Detekcia Anomálií v Transakčných Údajoch Pomocou Autoenkóderov

Martin Mocko, Jakub Ševcech

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

martin.mocko1@gmail.com, jakub.sevcech@stuba.sk

Abstrakt. Základnou myšlienkou väčšiny algoritmov na detekciu anomálií je nájsť pozorovania, ktoré sa vymykajú typickým vzorom. Často sa na takéto účely používa zhlukovanie, hľadanie frekventovaných vzorov alebo rôzne prístupy založené na podobnosti pozorovaní. V tejto práci rozvíjame myšlienku detekcie anomálií na základe rekonštrukčnej chyby vznikajúcej pri rekonštrukcii zo stratovej kompresie vytvorenej autoenkóderom. Vyhodnocujeme úspešnosť detekcie na základe tejto rekonštrukčnej chyby a porovnávame úspešnosť takejto detekcie s rôznymi klasifikátormi na sade údajov reprezentujúcej rôzne útoky na počítačovú sieť.

Kľúčové slová: detekcia anomálií, autoenkóder, transakčné údaje

1 Úvod

Detekcia anomálií je dôležitou podoblasťou strojového učenia. Anomálie môžeme zadefinovať ako vzory v dátach, ktoré nezodpovedajú definovanému normálnemu správaniu [1]. Pri úlohe detekcie anomálií je pri väčšine existujúcich prístupov nevyhnutné mať vytvorený nejaký model, ktorý opisuje, ako vyzerá normálne správanie. Na základe takéhoto modelu je následne možné detegovať pozorovania, ktoré sa z neho vymykajú a priraďovať im skóre "anomálnosti".

Rôzne algoritmy strojového učenia typicky rozdeľujeme do troch tried: učenie s učiteľom, bez učiteľa a učenie s čiastočnou podporou učiteľa (v angličtine nazývané semi-supervised learning). Pri detekcii anomálií je možné použiť metódy zo všetkých troch týchto tried [1]. Medzi populárne techniky, ktoré sa používajú na riešenie problému detekcie anomálií sú napríklad techniky založené na klasifikácii známych tried anomálií, kombinácia zhlukovania (K-means, DBSCAN, ...) a pozorovania vzdialenosti od najbližšieho zhluku, techniky zohľadňujúce vzdialenosť najbližších susedov (kNN s použitím metriky Local Outlier Factor [3]), techniky založené na odchýlkach od asociačných pravidiel a frekventovaných množín ale aj techniky založené na stratovej kompresii a sledovaní rekonštrukčnej chyby vzniknutej pri rekonštrukcii pozorovania z jeho komprimovanej podoby [4]. Práve na metódu z tejto poslednej skupiny sa sústreďujeme v tejto práci.

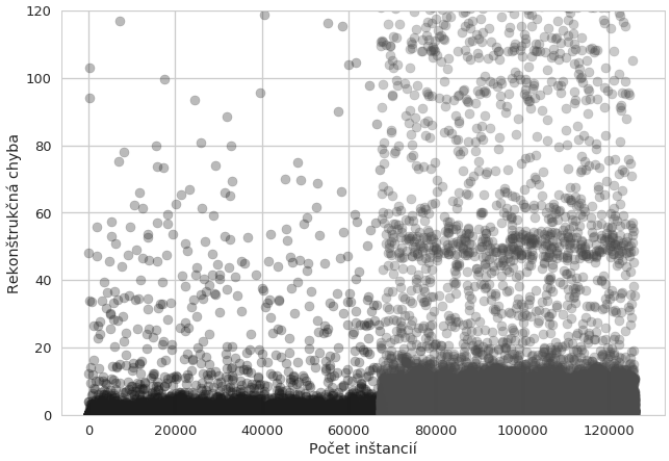
Zamerali sme sa na použitie autoenkóderov na naučenie sa štruktúry spracovávaných údajov a na ich použitie pre označenie pozorovaní pomocou skóre anomálnosti. Hlavnou myšlienkou autoenkóderov je natrénovať neurónovú sieť tak, aby dokázala zrekonštruovať vstupné údaje na výstupnej vrstve aj napriek obmedzenej veľkosti skrytých vrstiev (skrytá vrstva je menšia ako vstupná alebo výstupná) [2]. Snaha je teda o čo najväčšiu minimalizáciu rekonštrukčnej chyby autoenkódera.
Následné použitie autoenkódera na detekciu anomálií spočíva vo využití predpokladu, že normálne správanie, ktorého by v trénovacích dátach mala byť veľká väčšina, dokáže autoenkóder zrekonštruovať s menšou rekonštrukčnou chybou ako anomálne pozorovania.

Detekcia anomálií sa používa v mnohých doménach, ako prostriedok na detekciu narušenia, monitorovanie zdravia systému, detekcia netypických udalostí v sieťach senzorov a v neposlednom rade na detekciu podvodov. V našej práci sme si zvolili doménu monitorovania počítačovej siete ako príklad oblasti, kde hľadáme anomálie v agregovaných údajoch o udalostiach generovaných nejakým systémom alebo skupinou agentov. V tomto prípade ide o charakteristiky spojení medzi počítačmi pripojenými do siete.

2 Metóda detekcie anomálií pomocou autoenkóderu

Hlavnou myšlienkou použitia autoenkóderov (ako aj ďalších metód pre stratovú kompresiu ako je napríklad PCA) je použiť rekonštrukčnú chybu jednotlivých pozorovaní ako anomálne skóre. Základný predpoklad je, že anomálie sú zriedkavé, generuje ich iný proces ako normálne údaje a odlišujú sa od normálnych pozorovaní. Kompresný algoritmus preto nedokáže vytvoriť spoľahlivý model na ich kompresiu a teda ich rekonštrukčná chyba bude väčšia ako pri typických údajoch.
Toto je vidieť na obrázku číslo 1, ktorý zobrazuje rekonštrukčné chyby pozorovaní z dátovej sady normálnych a útočných spojení v počítačovej sieti. Na ľavej strane sú zobrazené bežné spojenia a na pravej tie, ktoré vznikali počas rôznych typov útokov na sieť. Pozícia bodu na y-ovej osy zobrazuje rekonštrukčný chyby po spracovaní natrénovaným autoenkóderom. Odtieňom šedej farby sú v ľavej polovici obrázku zobrazené normálne pozorovania a v pravej polovici anomálne pozorovania. Jasne je vidieť rozdiel v rekonštrukčnej chybe pre tieto dve skupiny pozorovaní a jasne je vidieť hranicu, kde sa tieto dve množiny stretajú. Pre normálne pozorovania nadobúda rekonštrukčná chyba oveľa menšie hodnoty ako pre tie útočné a teda by bolo možné ich na základe prahovej hodnoty rekonštrukčnej chyby rozdeliť do týchto dvoch tried.

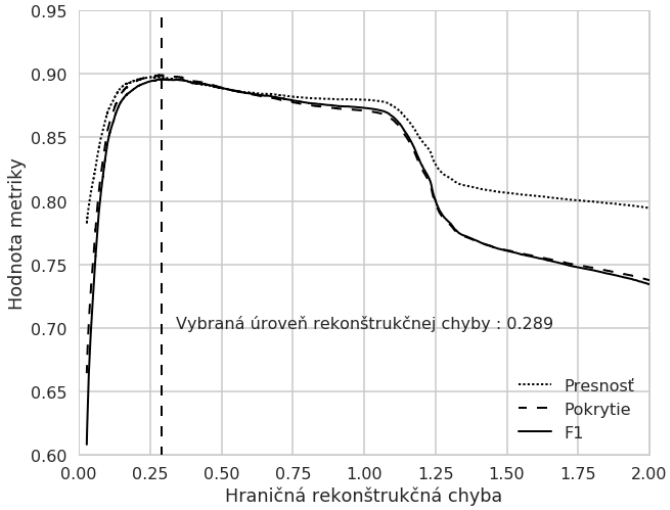
Vytvorené anomálne skóre (rekonštrukčnú chybu) je možné použiť priamo na označenie pozorovaní za anomálne na základe zvolenej prahovej hodnoty, ale je možné tiež použiť toto skóre ako vstupný atribút do ďalších fáz spracovania údajov. V tomto príspevku sa sústreďujeme na priame použitie anomálneho skóre samotného na detekciu anomálií a preto prezentujeme použitie prahovej hodnoty tohto skóre na označenie anomálnych pozorovaní.



Obr. 1. Rekonštrukčná chyba na testovacej sade pre typické a anomálne pozorovania v dátovej sade útokov na počítačovú sieť (KDDTest+)

Postup výpočtu anomálneho skóre a označenia pozorovaní za anomálne je nasledovný:

1. Natrénovanie autoenkóderu na trénovacej vzorke, kde minimalizujeme rekonštrukčnú chybu.
2. Výpočet rekonštrukčnej chyby pre celú trénovaciu sadu.
3. Výpočet úspešnosti označovania anomálií pre rôzne prahové hodnoty rekonštrukčnej chyby a výber optimálnej prahovej hodnoty (Obrázok 2) na základe zvolenej metriky (v našich experimentoch sme zvolili metriku F1 makro).



Obr. 2. Výber prahovej rekonštrukčnej chyby pre označenie anomálnych pozorovaní

Takto natrénovaný autoenkóder a zvolená prahová hodnota rekonštrukčnej chyby je použitá na spracovanie a predikciu na testovacej sade.

3 Návrh experimentu a výsledky

Na vyhodnotenie úspešnosti autoenkóderu na detekciu anomálií sme použili dataset NSL-KDD [5], ktorý obsahuje údaje o normálnych (neškodných) spojeniach v počítačovej sieti ako aj údaje vytvorené pri viacerých rôznych typoch útokov na sieť. Pozorovania sú označené týmito typmi útokov. V našom experimente sme porovnávali úspešnosť autoenkóderu a viacerých klasifikačných algoritmov pri detekcii týchto útokov.

Keďže pôvodný dataset obsahuje normálne a škodlivé pozorovania vo vyrovnanom pomere a my pri trénovaní modelu detekcie anomálií využívame predpoklad, že anomálie sú zriedkavé, rozhodli sme sa upraviť pomer normálnych a škodlivých pozorovaní v trénovacej sade na 95% a 5%. Použili sme pri tom 100% normálnych pozorovaní a doplnili ich náhodnou podmnožinou anomálnych pozorovaní (náhodný výber bez opakovania spomedzi anomálnych pozorovaní v trénovacej sade) tak, aby tieto vo výslednej dátovej sade tvorili 5% všetkých pozorovaní. Výsledná trénovacia sada obsahovala viac ako 70 000 pozorovaní.

Pri overovaní sme používali obe testovacie sady NSL-KDD datasetu, nazvané KDDTest+ a KDDTest-21. Testovaciu dátovú sadu KDDTest-21 vytvorili autori NSL-KDD datasetu odstránením triviálnych pozorovaní zo sady KDDTest+. Natrénovali 21 rôznych klasifikátorov na detekciu útokov a odstránili tie pozorovania, ktoré boli úspešne detegované všetkými týmito klasifikátormi. Vznikla tak jedna úplná sada a jedna, ktorá obsahovala len tie „ťažšie“ príklady.

Na nastavenie hyper-parametrov autoenkóderu sme používali vyčerpávajúce prehľadávanie mriežky nastavení, pričom sme používali nastavenia zosumarizované v tabuľke číslo 1. V práci prezentujeme výsledok, ktorý dosiahol najvyššiu hodnotu metriky F1 spomedzi všetkých overovaných nastavení.

Tabuľka 1. Nastavenia hyper-parametrov autoenkóderu

Hyper-parameter	**Nastavenie**
Aktivačná funkcia	{tanh, sigmoid, linear}
Počet skrytých vrstiev	{1,2,3}
Veľkosť skrytých vrstiev	{10, 15, 20, 25, 30}, veľkosti vrstiev boli symetrické podľa strednej vrstvy, pričom stredná vrstva bola vždy menšia ako vonkajšie napr. 30, 10, 30.

Úspešnosť autoenkóderu pri detekcií útokov na počítačovú sieť sme porovnávali s viacerými bežne používanými klasifikátormi, pre ktoré sme našli najlepšie parametre pomocou vyčerpávajúceho prehľadávania mriežky nastavení. Výsledné nastavenia rôznych klasifikačných algoritmov, ktoré sme použili pri porovnaní, sú zosumarizované v tabuľke 2.

Pre porovnávané klasifikátory sme využili ich implementácie v jazyku Python v knižnici scikit-learn a pre autoenkóder knižnicu Keras v kombinácii s TensorFlow. Dosiahnuté výsledky sú zhrnuté v tabuľke číslo 3. Pre všetky prezentované metriky v tabuľke 3 prezentujeme ich makro verziu aby sme dosiahli robustnosť výsledkov na nevyváženej dátovej sade.

Tabuľka 2. Výsledné nastavenia porovnávaných klasifikátorov získané vyčerpávajúcim prehľadávaním priestoru hyperparametrov

Algoritmus	Nastavenie parametrov
Autoenkóder	activation_function:tanh, layers:(25,15,25)
Logistická regresia	max_iter: 500, penalty: l2, solver: newton-cg
SVM (rbf)	C: 1.0, gamma: auto
kNN	n_neighbors: 3, weights: distance
Rozhodovací strom	criterion: entropy, max_features: None, max_depth: 10
Náhodný les	n_estimators: 150, criterion: entropy, max_features: None, max_depth: 10
MLP	hidden_layer_sizes: (100,), alpha: 0.1, activation: tanh
SVM (linear)	penalty: l2, C: 0.1
Naive Bayes	-

Tabuľka 3. Výsledky detekcie útokov pre dve rôzne testovacie sady

	KDDTest+			KDDTest-21		
Algoritmus	F1	Presnosť	Pokrytie	F1	Presnosť	Pokrytie
Autoenkóder	0.76	0.79	0.78	**0.52**	0.57	0.61
Logistická regresia	0.65	0.77	0.70	0.37	0.59	0.60
SVM (rbf)	0.71	0.79	0.75	0.45	**0.60**	0.65
kNN	0.71	0.79	0.75	0.46	**0.60**	0.65
Rozhodovací strom	0.74	0.78	0.77	0.48	0.56	0.60
Náhodný les	0.74	**0.80**	0.77	0.49	**0.60**	**0.66**
MLP	0.72	0.79	0.76	0.47	0.59	0.64
SVM (linear)	0.70	0.75	0.73	0.42	0.52	0.54
Naive Bayes	**0.77**	0.79	**0.79**	**0.52**	0.55	0.59

Ako je z tabuľky 3 vidieť, pomocou autoenkóderu sa nám podarilo dosiahnuť najlepší výsledok v porovnaní s ostatnými klasifikátormi len na jednej vyhodnocovanej metrike a len v prípade jednej porovnávanej dátovej sady. Takmer vždy však boli výsledky autoenkóderu medzi najúspešnejšími metódami. Tento výsledok sme dosiahli napriek tomu, že sme porovnávali metódu učenia bez učiteľa (autoenkóder) a metódami, ktoré využívali informáciu o skutočnej triede trénovacích údajov na vytvorenie modelu.

4 Záver

V prezentovanej práci sme vyhodnotili použiteľnosť autoenkóderu ako prostriedku na detekciu anomálií (útokov) v údajoch opisujúcich rôzne typy útokov na počítačovú sieť. Podarilo sa nám dosiahnuť porovnateľné výsledky pri použití autoenkóderu na výpočet rekonštrukčnej chyby ako anomálneho skóre, ako v prípade použitia rôznych klasifikátorov na identifikáciu útokov. Porovnateľné výsledky sme dosiahli napriek tomu, že sme porovnávali úspešnosť algoritmu učenia bez učiteľa s výsledkami algoritmov, ktoré dokázali použiť informáciu o skutočnej triede pozorovaní pri vytváraní modelu. Overili sme teda predpoklad, že rekonštrukčná chyba získaná takýmto spôsobom je dobrý indikátor anomálnosti pozorovania. V ďalšej práci overíme použiteľnosť tohto anomálneho skóre ako vstupnej premennej do ďalšieho spracovania pomocou bežne používaných klasifikátorov. Predpokladáme, že týmto spôsobom by sme mohli ešte viac zvýšiť presnosť detekcie anomálií a znížiť množstvo nesprávne označených normálnych pozorovaní za anomálne.

Poďakovanie. Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV-15-0508 a VG 1/0646/15.

Literatúra

1. Agrawal S, Agrawal J (2015) Survey on anomaly detection using data mining techniques. In: Procedia Computer Science, vol 60, pp 708-713
2. Bengio Y (2009) Learning deep architectures for AI. In: Foundations and trends in Machine Le-arning, vol 2, no 1, pp 1-127
3. Dokas P, Ertoz L, Kumar V, Lazarevic A, Srivastava J, Tan P N (November 2002) Data mining for network intrusion detection. In: Proceedings of NSF Workshop on Next Genera-tion Data Mining, pp 21-30
4. Pevny T, Rehák M, Grill M (December 2012) Detecting anomalous network hosts by means of pca. In: Proceedings of Information Forensics and Security (WIFS), IEEE, pp 103-108
5. Tavallaee M, Bagheri E, Lu W, Ghorbani A A (July 2009) A detailed analysis of the KDD CUP 99 data set. In: Computational Intelligence for Security and Defense Applications, IEEE, pp 1-6