

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-0000-73688

Bc. Matúš Cuper

Identifikácia neštandardného správania odberateľov v energetickej sieti

Diplomová práca

Vedúci práce: Ing. Marek Lóderer

máj 2019

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-0000-73688

Bc. Matúš Cuper

Identifikácia neštandardného správania odberateľov v energetickej sieti

Diplomová práca

Študijný program: Inteligentné softvérové systémy

Študijný odbor: 9.2.5 Softvérové inžinierstvo

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave

Vedúci práce: Ing. Marek Lóderer

máj 2019

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Študijný program: Inteligentné softvérové systémy

Autor: Bc. Matúš Cuper

Bakalárska práca: Identifikácia neštandardného správania odberateľov v energetickej sieti

Vedúci práce: Ing. Marek Lóderer

máj 2019

V práci sme sa zamerali

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree Course: Intelligent software systems

Author: Bc. Matúš Cuper

Bachelor thesis: Identification of abnormal behavior of customers in the power grid

Supervisor: Ing. Marek Lóderer

May 2019

In the thesis we focused

ČESTNÉ PREHLÁSENIE

Čestne prehlasujem, že diplomovú prácu som vypracoval samostatne pod vedením vedúceho diplomovej práce a s použitím odbornej literatúry, ktorá je uvedená v zozname použitej literatúry.

.....
Matúš Cuper

POĎAKOVANIE

Ďakujem vedúcemu diplomovej práce Ing. Marekovi Lódererovi za odborné vedenie, cenné rady a pripomienky pri spracovaní diplomovej práce.

Obsah

1	Úvod	1
2	Analýza problému	2
2.1	Anomálie v energetickej sieti	2
2.2	Typy anomálií	3
2.2.1	Bodové anomálie	3
2.2.2	Kontextové anomálie	3
2.2.3	Skupinové anomálie	4
2.3	Prístupy k identifikácii anomálií	4
2.3.1	Detekcia bez učiteľa	4
2.3.2	Detekcia s učiteľom	4
2.4	Techniky detekcie anomálií	5
2.4.1	Klasifikácia	5
2.4.2	Analýza najbližšieho suseda	5
2.4.3	Klastrovanie	5
3	Rozpracovanie problému	7

1 Úvod

Jedným z problémov, ktorým v súčasnosti čelia distribučné spoločnosti, je detekcia neštandardného správania odberateľov. Jej úlohou je identifikovať profily zákazníkov, ktorí svojím správaním porušujú stanovené podmienky a manipulujú s hodnotami nameranými meračmi za cieľom obohatenia sa. Samozrejme tiež dochádza k prípadom, kedy je presnosť meracieho zariadenia nižšia aj bez zapríčinenia zákazníka. Oba prípady sú pre distribučnú spoločnosť nežiadúce a je v záujme zníženia strát ich, čo najskôr identifikovať. Obvykle sú za týmto účelom vykonávané náhodné kontroly, ktoré pokrývajú iba nízky počet zákazníkov s anomálnym správaním. Na základe množstva dát získavaných z inteligentných meračov je možné modelovať správanie zákazníkov. Distribučné spoločnosti tak môžu znižovať svoje straty a preverovať iba odberateľov, ktorí svojím profilom nezapadajú medzi odberateľov so štandardným správaním.

2 Analýza problému

Tak ako je spomenuté v článku [4], straty v distribučných sieťach v niektorých krajinách tvoria až 30% z celkového objemu distribuovanej energie. Väčšinu strát vytvára svojimi vlastnosťami samotná sieť, no nezanedbateľnú časť tvoria aj nelegálne odbery. Pravidelná kontrola všetkých odberateľov by bola časovo aj finančne náročná, preto je potrebné správne identifikovať zákazníkov s neštandardnou spotrebou energie, čím sa minimalizujú náklady spojené s kontrolami. Zatiaľ čo v minulosti bola možná identifikácia nelegálnych odberov len fyzickou kontrolou, dnes vieme obmedziť okruh podozrivých aj na diaľku, keďže inteligentné merače nám poskytujú dáta v pravidelných intervaloch s minimálnou odchýlkou.

Vďaka tomu vznikajú nové možnosti identifikácie neštandardného správania využitím dátovej analitiky a strojového učenia. Zatiaľ čo väčšina algoritmov na identifikáciu anomálií pracuje s nízkorozmernými dátami, časové rady predstavujú presný opak a použité metódy sa líšia od tých klasických. Výzvou pri skupinových a kontextových anomáliách je aj vhodný výber premenných, na základe ktorých budú anomálie identifikované. Zvýšenie presnosti pri hľadaní anomálií môžeme docieľiť kombinovaním rôznych zdrojov dát, či už by sa jednalo o počasie alebo údaje z inteligentných meračov iných druhov energie. Cieľom tejto kapitoly je preto analyzovať a porovnať používané metódy pri detekcii anomálií v časových radoch a zamerať sa najmä na vhodnú reprezentáciu jednotlivých odberovateľov pomocou získaných dát.

Anomálne správanie alebo anomália je definovaná ako vzor v správaní, ktorý nezodpovedá štandardnému správaniu. Pri dátach z inteligentných meračov, anomália zodpovedá meraniu, ktoré sa nenachádza v oblasti normálnych dát.

Na detekciu anomálií sú používané aj algoritmy určené na klasifikáciu, ako je napríklad naivný Bayesovský klasifikátor (angl. Naive Bayes), k-najbližší susedia (angl. k-nearest neighbors), rozhodovacie stromy (angl. decision tree), náhodné lesy (angl. random forests), neurónové siete so spätnou propagáciou (angl. neural networks with backpropagation) alebo metóda podporných vektorov (angl. support vector machine) [2].

Pri identifikácii anomálií je najskôr potrebné zamyslieť sa nad nasledovnými problémami:

- **Definovanie oblasti normálnych dát** je veľmi náročné, nakoľko hranica medzi normálnymi dátami a anomáliami je nepresná a môže tak dojsť k nesprávnemu označeniu meraní,
- **Anomálie vytvorené škodlivou činnosťou** sa javia ako normálne dáta, čo sťažuje definíciu normálneho správania,
- **Evolúcia dát** spôsobuje, že definícia normálneho správania sa môže časom zmeniť,
- **Presná predstava o anomálii** je často rôzna naprieč viacerými odbormi, a preto neexistuje univerzálny spôsob na určovanie anomálií,
- **Dostupnosť označených dát** zlepšuje presnosť identifikácie anomálií, avšak často takéto dáta neexistujú alebo ich je potrebné označiť,
- **Biely šum** vyskytujúci sa v dátach má tendenciu skresľovať normálne dáta, ktorých identifikácia je následne zložitá [1].

2.1 Anomálie v energetickej sieti

V distribučných sieťach vznikajú straty, ktoré vo všeobecnosti môžeme rozdeliť na technické a netechnické straty. Technické straty sú spôsobené vlastnosťami obvodu ako napr. odporom materiálu či únikmi cez poškodenú izoláciu a môžu sa meniť pri rôznych teplotách či počasí.

Medzi netechnické straty patria najmä nelegálne odbery. V práci sa budeme zaoberať ich identifikáciou na základe anomálneho správania spotrebiteľa. Keďže je časovo a finančne náročné pravidelne kontrolovať odberateľov tak, aby sa predišlo nelegálnemu odberu, je potrebné znížiť počet podozrivých odberateľov na minimum a zároveň maximalizovať pravdepodobnosť, s ktorou budú kontrolovaní iba odberatelia s neštandardnými odbermi [2, 6].

Najčastejšími metódami používanými pri nelegálnom odbere je obídenie meračov spotreby energie či samotná manipulácia s nimi. Merače tak poskytujú nesprávne informácie o spotrebovanej energii odberateľmi, čo je možné detegovať až po identifikácii celkových netechnických strát v sieti. Ďalšou populárnou metódou používanou na detekciu nelegálnych odberov je analýza spotrebiteľského profilu zákazníka, kedy je našou snahou identifikovať nepravidelné vzory v nameraných spotrebiteľských dátach [6]. Tak ako je spomenuté v práci [3], nelegálne odbery môžu prebiehať iba v určitom čase prípadne iba pri zvýšenej spotrebe. Identifikácia takýchto nelegálnych odberov je náročná a prípadná kontrola nemusí odhaliť manipuláciu s meracím zariadením.

Vďaka inteligentným meračom je možné detegovať nelegálne odbery omnoho rýchlejšie, najmä kvôli vysokej frekvencii zberania údajov. Tak sú identifikované aj také odbery, ktoré by sa pri klasických meraniach stratili v týždenných alebo mesačných agregáciách. Úspešnosť detekcie nelegálnych odberov je výrazne vyššia najmä pri neštandardných spotrebách alebo ak sa jedná o neopakujúcu sa udalosť. Problém vzniká ak odberateľ systematicky mení nelegálnu spotrebu a kopíruje vzory, ktoré vznikajú v dátach pri legálnom odbere. Vtedy je potrebné mať k dispozícii väčšie množstvo dát a zároveň použiť zložitejšie algoritmy detekcie anomálií, ktoré sú popísané v súvisiacej práci [5].

V súvisiacich prácach sa autori zaoberali určením netechnických strát v elektrických distribučných sieťach s použitím rôznych štatistických metód alebo strojového učenia. Dostupné dáta od distribútorov pochádzali najmä z jedného zdroja, lokality a zameriavali sa na jeden zdroj energie. Dáta, ktoré budeme mať k dispozícii disponujú podobnými vlastnosťami. V súvisiacej práci [2] boli použité viaceré zdroje dát a energie, následkom čoho bola zvýšená presnosť identifikácie anomálneho správania odberateľa. Ďalším zdrojom dát môžu byť agregované hodnoty meraní z klasických meračov, prípadne spätná väzba zo samotných kontrol odberateľov.

2.2 Typy anomálií

Dôležitým aspektom pri uplatnení detekciách anomálií je charakter anomálie. Z toho dôvodu je anomálie deliť do nasledujúcich troch skupín:

2.2.1 Bodové anomálie

Predstavujú inštancie, ktoré sa nenachádzajú v oblasti normálnych dát a je možné ich detegovať jednotlivo. Jedná sa o najjednoduchší typ anomálie a sústreďuje sa naň väčšina výskumov. Príkladom zo skutočného života môže byť detekcia podvodov s kreditnými kartami, kedy transakcia výrazne väčšieho objemu peňazí ako ostatné transakcie nachádzajúce sa v normálnom rozsahu [1].

2.2.2 Kontextové anomálie

Ide o inštancie, ktoré sa nachádzajú v oblasti normálnych dát, ale v špecifickom kontexte sú považované za anomáliu. Kontext je daný kontextovými atribútmi v dátach, na základe

ktorých sa určujú susedné inštancie. Nekontextové atribúty, nazývané aj behaviorálne, reprezentujú meranú veličinu. Napríklad pri meteorologických meraniach, budú informácie o polohe alebo nadmorskej výške predstavovať kontextové atribúty, zatiaľ čo množstvo zrážok alebo slnečných hodín budú behaviorálne atribúty.

Anomálne správanie inšancií je dané behaviorálnymi atribútmi v určitom kontexte. Čiže ak inštancia s danými behaviorálnymi atribútmi je považovaná za normálnu, iná inštancia s rovnakými behaviorálnymi ale s rôznymi kontextovými atribútmi môže byť považovaná za anomáliu. Kontextové anomálie boli najčastejšie identifikované v časových radoch. Príkladom môžu byť opäť transakcie väčšieho objemu peňazí, ktoré sú bežné v období pred Vianocami, ale neštandardné napr. na jar.

Zatiaľ čo v niektorých prípadoch je definovanie kontextu priamočiare, existujú domény, kde to nie je jednoduché. Dôležité je aby kontextové atribúty boli zmysluplne určené v cieľovej doméne ich aplikácie [1].

2.2.3 Skupinové anomálie

Predstavujú anomálie, kedy sa jednotlivé inštancie nachádzajú v oblasti normálnych dát, ale skupina týchto inšancií tvorí spolu anomáliu. Vzniknutá anomália obsahuje sekvenciu inšancií, ktorá by pri inom zoradení nepredstavovala anomáliu a taktiež jednotlivé inštancie sa nachádzajú v rozsahu normálnych dát. Príkladom môžu byť systémové volania operačného systému, ktoré sú označené ako činnosť škodlivého softvéru v prípade dodržania určitej postupnosti.

Zatiaľ čo bodové anomálie sa môžu vyskytovať v každom datasete, skupinové sa vyskytujú iba v datasetoch, kde existuje medzi inštanciami vzťah. Pri kontextových anomáliách je potrebné určiť kontextové atribúty, ktoré sa v niektorých datasetoch ani nemusia nachádzať. Problém detekcie bodových a skupinových anomálií je možné transformovať na problém detekcie kontextových anomálií, v prípade, že sa prihliada na kontext jednotlivých inšancií. Techniky používané pri detekcii skupinových anomálií sa značne líšia od techník používaných pri bodových a kontextových anomáliách [1].

2.3 Prístupy k identifikácii anomálií

V praxi sa stretávame s datasetmi, ktoré sa líšia v množstve označených dát, počte typov anomálií, ktoré budeme detegovať alebo aj pomerom medzi normálnymi inštanciami a tými neštandardnými. Často je označovanie inšancií vykonávané manuálne ľudskými expertmi drahé a neefektívne. Taktiež proces spätnej väzby môže byť zdĺhavý a nepraktický. Z toho dôvodu je dôležité zvoliť správny prístup pri identifikácii anomálií. V súčasnosti existujú 3 prístupy, a to detekcia anomálií s učiteľom, bez učiteľa a ich kombinácia [1].

2.3.1 Detekcia bez učiteľa

Táto metóda nepotrebuje označené trénovacie dáta, vďaka čomu je široko aplikovateľná a často používaná. Vychádza z predpokladu, že normálne inštancie majú majoritné zastúpenie v množine. Ak táto podmienka nie je splnená, dochádza tak často k falošnému alarmu [1].

2.3.2 Detekcia s učiteľom

Metóda potrebuje trénovacie dáta s označenými inštanciami ako normálnymi, tak aj anomálnymi. Cieľom je vytvoriť prediktívny model, ktorého úlohou je určiť triedu inšancie.

Problémom je, že anomálnych inštancií v porovnaní s normálnymi je omnoho menej a označenie dát ľudským expertom môže byť pri anomálnej inštancii náročné [1].

Kombinácia týchto dvoch prístupov počíta s označenou iba jednou triedou inštancií. Typicky sú označené normálne inštalácie, keďže ich identifikácia je menej náročná. V takom prípade je vytvorený model pre normálnu triedu a identifikácia anomálií prebieha v testovacej vzorke dát [1].

2.4 Techniky detekcie anomálií

Detegovať anomálie rôznych typov môžeme niekoľkými spôsobmi, čo závisí aj od samotných dát. Ich úplnosť, množstvo a oblasť, v ktorej boli zozbierané sú kritické pre správny výber techniky, pomocou ktorej budú identifikované anomálie. Nás budú zaujímať najmä detekcie anomálií v časových radoch. Popísané metódy sú najmä z oblasti strojového učenia a dátovej analýzy, ale pre úplnosť sú spomenuté aj iné používané metódy.

2.4.1 Klasifikácia

Pomocou naučeného modelu, nazývaného aj klasifikátor, sú rozoznávajú triedy jednotlivých inštancií. Pri detekcii anomálneho správania, klasifikátor rozlišuje iba medzi dvoma triedami, triedou normálnych dát a anomálií. Vzhľadom na to, že na natrénovanie klasifikátora sú potrebné označené dáta, ide o učenie s učiteľom. Na implementovanie klasifikátora môžeme použiť techniky založené na rôznych typoch neurónových sietí, Bayesových sieťach, pravidlových systémoch či SVM [1, 8].

2.4.2 Analýza najbližšieho suseda

Metóda určí na základe vzdialenosti alebo podobnosti medzi dátovými inštanciami, či sa jedná o normálnu inštanciu alebo anomáliu. To je vypočítané pomocou vzdialeností medzi testovanou inštanciou a všetkými bodmi, alebo iba k najbližšími bodmi. Pri viacrozmerných dátach je vzdialenosť určovaná pre každú dimenziu zvlášť. Metóda je založená na predpoklade, že zatiaľ čo normálne inštalácie sa nachádzajú pri sebe a sú husto usporiadané, anomálie sú vzdialenejšie, prípadne na okraji vzniknutých oblastí. Aplikácia je možná pomocou techník založených na relatívnej hustote alebo vzdialenosti najbližších k susedných inštancií [1, 8].

2.4.3 Klastrovanie

Jedná sa o učenie bez učiteľa, keďže klastre inštancií sú vytvorené na základe ich vzdialenosti či podobnosti. Techniky ďalej delíme do kategórií na základe predpokladu o dátových inštanciách [1, 8].

Prvá kategória používa klastrovacie algoritmy ako DBSCAN alebo ROCK a vychádza z predpokladu, že normálne inštalácie patria do klastra, zatiaľ čo anomálne nepatria do žiadneho. Keďže sa jedná o klastrovacie algoritmy, nevýhodou môže byť neoptimálne použitie pri detekcii anomálií [1].

Druhá kategória používa neurónové siete (konkrétne SOM) alebo algoritmus k-means. Vychádza z predpokladu, že normálne inštalácie ležia v blízkosti najbližšieho centroidu, anomálne inštalácie sú od neho vzdialené [1].

Posledná kategória pracuje s predpokladom, že normálne inštalácie sú súčasťou veľkých a hustých klastrov, na druhej strane anomálie patria do malých a riedkych klastrov. Používanými algoritmami sú napr. CBLOF (*angl. Cluster-Based Local Outlier Factor*) alebo k -d

stromy. V princípe algoritmy najskôr vytvárajú klastre a až potom určujú, na základe ich hustoty, či sa jedná o normálne klastre alebo anomálie. Klaster je vytvorený iba v prípade, že inštancia sa nachádza mimo preddefinovaného rádiusu od centra daného klastra [7].

3 Rozpracovanie problému

Pomocou metód strojového učenia a dátovej analitiky sa zameriame na identifikáciu anomálií v oblasti distribučných spoločností. Na základe dát, ktoré máme k dispozícii zvolíme vhodnú metódu detekcie anomálií. Keďže dáta sú z domény distribúcie elektrickej energie, ich označenie by bolo finančne aj časovo náročné. Rovnako aj spätná väzba pri identifikácii anomálií je časovo náročná a jej spracovanie môže trvať niekoľko týždňov až mesiacov.

Zameriavať sa budeme najmä na identifikáciu anomálií v časových radoch, čo spadá pod skupinové a kontextové typy anomálií. Úlohou bude taktiež identifikovať výhody a nevýhody uplatnenia jednotlivých prístupov k identifikácii. Zároveň vzniká potreba nájsť najvhodnejšie techniky detekcií anomálií pre dáta zbierané z distribučných sietí pomocou inteligentných meračov. Najmä z dôvodu, že každá doména, v ktorej je potrebné identifikovať anomálie sa vyznačuje špecifickými potrebami na použitý model.

Literatúra

- [1] Chandola, V.; Banerjee, A.; Kumar, V.: Anomaly Detection: A Survey. *ACM Comput. Surv.*, ročník 41, č. 3, jul 2009: s. 1–58, ISSN 0360-0300, doi:10.1145/1541880.1541882.
- [2] Coma-Puig, B.; Carmona, J.; Gavalda, R.; aj.: Fraud detection in energy consumption: A supervised approach. *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, 2016: s. 120–129, doi:10.1109/DSAA.2016.19.
- [3] Depuru, S. S. S. R.: *Modeling, detection, and prevention of electricity theft for enhanced performance and security of power grid*. Dizertačná práca, The University of Toledo, 2012.
- [4] Meffe, A.; de Oliveira, C. C. B.: Technical loss calculation by distribution system segment with corrections from measurements. In *CIREN 2009 - 20th International Conference and Exhibition on Electricity Distribution - Part 1*, June 2009, ISSN 0537-9989, s. 1–4, doi:10.1049/cp.2009.0962.
- [5] Nikovski, D. N.; Wang, Z.; Esenther, A.; aj.: Smart Meter Data Analysis for Power Theft Detection. *Machine Learning and Data Mining in Pattern Recognition*, 2013: s. 379–389, ISSN 03029743, doi:10.1007/978-3-642-39712-7_29.
- [6] Sahoo, S.; Nikovski, D.; Muso, T.; aj.: Electricity theft detection using smart meter data. *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2015: s. 1–5, doi:10.1109/ISGT.2015.7131776.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7131776>
- [7] Salvador, S.; Chan, P.: Learning states and rules for detecting anomalies in time series. *Applied Intelligence*, ročník 23, č. 3, 2005: s. 241–255, ISSN 0924669X, doi:10.1007/s10489-005-4610-3.
- [8] Tan, P.-N.; Steinbach, M.; Kumar, V.: *Introduction to Data Mining*. Addison Wesley, used vydanie, May 2005, ISBN 0321321367.