њих извуку највредније информације. Методологија екстракције података је детаљно приказана у наставку дијела А. Екстрактоване су двије врсте својстава из *BGP update* порука: својства која су везана за обим *BGP update* порука, као и својства која су везана за *ASPATH* атрибут *BGP update* порука. Формирана матрица својстава у својим колонама садржи својства система који се посматра, а у редовима временске инстанце посматраног догађаја. Чишћење података је сљедећи корак ком се приступа након формирања матрице својстава. Неке од техника чишћења података које смо примјенили су нормализовање вриједности нумеричких својстава и *data smoothing*. С обзиром да посматрамо догађаје од 2001. до 2016. године нормализација података смањује утицај ефекта раста Интернета на податке. Такође је битно, за класификаторе машинског учења да, с обзиром да зависе од међусобне удаљености инстанци и дискриминативне функције, узимамо у обзир релативне, а не апсолутне вриједности својстава. Дискретизација нумеричких својстава је кориштена у процесу чишћења података као *data smoothing* техника.

Посебан корак који није експлицитно приказан на слици 2, а налази се у првој фази дијаграма тока детекције аномалија у *BGP*-у, јесте форматирање података. Неријетко се доста времена уложи на трансформацију података у формат који је диктиран одабраном техником моделовања: редослијед својстава (укључујући и циљни атрибут код бинарне класификације) који диктира метода машинског учења, синтакса улазних фајлова се разликује од једне до друге методе, итд. Балансирање података је

метода која се у случају израде методологије примијенила након завршене прве фазе и обраде података са циљем побољшања мјера евалуације. Балансирање података се могло извести паралелно оригиналним скуповима података. Дакле, од овог корака зависи и комплексност обраде података у другој фази.

Друга фаза се односи на примјену метода селекције својстава и метода машинског учења за интелигентну обраду података прикупљених у првој фази. У овој фази се прије развоја модела дефинише процедура процјењивања и бирања мјера квалитета модела. На основу овога одређујемо на који начин раздвајамо скупове података за учење модела и за његово тестирање. Сљедећи процес у овој фази се односи на развој модела машинског учења, који је итеративан по својој природи. С обзиром да га дефинишу одређени параметри, процес проналаска најбољег модела заснива се на оптималној комбинацији параметара модела машинског учења, која има најбоље перформансе на скупу података за тестирање. Детаљнији опис методологије рада друге фазе је дат у дијелу Б и односи се на селекција својстава и методе машинског учења.

У посљедњој фази детекције аномалија у *BGP*-у интерпретирамо добијене резултате у складу са мјерама за евалуацију модела учења. У овој фази рада модел се оцјењује везано за поузданост резултата на скупу података за тестирање. Мјере евалуације су детаљно приказане у поглављу Ц.

У наставку се изводе закључци у погледу могућности успјешне реализације модела машинског учења за детекцију аномалија у *BGP*-у и кориштења развијених модела у пракси. Такође, један аспект евалуације је идентификовање могућих грешака у процесу, а самим тим и могуће побољшање модела.

А. Прва Фаза

Пројекат информацијске базе рутирања (енг. *Routing Information Service, RIS*) [13] формиран је од стране RIPE NCC (енг. *Network Coordination Centre*), а *Route Views* [14] је пројекат формиран на Универзитету у Орегону, САД, у сврху колекције и похрањивања података о рутирању на Интернету. Приступ хронолошким подацима о рутирању је од користи научницима за истраживање, а и мрежним администраторима за рјешавање текућих проблема. Једна од занимљивијих примјена података о рутирању на Интернету у истраживачкој заједници је у области детекције аномалија у *BGP* протоколу.

RRC (енг. *Remote Route Collector*) софтверски рутери, реализовани на Линух платформи прикупљају *BGP* update поруке , и похрањују у MPT (енг. *Multi-threaded Routing Toolkit*) формату [15]. *BGP* даемон *Zebra/Quagga* се користе за интерфејс са *BGP* рутерима. Поред *BGP update* порука доступне су и комплетне табеле рутирања, које се генеришу сваких осам сати, и нису кориштене за потребе овог истраживања. Сједишта за размјену Интернет протока (енг. Internet eXchange Point, IXP) су инфраструктура путем које се директно повезују посредници Интернет услуга (енг. Internet Service Providers, ISPs) између аутономних система. Предности директног повезивања су, између осталог: цијена, кашњење и пропусни опсег. Сједишта за размјену Интернет протока се налазе на преко шест стотина локација широм планете [16], од којих су неки глобални мега центри попут PA-IX, DE-CIX, AMS-IX, LINX, NL-IX, и MSX-IX а неки су мањи центри регионалног карактера. *RIS* распоређује *RRC* на многа сједишта за размјену Интернет протока.

До јула 2003. године подаци у *RIB-у* регистровани су сваких петнаест минута а послије је тај интервал смањен на пет минута. Интернет стандарде је развила *IETF* која је, између осталог, дефинисала и *MRT* формат, који се користи за експортовање порука протокола рутирања и садржаја *RIB*-а.

Кад два *AS*-а желе размјењивати саобраћај требају успоставити *BGP* сесију која се остварује путем *TCP* протокола на порту 179. За сваки линк који директно повезује два рутера из различитих *AS*-ова постоји *BGP TCP* конекција. У свом раду *BGP* рутери могу размијенити четири врсте порука: *Open*, *Update*, *KeepAlive* и *Notification*. Порука типа *Open* садржи основне информације попут идентификатора рутера, кориштене верзије *BGP*-а, броја *AS*-а и њом се успоставља *BGP* веза. Порука типа *Notification* се користи ако постоји неусаглашеност у конфигурацијским параметрима као нпр. различити *AS* бројеви или *IP* адресе. Дакле, *BGP* сесија се не успоставља на прописани начин и генерише се одговарајућа порука обавјештења. Када се успостави *BGP* сесија, рутери размјењују све познате путање, користећи *update* поруку, а након тога само када дође до промјене *BGP* путања у рутинг табелама.

MRT формат је развијен за потребе стандардизованог приказа података који се између осталог користе при анализи мрежа у истраживању протокола рутирања. Све поруке *MRT* типа имају јединствено заглавље. Користи се *UNIX* формат времена и са 32-бита се може представити *UTC* вријеме у секундама од 00:00:00 (хх:мм:сс), 1. јануара 1970. године до 03:14:07, 19. јануара 2038. године. Један од типова *MRT* порука је и *BGP4MP* [17] тип поруке, оригинално дефинисан у Зебра софтверском пакету за *BGP* протокол, са подршком за мултипротоколску екстензију дефинисану у *RFC*4760, чији је подтип поруке *BGP4MP_MESSAGE*. Различити подтипови *BGP4MP_MESSAGE* порука су дефинисани за *MRT BGP* тип поруке. *Update* подтип *BGP* поруке је од интереса у овом истраживању.

Подаци који се користе у оквиру истраживања се требају прикупити са једног или више *RRC*-ова и/или *Route Views* рутера. *BGP update* поруке које су похрањене за вријеме дешавања познатих Интернет аномалија су од интереса. Поруке за познате Интернет аномалије се прикупљају у одређеним *IXP*-овима али ако посматрамо догађаје чија је видљивост на глобалном нивоу у том случају је небитно који је изворни аутономни систем, као и који *IXP* користимо. Поменути догађаји креирају нестабилности у међудоменском рутирању, које се манифестују, између осталог, оштрим и непрекидним порастом у броју објављених и повучених путања које се размјењују порукама *BGP* рутера.

Примјер једне *BGP update* поруке која се добије послије трансформације из *MRT*-а у *ASCII* формат

приказан је у табели 1. С обзиром да се поруке генеришу на нивоу сваке секунде, потребно их је објединити у једну повезану структуру за временски период од интереса. Дакле, над порукама у *ASCII* формату прво се извршава конкатенација да би се лакше манипулисало учитавањем *BGP update* порука у базу података. *SQL loader* алат је кориштен за учитавање *BGP update* порука у табеле базе података, а само учитавање је вршено секвенцијално. У току учитавања података могу се идентификовати оштећени и самим тим неупотребљиви подаци, а такође се могу у одређеним временским инстанцама регистровати подаци који недостају. Задњи проблем су идентификовали и аутори у [18] и понудили организовану презентацију неуспјелих *BGP* сесија, похрањених у склопу *RIS* и *Route Views* пројеката, које могу бити од користи истраживачима за одабир и чишћење података прије саме анализе.

ТАБЕЛА I Конверзија *BGP update* поруке из *MRT*-а у *ASCII* формат путем *bgpdump* алата

Поље	**Вриједност**
TIME	01/25/03 15:45:53
TYPE	*BGP4MP/MESSAGE/Update*
FROM	192.65.184.3 AS513
TO	192.65.185.40 AS12654
ORIGIN	*IGP*
ASPATH	513 3320 209 16738
NEXT_HOP	192.65.185.4
ANNOUNCE	198.3.128.0/24

У бази података, користећи упите се могу генерисати својства *BGP update* порука. Генерисање својстава се може извести на нивоу неког временског оквира, нпр. на нивоу сваке минуте, и у одређеном периоду нпр. у току пет дана, за све познате нападе на Интернету. У случају оваквог временског оквира трећи дан се сматра даном напада. Два дана која му претходе и два дана послије напада су, такође, кориштени за генерисање својстава и сматрају се редовним данима. Нека од својстава која се могу генерисати на основу података из *BGP update* поруке су приказана испод.

- *BGP* порука типа ‘објављена’ (*announce*) – број *BGP update* порука у току једне минуте које оглашавају доступне путање за доставу пакета
- *BGP* порука типа ‘повучена’ (*withdrawn*) – број *BGP update* порука у току једне минуте које оглашавају повучене путање за доставу пакета
- Објављени *IP* префикси – број *IP* префикса у *BGP update* поруци типа ‘објављена’ у току једне минуте
- Повучени *IP* префикса – број *IP* префикса у *BGP update* поруци типа ‘повучена’ у току једне минуте
- Дупле *BGP* поруке типа ‘објављена’ (*announce*) – број дуплих *BGP update* порука у току једне минуте које оглашавају доступне путање за доставу пакета
- Дупле *BGP* поруке типа ‘повучена’ (*withdrawn*) – број дуплих *BGP update* порука у току једне минуте које оглашавају повучене путање за доставу пакета
- Имплицитно повучене *BGP* порука – број *BGP update* порука које су типа ‘објављена’ (*announce*), али са различитим *ASPATH* атрибутом за већ објављене *IP* префиксе у току једне минуте
- *IGP* – број *BGP update* порука генерисаних од *IGP* протокола у току једне минуте
- *EGP* – број *BGP update* порука генерисаних од *EGP* протокола у току једне минуте
- *Incomplete* - број *BGP update* порука генерисаних од непознатог извора у току једне минуте

Дупле *BGP* поруке типа ‘објављена’/‘повучена’ дефинишу се на сљедећи начин: ако у тренутку t0 имамо објављену/повучену *IP* адресу са одређеним *ASPATH* атрибутом и ако у тренутку t1 имамо објављену/повучену *IP* адресу са истим *ASPATH* атрибутом, онда се поруке сматрају дуплим. Имплицитно повучене *BGP* поруке се дефинишу овако: ако у тренутку t0 имамо објављену *IP* адресу са једним *ASPATH* атрибутом, а у тренутку t1 објављена иста *IP* адреса са другим *ASPATH* атрибутом, онда се порука сматра имплицитно повученом.

Наредна својства су генерисана из *ASPATH* атрибута *BGP update* поруке. Парсирање *ASCII* фајла и генерисање својстава је изведено путем филтрирања података у бази. Додатни прикази се креирају у бази са циљем постизања различитих захтјева. Комплекснији задаци захтјевају писање *PL/SQL* кода. Сљедећа својства су генерисана на основу података о *ASPATH* атрибуту из *BGP update* поруке:

- Просјечна дужина атрибута *ASPATH* – просјечна дужина *ASPATH*-ова свих *BGP update* порука у току једне минуте
- Максимална дужина атрибута *ASPATH* - максимална дужина *ASPATH*-ова свих *BGP update* порука у току једне минуте
- Просјечна дужина јединствених атрибута *ASPATH* – просјечна дужина јединствених *ASPATH*-ова свих *BGP update* порука у току једне минуте
- Просјечна едит дистанце – просјечна едит дистанце између свих порука у току једне минуте
- Максимална едит дистанце – максимална едит дистанце између свих порука у току једне минуте

Levenshtein-ова удаљеност је мјера сличности између два низа. Удаљеност је минимални број неопходних избацивања, убацивања или замјена да би се један низ трансформисао у други. Функција креирана за рачунање *Levenshtein*-ове удаљености [19] је унесена у базу података. *PL/SQL* код је написан за кориштење функције [19] и да би *ASPATH* атрибути користили као улази у *Levenshtein*-ову функцију. *SQL* упити рачунају просјечну и максималну вриједност едит дистанце у току једне минуте.

B. Друга Фаза

Машинско учење се бави аутоматским препознавањем образаца у подацима и њихове што тачније предикције базиране на претходним искуствима. Класификација је процес у коме се улазни подаци разврставају у неколико предефинисаних класа и користи се у разним сферама људског дјеловања. Управљање и анализа података представља изазов данашњице како у истраживачком тако и у приватном сектору. Проблем дименизионалности података укључених у процес машинског учења је

проблем у коме се праве помаци и нуде рјешења. Циљ је смањивање количине података који се обрађују и креирање модела који могу доносити квалитетне закључке на редукованом скупу података. Технике редукције димезионалности су подобласт машинског учења. Њима се врши селекција података у сврху поједностављивања обраде истих. За посљедицу имају моделе који боље генерализују податке, а самим тим смањују комплексност прорачуна и ресурсе неопходне за прорачуне. Једна од техника редукције димензионалности којом се број својстава смањује јесте селекција својстава [20]. Са присуством великог броја својстава, дакле без селекције својстава, учећи модел се претјерано прилагођава подацима (*overfitting*), што има за посљедицу деградацију перформанси. Селекција својстава у машинском учењу је процес који се односи на селекцију релевантних својстава, тј. селекцију подскупа релевантних својстава. Процес селекције својстава је неопходан корак при конструкцији класификацијског модела. Гледано из перспективе процеса анализе података, показује која су улазна својства битна за предикцију и у каквом су међусобном односу. С обзиром да је фокус на подацима који су најкориснији, перформансе учећих алгоритама и будућих предвиђања се побољшавају.

У [21] аутори представљају четири корака која чине генералну структуру селекције својстава. У првом кораку се генерише подскуп својстава, који може почети са празним, пуним или подскупом са својствима насумично одабраним. Сљедећи корак је процес евалуације у којој се оцјењује одабрани подскуп својстава и пореди са најбољим претходним подскупом. Увијек се најбоље оцијењени подскуп задржава за будућа поређења. Неопходност постојања критерија заустављања, као трећег корака, огледа се у томе да се процес селекције својстава мора ограничити на неки начин. Критериј заустављања може се базирати на процесу генерисања подскупа својстава, путем унапријед дефинисаног броја могућих својстава која ће се одабрати, или унапријед дефинисаног броја итерација које ће се извршавати. Такође, критериј заустављања може се базирати на процесу евалуације, у случају када се не постигне бољи подскуп (приликом елиминације или додавања својстава), или када је остварен оптимални подскуп у складу са критеријем процјене. Посљедњи корак се односи на процес валидације.

У току процеса селекције својстава се полази од тога да сувишна (енг. *redundant*) или безначајна (енг. *irrelevant*) својства треба одстранити из матрице података, и то на начин који неће проузроковати губитак информација. Такође је пожељно из матрице података одстранити елементе шума.

Селекција својстава може се такође подијелити на филтер методе и методе претходног учења (енг. *wrapper*) [22]. Филтер методе раздвајају процес селекције својстава и процес алгоритма учења, тако да апроксимиране грешке (енг. *bias*) учећег алгоритма и алгоритма селекције својстава не утичу једна на другу. У случају методе претходног учења, прво се одреди најбољи подскуп својстава, затим се за сваки подскуп уче правила, а потом одабире подскуп изабраних својстава према процјени тачности предвиђања алгоритма учења. Два најчешће кориштена метода машинског учења су индуктивно учење или надгледано учење (енг. *supervised learning*) и самообучавање или ненадгледано учење (енг. *unsupervised learning*) [23]. Разлика између два поменута метода машинског учења је у сљедећем: код надгледаног учења циљни атрибут је познат, а код ненадгледаног учења непознат. У случају надгледаног учења подаци за обучавање модела морају садржавати и циљни атрибут. Дакле, у надгледаном учењу, алгоритми машинског учења на основу података реализују класификатор (ако су циљни атрибути дискретне вриједности); или регресијску функцију (ако су циљни атрибути непрекидне вриједности).

Основни појмови надгледаног проблема машинског учења су:

- улазни вектор својстава $x = (x_1, x_2, ..., x_m)$, где је *m* број улазних својстава;
- излазно (циљно) својство у;
- примјер за обучавање (тестирање) $(x_1, x_2, ..., x_m, y)$;
- скуп N примјера за обучавање $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$;
- непозната циљна функција *f(x)*.

За дати скуп примјера за учење $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$ где су x_i улазни вектор својстава, y_i излазно (циљно) својство за *i*-ти примјер, и N број примјера за обучавање, треба научити функцију $h: X \rightarrow Y$, тако да $h(x) \approx f(x)$ тј. да $h(x)$ буде добар предиктор за одговарајућу вриједност циљног својства у. Циљно својство дефинише понашање система који учи. Функција h је један елемент простора хипотезâ H која је сачињена од скупа свих прихватљивих хипотеза и њом апроксимирамо циљну функцију.

C. Трећа Фаза

Бинарна класификација има за циљ да идентификује једну класу. У случају детекције аномалија то је класа ‘аномалија’. Подаци који нису кориштени за учење модела класификатора у току класификације једнозначно се придружују скуповима података ‘аномалија’ или ‘регуларна’. Одлука коју доноси класификатор представља се матрицом конфузије гдје је:

- *TP* (енг. *True Positive*): број инстанци тренинг скупа ‘аномалија’ које су класификоване као аномалија
- *FP* (енг. *False Positive*): број инстанци тренинг скупа ‘регуларна’ које су класификоване као аномалија
- *FN* (енг. *False Negative*): број инстанци тренинг скупа ‘аномалија’ које су класификоване као регуларна
- *TN* (енг. *True Negative*): број инстанци тренинг скупа ‘регуларна’ које су класификоване као регуларна.

Класификатори су обучени на ограниченом скупу података и тестирани на подацима који нису кориштени при обучавању. Евалуација перформанси класификатора обавља се како би се провјерила његова способност да генерализује. Циљ је класификовати класу ‘аномалија’,

која се у скуповима података за обучавање и тестирање налази у мањем постотку. Ако постоји несразмјерност у подацима требају се идентификовати мјере евалуације класификатора, које ће рефлектовати тачност и прецизност класификатора над подацима класе ‘аномалија’. Треба имати на уму да тачност као мјера евалуације претпоставља једнаку цијену за погрешне класификације, без обзира на класу из које инстанце долазе.

Постоји низ мјера евалуације које се могу израчунати: Тачност, Одзив, Прецизност, *FPR* (енг. *False Positive Rate*), *F*-мјера, *MCC*. *MCC* се користи када класе имају различит број инстанци, тј. када подаци у класама нису избалансирани. Треба примијетити да ни одзив, а ни прецизност не узимају у обзир исправно класификоване инстанце класе ‘регуларна’ (*TN*). Ове двије мјере обично дјелују тако да се са повећањем једне смањује друга и обрнуто. На примјер, строжији класификатор ће имати повећану вриједност прецизности, а смањену вриједност одзива. Одзив и прецизност имају максималну вриједност када су погрешно идентификоване инстанце класе ‘регуларна’ (*FN*) и класе ‘аномалија’ (*FP*) минималне, респективно. Поред набројаних мјера за евалуацију модела машинског учења користе се *ROC* (енг. *Receiver Operating Characteristics*) криве које представљају TPR (одзив) у функцији од *FPR* за различите параметре модела машинског учења и *PR* (енг. *Precision-Recall*) криве које представљају прецизност у функцији од одзива за различите параметре модела машинског учења.

III. Закључак

Постојање методолошког приступа је неопходно у домену детекције аномалија BGP-а. У раду је приказан дијаграм тока детекције аномалија у BGP-у и детаљно су разрађене све фазе процеса. Одабир података, обрада података и екстракције својстава је први корак процеса. Затим се анализирају прикупљени и обрађени подаци примјеном метода селекције својстава и метода машинског учења. Посљедња фаза је интерпретација добијених резултата у складу са мјерама евалуације модела учења.

Литература

[1] Y. Rekhter, T. Li, S. Hares, "A Border Gateway Protocol 4 (*BGP*-4)," RFC 4271, IETF, 2006 [Online]. Available: http://ietf.org/rfc/rfc4271 [last visited 17.06.2017]

[2] J. Hawiknson, T. Bates, "Guidelines for creation, selection, and registration of an Autonomous System (AS), "RFC 1930, IETF, 1996 [Online]. Available: http://tools.ietf.org/html/rfc1930 [last visited 17.06.2017]

[3] The Internet Assigned Numbers Authority (IANA) [Online]. Available: http://www.iana.org/ [last visited 17.06.2017]

[4] T. Bates, P. Smith, G. Huston, "CIDR report," 2016 [Online]. Available: http://www.cidr-report.org/as2.0/ [last visited 17.06.2017]Y. Rekhter, et al. “A Border Gateway Protocol 4 (*BGP*-4),” RFC 4271, IETF, 2006 [Online]. Available: http://ietf.org/rfc/rfc4271

[5] K. Butler, T. R. Farley, P. McDaniel, "A survey of *BGP* security issues and solutions,"Technical report, AT&T Labs - Research, Florham Park, NJ, February 2004

[6] M. Ćosović, S. Obradović, "Sigurnost u *BGP* Protokolu," INFOTEH-JAHORINA, Vol. 13, Ref. KST 3 6, pp. 496 500, March 2014.

[7] S. Kent, C. Lynn, K. Seo, "Secure Border Gateway Protocol (S*BGP*)," IEEE Journal on Selected Areas in Communications, Vol. 18, No. 4, pp. 582 592, April 2000.Y. Rekhter, T. Li, “A Border Gateway Protocol 4 (*BGP*-4),” RFC 1771, IETF, 1995 [Online]. Available: http://ietf.org/rfc/rfc1771

[8] The Internet Engineering Task Force [Online]. Available: https://www.ietf.org/ [last visited 17.06.2017]

[9] The Inter-Domain Routing Working Group [Online]. Available: https://datatracker.ietf.org/wg/idr/about/ [last visited 17.06.2017]

[10] The Global Routing Operations Working Group [Online]. Available: https://datatracker.ietf.org/wg/grow/about/ [last visited 17.06.2017]

[11] The North America Network Operators’ Group [Online]. Available: https://www.nanog.org/ [last visited 17.06.2017]

[12] Mailing list Archive of The North America Network Operators’ Group [Online]. Available: https://www.nanog.org/list/archives/historical [last visited 17.06.2017]

[13] RIPE RIS raw data [Online]. Available: http://www.ripe.net/data-tools/stats/ris/ris-raw-data [last visited 17.06.2017]

[14] University of Oregon Route Views project [Online]. Available: http://www.routeviews.org/ [last visited 17.06.2017]

[15] T. Manderson, "Multi-threaded routing toolkit (MRT) Border Gateway Protocol (BGP) routing information export format with geo-location extensions," RFC 6397, IETF, [Online]. Available: http://www.ietf.org/rfc/rfc6397.txt [last visited 17.06.2017]

[16] The Internet Exchange Map [Online]. Available: http://internetexchangemap.com/ [last visited 17.06.2017]

[17] L. Blunk, M. Karir, C. Labovitz, "Multi-Threaded Routing Toolkit (MRT) Routing Information Export Format," October 2011 [Online]. Available: https://tools.ietf.org/html/rfc6396 [last visited 17.06.2017]

[18] P. -C. Cheng, X. Zhao, B. Zhang, L. Zhang, "Longitudinal study of BGP monitor session failures," Computer Communication Review, Vol. 40, No. 2, pp. 34 42, April 2010.

[19]]Levenshtein algorithm PL/SQL [Online]. Available: http://richmurnane.blogspot.com/2006/02/ levenshtein-distance-algorithm-oracle.html [last visited 17.06.2017]

[20] M. Cosovic, S. Obradovic, and Lj. Trajkovic, "Feature selection techniques for machine learning," in Proceedings of International Scientific Conference, UNITECH 2013, Gabrovo, Bulgaria, November 2013, No. 1, pp. 85-89

[21] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, Vol. 1, No. 3, pp. 131 156, May 1997.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, Vol. 97, No. 1 2, pp. 273 324, December 1997.

[23] M. Cosovic, S. Obradovic, and Lj. Trajkovic, "Algorithms for investigation of abnormal BGP events," in Proceedings of International Scientific Conference, UNITECH 2013, Gabrovo, Bulgaria, November 2013, No. 2, pp. 253 257

ABSTRACT

The paper describes BGP anomaly detection methodology. Detection of anomalies is present in different domains. The importance of methodological approach is of great importance in the domain of anomaly detection in BGP for both researchers and network operators. The flow diagram of the anomaly detection in BGP is displayed and three phases of the process are elaborated in detail. The first phase consists of selecting data, processing data and extracting properties from BGP update messages. The analysis of the collected and processed data from the first phase follows using the feature selection methods and machine learning models. At the last stage, the obtained results are interpreted in accordance with the learning model evaluation measures.

ANOMALY DETECTION METHODOLOGY IN *BGP*

Marijana Ćosović, Slobodan Obradović