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

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
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

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 [6], 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 lýšia od tých klasických. Výzvou pri skupinových a kontextových anomáliach je aj vhodný výber premenných, na základe ktorých budú anomálie identifikované. Zvýšenie presnosti pri hľadaní anomálií môžeme docieliť 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.

2.1 Časové rady

2.2 Detakcia anomálií

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].

2.2.1 Typy anomálií

Dôležitým aspektom pri uplatnení detekcie anomálií je charakter anomálie. Z toho dôvodu môžeme anomálie rozdeliť do nasledujúcich troch skupín.

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 predstavuje podvod, zatiaľ čo ostatné transakcie, nachádzajúce sa v normálnom rozsahu predstavujú normálne dáta, ktoré nie sú anomáliou [1].

Kontextové anomálie predstavujú 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 [1].

Anomálne správanie inštancií 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 v inom ročnom období [1].

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

Skupinové anomálie sa nachádzajú v oblasti normálnych dát, ale skupina týchto inštancií tvorí spolu anomáliu. Vzniknutá anomália obsahuje sekvenciu inštancií, ktorá by pri inom zoradení nepredstavovala anomáliu. Taktiež sa jednotlivé inštancie môžu nachádzať v rozsahu normálnych dát. Príkladom môžu byť systémové volania operačného systému, ktoré sú v prípade dodržania určitej postupnosti označené ako činnosť škodlivého softvéru [1].

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 inštancií. Techniky používané pri detekcii skupinových anomálií sa značne líšia od techník používaných pri bodových a kontextových anomáliách [1].

2.2.2 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 inštancií 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 identifikácií anomálií. V súčasnosti existujú 3 prístupy, a to detekcia anomálií s učiteľom, bez učiteľa a ich kombinácia [1].

Detekcia bez učiteľa 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].

Detekcia s učiteľom potrebuje trénovacie dáta s označenými inštanciami ako normálnymi, tak aj anomálnymi. Cieľom je vytvoriť prediktívny model, ktorého úlohou je určiť triedu inštancie. Problémom je, že anomálnych inštancií v porovnaní s normálnymi je omnoho menej a označenie dát ľudským expertom môže byť pri anomálnej inštancií náročné [1].

Kombinácia týchto dvoch prístupov počíta s označenou iba jednou triedou inštancií. Typicky sú označené normálne inštancie, keďže ich identifikácia je menej náročná. V takom

prípade je vytvorený model pre normálnu triedu a identifikácia anomálií prebieha v testovacej vzorke dát [1].

2.3 Zhlukovanie časových radov

Cieľom zhlukovania je rozdeliť dátové inštacie do k zhlukov na základe spoločných čŕt. V prípade, že inštancie sú reprezentované nízkodimenzionálnym vektorom v Euklidovskom priestore, môžu byť na zhlukovanie použité klasické techniky ako napr. *k-means*. Ak sú inštancie reprezentujú časový rad, nasadenie takýchto prístupov je zriedkavé [5].

Metódy používané na meranie vzdialenosti medzi časovými radmi môžeme rozdeliť do 3 skupín, založených na atribútoch, na modeloch a na tvare krivky. Pri atribútových metódach je pre každý časový rad vypočítaný atribútový vektor, na základe, ktorého je vypočítaná Euklidovská vzdialenosť medzi jednotlivými inštanciami. Modelové techniky používajú parametrický model, do ktorého vstupujú časové rady. Vzdialenosťou je potom definovaná ako vzdialenosť medzi jednotlivými modelmi. Metódy porovnávajúce tvary kriviek sa snažia prispôsobiť výsledný tvar časového radu nelineárnym rozťahovaním a kontrakciou časových osí [5].

2.3.1 Semi-supervised learning

Časové rady sú charakteristické svojou vysokou dimenzionalitou, spojitou povahou a číselnou reprezentáciou. Prístupy používané na ich klasifikáciu preto možno rozdeliť do dvoch kategórií, podľa toho či sú založené na atribútoch (angl. feature-based) alebo samotných inštanciách (angl. instance-based). Zatiaľ čo prvý prístup pôvodné dáta transformuje do nového priestoru, kde môžu byť použité konvenčné klasifikátory, druhý sa zameriava na vytváranie klasifikátorov špeciálne navrhnutých pre časové rady. Tento prístup je založený na vhodnej reprezentácií časového radu a identifikácií podobnosti medzi nimi [4].

Existuje niekoľko techník, používaných pri semi-supervised klasifikátoroch. Medzi najznámejšie prístupy patria cluster-then-label, prístupy založené na podobnosti nazývané aj geometrické párovanie šablón (angl. geometric template matching), rozšírenie klasických prístupov ako je súbor podporných vektorov alebo prístupy založené na grafoch. Ďalším prístupom sú samooznačovacie techniky (angl. self-labeled techniques), ktorých snahou je zväčšiť pôvodnú množinu označených dát o neoznačené dáta, ktorým bola priradená najvyššia miera pravdepodobnosti výskytu v danej triede. Tie sa ďalej delia a medzi najpopulárnejšie patria samotrénovacie (angl. self-training) techniky a spolutrénovacie (angl. co-training) techniky [4].

Samotrénovacie techniky iteratívne klasifikujú neoznačené dáta za predpokladu, že vyššia prenosť predikcie má tendenciu byť správna. V prípade, že k dispozícii dostupné iba dáta z jednej triedy, hovoríme o pozitívnom neoznačenom učení. Algoritmus k-najbližších susedov je typicky používaný na klasifikáciu takýchto úloh. Naopak, spolutrénovacie semi-supervised klasifikátory potrebujú dva podmienene nezávislé pohľady, ktoré sú použité pri učení klasifikátora. Opäť sú neoznačené inštancie s najvyššou mierou pravdepodobnosti označené a použité pri trénovaní ďalšieho pohľadu. Keďže väčšina prípadov meraní spolú úzko koreluje, je náročné nájsť nezávislé pohľady na dáta [4].

Samotrénovacie prístupy sa lýšia najmä v nasledujúcom:

Najčastejšími **mechanizmami pridávania** je inkrementovanie a zmena. Zatiaľ čo prvý, krok po kroku označuje najpravdepodobnejšie inštancie, druhý usmerňuje už označené inštancie a odznačuje tie, ktoré by mohli mohli v tejto množine predstavovať šum [4].

Samooznačovacie techniky môžu byť založené rôznych **klasifikačných modeloch**, ktoré pozostávajú z jedného alebo viacerých klasifikátoroch. Tie určujú triedu neoznačených inštancií. Zatiaľ čo modely s jedným klasifikátorom určia pre každú inštanciu triedu, do ktorej spadá, multi-klasifikátorové modely kombinujú hypotézy viacerých klasifikátorov a triedu určujú kombináciou ich výsledkov. S tým súvisí aj **prístup k učeniu**, ktorý môže byť single learning approach alebo multi learning approach, ktorý je aplikovateľný iba pri multi-klasifikátorových modeloch, kde každý klasifikátor vychádza z iného pristúpu [4].

Ukončovacie kritérium predstavuje mechanizmus, ktorý zabraňuje samooznačovaciemu procesu v označovaní inštancií, ktorým bola pridelená nízka miera pravdepodobnosti. Často je ukončovacím kritériom iba počet iterácií. Ďalším kritériom môže byť výskyt nemeniacich sa hypotéz počas trénovacieho procesu [4].

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].

2.4 Dátové atribúty a ich redukcia

2.5 Identifikácia anomálneho správania 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, 8].

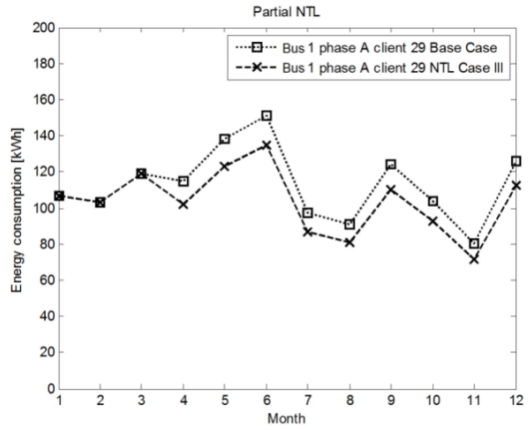
Najčastejšími metódami používanými pri nelegálnom odbere je obídenie meračov spotreby energie či samotná manipulácia s nimi. Merače tak poskytujú nesprávne informácie o spotrebovanej energie odberateľmi, čo je možné detegovať až po identifikácií celkových netechnických strát v sieti. Ďalšou populárnou metódou používanou na detekciu nelegálnych odberov je analýza spotrebiteľského profilu zákazníka, kedy je našou snahou identifikovať nepravidelné vzory v nameraných spotrebiteľských dátach [8]. 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. Takto sú identifikované aj také odbery, ktoré

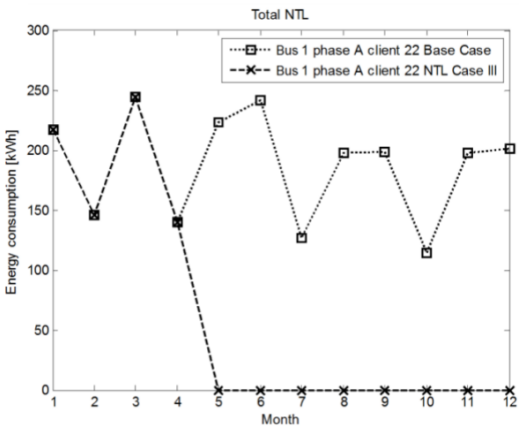
by sa pri klasických meraniach stratili v týždenných alebo mesačných agregáciách. Úspešnosť detekcie nelegálnych odberov je výrazne vyššia najmä pri neštandardných spotrebách alebo ak sa jedná o neopakujúcu sa udalosť. Problém vzniká ak odberateľ systematicky mení nelegálnu spotrebu a kopíruje vzory, ktoré vznikajú v dátach pri legálnom odbere. Vtedy je potrebné mať k dispozícií väčšie množstvo dát a zároveň použiť zložitejšie algoritmy detekcie anomálií, ktoré sú popísané v súvisiacej práci [7].

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 dispozícií 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.

Typickou črtou netechnických strát je negatívny skok v spotrebe elektrickej energie. Nasleduje po poškodení inteligentného meracieho zariadenia alebo pri začatí nelegálneho odberu. Následkom je nižšia nameraná spotrebe energie v dlhšom horizonte. Zníženie spotreby môže byť čiastočné alebo úplne, ako môžeme vidieť na obrázkoch 1 a 2 [11].



Obr. 1: Čiastočné zníženie spotreby elektrickej energie [11].



Obr. 2: Úplné zníženie spotreby elektrickej energie [11].

2.6 Vyhodnocovacie metriky

2.6.1 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.

Klasifikácia Pomocou naučeného modelu, nazývaného aj klasifikátor, sú rozoznávané 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, Bayesovych sieťach, pravidlových systémoch či SVM [1, 10].

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 inštancie 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, 10].

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, 10].

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

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

Posledná kategória pracuje s predpokladom, že normálne inštancie 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 [9].

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 dispozícií 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 identifikácií 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 identifikácií. 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. 15:1–15:58, 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 electricty theft for enhanced performance and security of power grid*. Dizertačná práca, The University of Toledo, 2012.

[4] González, M.; Bergmeir, C.; Triguero, I.; aj.: Self-labeling techniques for semi-supervised time series classification: an empirical study. *Knowledge and Information Systems*, 2017: s. 1–36, ISSN 02193116, doi:10.1007/s10115-017-1090-9.

[5] Hautamaki, V.; Nykanen, P.; Franti, P.: Time-series clustering by approximate prototypes. *2008 19th International Conference on Pattern Recognition*, 2008: s. 1–4, ISSN 1051-4651, doi:10.1109/ICPR.2008.4761105.
URL `http://ieeexplore.ieee.org/document/4761105/`

[6] Meffe, A.; de Oliveira, C. C. B.: Technical loss calculation by distribution system segment with corrections from measurements. In *CIRED 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.

[7] 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.

[8] 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.

[9] 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.

[10] Tan, P.-N.; Steinbach, M.; Kumar, V.: *Introduction to Data Mining*. Addison Wesley, us ed vydanie, May 2005, ISBN 0321321367.

[11] Trevizan, R. D.; Bretas, A. S.; Rossoni, A.: Nontechnical Losses detection: A Discrete Cosine Transform and Optimum-Path Forest based approach. *2015 North American Power Symposium, NAPS 2015*, October 2015, doi:10.1109/NAPS.2015.7335160.