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 [8], 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 analytiky 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 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 odberateľov pomocou získaných dát.

2.1 Časové rady

2.1.1 Delenie časových radov

Výraznými vlastnosťami časových radov sú aj periodicita a synchrónnosť. Vznikajú tak 4 nasledujúce kategórie [14]:

- **Periodické a synchrónne časové rady** predstavujú najjednoduchšiu kombináciu, keďže každý časový rad má konštantnú časovú periódu a zároveň sú všetky časové rady časovo zarovnané na konkrétny časový bod.

- **Neperiodické a synchrónne časové rady** nemajú žiadnu periodicitu, ale opäť sú časovo zarovnané.

- **Periodické a asynchrónne časové rady** nie sú časovo zarovnané, ale obsahujú periodicitu, čiže začiatok periódy v každom časovo rade je iný.

- **Neperiodické a asynchrónne časové rady** predstavujú skupinu, do ktorej spadajú ostatné časové rady, ktoré neobsahujú periodicitu a ani synchrónnosť.

2.2 Detekcia 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.

Pri identifikácii anomálií je najskôr potrebné zamyslieť sa nad nasledovnými problémami [1]:

- **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.

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*) [3].

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 (angl. *supervised learning*), bez učiteľa (angl. *unsupervised learning*) a ich kombinácia (angl. *semi-supervised learning*) [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].

Kombinované učenie je kombináciou predchádzajúcich dvoch prístupov a 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].

2.2.3 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, Bayesových sieťach, pravidlových systémoch či metóde podporných vektorov [1, 13].

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, 13].

Zhlukovanie Jedná sa o učenie bez učiteľa, keďže zhluky inštancií sú vytvorené na základe ich vzdialenosti či podobnosti. Techniky ďalej delíme do kategórií na základe predpokladu o dátových inštanciách [1, 13].

Prvá kategória predpokladá, že normálne inštancie patria do zhluku, zatiaľ čo anomálne nepatria do žiadneho. Používané sú zhlukovacie algoritmy ako DBSCAN alebo ROCK, pri ktorých nie nutne každá inštancia musí patriť do zhluku. Nevýhodou algoritmov môže byť neoptimálne použitie pri detekcií anomálií, keďže sú primárne určené na riešenie zhlukovacích problémov [1].

Druhá kategória predpokladá, že normálne inštancie ležia v blízkosti najbližšieho centroidu a anomálne inštancie sú od neho vzdialené. Algoritmy väčšinou pozostávajú z dvoch krokov, v prvom sú inštancie pridelené do zhluku a v druhom je vypočítané ich anomálne skóre na základe vzdialenosti od centroidu daného zhluku. Používanými algoritmami sú neurónové siete (konkrétne SOM) alebo algoritmus k-means, ktoré sa môžu učiť aj pomocou kombinovaného učenia [1].

Posledná kategória pracuje s predpokladom, že normálne inštancie sú súčasťou veľkých a hustých zhlukov, na druhej strane anomálie patria do malých a riedkych zhlukov. Používanými algoritmami sú napr. CBLOF (angl. *Cluster-Based Local Outlier Factor*) alebo *k-d* stromy. V princípe algoritmy najskôr vytvárajú zhluky a až potom určujú, na základe ich hustoty, či sa jedná o normálne zhluky alebo anomálie. Zhluk je vytvorený iba v prípade, že inštancia sa nachádza mimo preddefinovaného rádiusu od centra daného zhluku [11].

2.3 Zhlukovanie časových radov

Cieľom zhlukovania je rozdeliť dátové inštancie 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 spomenuté v 2.2.3. Ak inštancie reprezentujú časový rad, nasadenie takýchto štandardných prístupov je zriedkavé [6].

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 [6].

2.3.1 Hierarchické zhlukovanie

Táto metóda produkuje vnorenú hierarchiu skupín podobných časových radov na základe vzdialenostných matíc jednotlivých inštancií. Výhodou je, že nie je nutné zadávať počet zhlukov, ktoré ideme identifikovať. Nevýhodou je obmedzenie výpočtu iba na menšie datasety, keďže táto výpočtová zložitosť tejto metódy je kvadratická [5].

Metóda hierarchického zhlukovania zoskupuje časové rady do stromu zhlukov. Vo všeobecnosti existujú dva typy týchto metód, aglomeratívne a deliace. Aglomeratívne metódy zo začiatku umiestňujú časové rady do samostatného zhluku, až potom ich postupne spájajú do väčších zhlukov až pokiaľ neexistuje jediný zhluk alebo nie je ukončovacou podmienkou práve k zhlukov. Deliace metódy sú pravým opakom, kedy sú jednotlivé zhluky postupne delené na menšie a umiestňované do hierarchického stromu. Na zlepšenie kvality zhlukovania pri hierarchickom zhlukovaní sú používané bežné zhlukovacie techniky [16].

Aglomeratívne zhlukovanie Vzdialenosť medzi dvoma zhlukmi je meraná pomocou dvojice najbližších časových radov umiestnených v rôznych zhlukoch, ktoré sú potencionálnymi kandidátmi na zlúčenie. Podobnosť môže určovať aj *Wardov algoritmus minimálnej variance*, ktorý zlúči zhluky s najmenším nárastom variancie. V každom kroku sú tak vyskúšané všetky kombinácie dvojíc zhlukov, až potom je vybrané minimum. Porovnávané časové rady nemusia mať vždy rovnakú dĺžku. Nevýhodou metódy je najmä vysoký počet operácií, ale aj neschopnosť spätne zmeniť rozhodnutie zlúčiť zhluky [16].

Deliace zhlukovanie Algoritmus nie je obmedzený iba na časové rady rovnakej dĺžky. Zároveň tiež nie je možné zmeniť delenie zhluku, ktoré už bolo vykonané. Na meranie vzdialenosti môžu byť použité metriky opísané v 2.3.3 [16].

2.3.2 Samoorganizované mapy

Trieda neurónových sietí, kde neuróny sú usporiadané v nízkodimenzionálnej štruktúre a trénované iteratívne a bez učiteľa. Trénovací proces začína pridelením náhodných hodnôt váhovým vektorom w. Každá iterácia trénovania pozostáva z 3 krokov a to náhodného výberu vstupného vektoru z trénovacej množiny, evaluácie siete a aktualizovaní váhových vektorov. Po natrénovaní je vypočítaná Euklidovská vzdialenosť medzi vstupným vzorom a váhovým vektorom. Následne je neurón s najmenšou vzdialenosťou označený ako t a ostatné váhy ostatných neurónov sú aktualizované v závislosti od vzdialenosti od neuróna t. Nevýhodou je opäť náročné spracovanie časových radov rôznych dĺžok, keďže dĺžka časového radu definuje aj dĺžku váhového vektora w [7, 16].

2.3.3 Metriky vzdialenosti

Kľúčovou záležitosťou pri zhlukovaní časových radov na základe ich podobnosti, je meranie vzdialenosti medzi nimi. Rovnako ako pri zhlukovaní bodových inštancií je potrebné definovať si metódy merania vzdialenosti. Najčastejšími metrikami sú Euklidovská a Manhattanovská

vzdialenosť. Vhodnosť aplikovania týchto klasických metód nízka, keďže nameraná vzdialenosť zachytáva aj použitú škálu v dátach. Pri porovnávaní časových radov nás spravidla zaujíma zmena krivky časového radu a rovnaká dĺžka porovnávaných časových radov [5, 16].

Korelácia Korelačný koeficient $r(X, Y)$ meria stupeň lineárnej závislosti medzi dvoma časovými radmi X a Y. Vyjadríme ho vzorcom 1. Korelácia blízka -1 znamená, že nárast kriviek časových radov je zrkadlový, pri korelácií rovnej 0 hovoríme o rozdielnych časových radoch a pri hodnote 1 o podobných. Na základe hodnoty korelácie, potom môžeme vyjadriť vzdialenosť vzorcom 2. Nevýhodou je, ak máme k dispozícií iba malú, prípadne krátku časť datasetu, podobnosť touto metrikou sa určuje len ťažko. Keďže korelácia zachytáva iba lineárnu podobnosť časových radov, pri aplikovaní metriky na dva nelineárne podobné časové rady, sú vyhodnotené ako vzdialené [5].

$$r\left(X,Y\right) = \frac{E\left[\left(X - E\left[X\right]\right) \cdot \left(Y - E\left[Y\right]\right)\right]}{E\left[\left(X - E\left[X\right]\right)^2\right] \cdot E\left[\left(Y - E\left[Y\right]\right)^2\right]} \tag{1}$$

$$D_r\left(X,Y\right) = \sqrt{0.5 \cdot \left(1 - r\left(X,Y\right)\right)} \tag{2}$$

Dynamické deformovanie času Ide o metódu (angl. *Dynamic Time Warping*), ktorá dokáže zachytiť nelineárne skreslenie medzi časovými radmi vďaka prideleniu viacerých hodnôt časového radu X druhému časovému radu Y. Takto metóda viac zodpovedá ľudskej intuícii. D_{DTW} je vypočítané pomocou dynamického programovania a vyjadrené pomocou vzorca 3 [5].

$$D_{DTW}\left(i,j\right) = \begin{cases} d\left(x_i, y_j\right) + min \begin{cases} D_{DTW}\left(i-1, j\right) \\ D_{DTW}\left(i, j-1\right) \text{ ak } i \neq 0 \text{ a } j \neq 0 \\ D_{DTW}\left(i-1, j-1\right) \end{cases} \\ 0 \text{ ak } i = 0 \text{ a } j = 0 \\ \infty \text{ inak} \end{cases} \tag{3}$$

Kvalitatívna vzdialenosť Metóda je založená na kvalitatívnom porovnávaní tvaru dvoch časových radov. Pre časové rady X a Y vyberieme dvojicu bodov i a j, ktoré označujú zmenu premennej v danom časovom rade. Tak vznikajú 3 možnosti, hodnoty v časovom rade rastú ($X_i < X_j$), nemenia sa ($X_i \approx X_j$) alebo klesajú ($X_i > X_j$). Vzdialenosť potom vyjadríme vzorcom 4, pomocou ktorého spočítame počet zhôd v raste časových radov. Práve funkcia $Diff(q_1, q_2)$ vyjadruje rozdiel v zmene rastu. Metóda nemá nevýhody, ktoré vznikali pri korelácií, na druhú stranu je aplikovateľná iba na krátke časové rady bez toho, aby sa dramaticky znížila kvalita odhadu vzdialenosti. Podobnosť tvarov kriviek je detegovaná aj v prípade, kedy neexistuje medzi časovými radmi lineárna alebo nelineárna závislosť [5].

$$D_q\left(X,Y\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2 \cdot Diff\left(q\left(X_i, X_j\right), q\left(Y_i, Y_j\right)\right)}{N \cdot \left(N-1\right)} \tag{4}$$

Euklidovská vzdialenosť Je používaná najmä pri klasických zhlukovacích problémoch. Ak zvolený časový rad má dĺžku n, vzdialenosť vypočítame vzorcom 5 [16].

$$D_E\left(X,Y\right) = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - Y_{jk}\right)^2} \tag{5}$$

Manhattanovská vzdialenosť Je rovnako ako Euklidovská vzdialenosť používaná najmä pri klasických zhlukovacích problémoch. Výpočet je tiež veľmi podobný, môžeme ho vyjadriť vzorcom 6

$$D_M(X,Y) = \sum_{k=1}^{n} |X_{ik} - Y_{jk}| \tag{6}$$

Pearsonov korelačný koeficient Vo vzorci 7 reprezentuje $\widetilde{X}$ aritmetický priemer časového radu X. Koeficient je používaný pri výpočte vzdialenosti, ktorá je založená na vzájomnej korelácií. Vzdialenosť vyjadríme vzorcom 7 [16].

$$r(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \widetilde{X}\right)\cdot\left(Y_i - \widetilde{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \widetilde{X}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(Y_i - \widetilde{Y}\right)^2}} \tag{7}$$

$$D_P(X,Y) = 2\cdot(1 - r(X,Y)) \tag{8}$$

Vzdialenosť medzi krátkymi časovými radmi Metóda (angl. *Short time series*) meria vzdialenosť ako sumu štvorcových rozdielov medzi krivkami jednotlivých časových radov. Na odstránenie nežiadúcich efektov škály sa používa z štandardizácia. Matematicky vzdialenosť vyjadrime vzorcom 9. Zložka t_k predstavuje čas [16].

$$D_{STS}(X,Y) = \sqrt{\sum_{k=1}^{n}\left(\frac{Y_{j(k+1)} - Y_{jk}}{t_{(k+1)} - t_k} - \frac{X_{i(k+1)} - X_{ik}}{t_{(k+1)} - t_k}\right)^2} \tag{9}$$

2.4 Redukcia dimenzií a výber atribútov

2.4.1 Výber atribútov

Väčšina dát pochádzajúcich z inteligentných meračov obsahuje iba stĺpce s časovou známkou a momentálnou spotrebou daného uzlu v sieti. Z týchto informácií ešte vieme vyčítať, mesiac, týždeň, deň prípadne deň v týždni alebo sviatok. Niektoré z extrahovaných atribútov úzko súvisia s funkciou spotreby elektrickej energie. Pri vytváraní presného modelu je preto nevyhnutné správne identifikovať takéto atribúty. Otestovanie všetkých kombinácií by bolo časovo a výpočtovo náročné. Najjednoduchším spôsobom je vytvorenie korelačnej matice jednotlivých vysvetľujúcich premenných a sledovanej veličiny [2].

2.4.2 Agregácia dát

Dáta z meračov sú dostupné v pravidelných intervaloch. Pre jednoduchšiu manipuláciu s časovými radmi a redukciu dimenzií môžu byť dáta agregované do väčších intervalov. Pri použití viacerých datasetov s rôznou frekvenciou zberu, je agregácia hustejšieho časového radu nutná, keďže by tak vzniklo množstvo chýbajúcich hodnôt. Agregácia dát tiež vyhladzuje malé odchýlky v časových radoch, čo môže sťažiť identifikáciu náhlej zmeny správania odberateľov. To môže viesť k nesprávnemu označeniu správania odberateľa za neštandardné [2].

2.5 Identifikácia abnormálneho správania

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 [3, 10].

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 [10]. Tak ako je spomenuté v práci [4], 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.

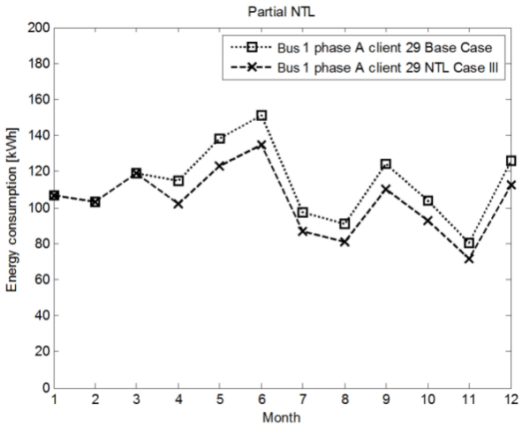
Vďaka inteligentným meračom je možné detegovať nelegálne odbery omnoho rýchlejšie, najmä kvôli vysokej frekvencii zberania údajov. Takto sú identifikované aj také odbery, ktoré by sa pri klasických meraniach stratili v týždenných alebo mesačných agregáciách. Úspešnosť detekcie nelegálnych odberov je 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 [9].

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 [3] 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.

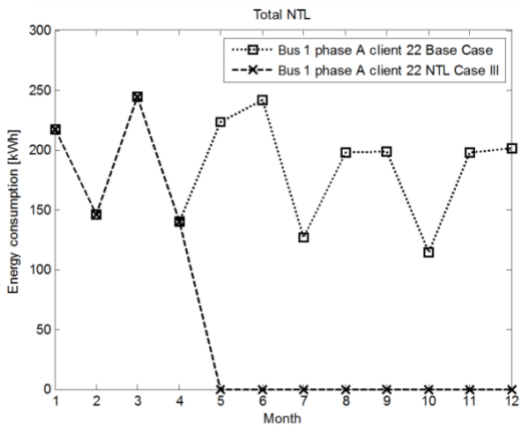
Typickou črtou netechnických strát je negatívny skok v spotrebe elektrickej energie. Nasleduje po poškodení inteligentného meracieho zariadenia alebo pri začatí nelegálneho odberu. Následkom je nižšia nameraná spotrebe energie v dlhšom horizonte. Zníženie spotreby môže byť čiastočné alebo úplne, ako môžeme vidieť na obrázkoch 1 a 2 [15].

2.6 Vyhodnocovacie metriky

Za predpokladu, že získané dáta budú obsahovať aj označené inštancie, prípadne budú označené dodatočne na základe výpočtov, môžeme na vyhodnotenie úspešnosti použiť aj CONFUSION MATRIX. V takom prípade budeme musieť predpovedať triedu jednotlivých inštancií, a teda či sa jedná o normálneho alebo anomálneho odberateľa. Jednoduchý klasifikátor označí prvých n odberateľov, ktorých miera pravdepodobnosti výskytu čierneho odberu je najvyššia, za anomálnych. Pri použití CONFUSION MATRIX potom riadky predstavujú predpovedanú triedu a stĺpce skutočnú. Vznikajú tak 4 kategórie, ktoré sú označované ako *TRUE POSITIVE*



Obr. 1: Čiastočné zníženie spotreby elektrickej energie [15].



Obr. 2: Úplné zníženie spotreby elektrickej energie [15].

(správne označení podozriví odberatelia), *FALSE POSITIVE* (nesprávne označení podozriví odberatelia), *TRUE NEGATIVE* (nesprávne označený normálny odberatelia) a *FALSE NEGATIVE* (správne označený normálny odberatelia). Kvalitu klasifikácie potom môžeme zmerať pomocou presnosti a citlivosti. Presnosť vypočítame vzorcom 10, kedy ide o pomer správne označených anomálií a celkový počet označených anomálií. Tým vypočítame percento odberateľov, ktorých sme správne klasifikovali ako podozrivých.

$$\text{Presnosť} = \frac{TP}{TP + FP} \tag{10}$$

Citlivosť označuje pomer správne označených anomálií a celkový počet skutočných anomálií. Vyjadríme ju pomocou vzorca 11.

$$\text{Citlivosť} = \frac{TP}{TP + FN} \tag{11}$$

Aby sa predišlo situácií, kedy sa v dátach nachádza iba malý počet anomálnych odberateľov a pre model by tak bolo výhodnejšie označovať iba tých, s ktorými si je takmer istý, je dôležité brať do úvahy aj túto metriku. Obe metriky sú vyjadrené v percentách [15].

Ďalšou používanou metrikou je aj tzv. F-skóre, ktoré obsahuje informácie oboch predchádzajúcich metrík. Keďže ide o súčet metrík, tiež je vyjadrené v percentách. Cieľom práce je maximalizovať túto metriku. F-skóre vyjadríme pomocou vzorca 12, kde P predstavuje presnosť a C predstavuje citlivosť [15].

$$F = 2 \cdot \left(P^{-1} + C^{-1}\right)^{-1} \tag{12}$$

2.7 Súvisiace práce v doméne energetiky

V [6] bola pri zhlukovaní použitá aj kombinácia viacerých metód, konkrétne k-means, metóda náhodnej výmeny a aglomeratívne zhlukovanie. Ako už bolo spomenuté v 2.2.3, úlohou algoritmu k-means namapovať existujúce inštancie do *k* zhlukov. Aj keď metóda náhodnej výmeny je obmedzená na zhlukovacie problémy v Euklidovskom priestore, bola použitá aj pri zhlukovaní časových radov a zabraňuje zaseknutiu zhluku v lokálnom minime. V princípe je náhodne vybraný zhluk, ktorý bude vymazaná a za centroid bude vybraný jeden časový rad z neho. Ak takéto riešenie je lepšie ako bez rozpustenia zhluku je nahradené pôvodným. Ako

bolo spomenuté v 2.3.1 cieľom aglomeratívneho zhlukovania je všetky časové rady označiť ako zhluky a následne ich iteratívne zhlukovať. V momente, keď je vytvorených k zhlukov, je vypočítaný centroid zhluku a určená hierarchia zhlukov.

V práci [2] boli pri určovaní podozrivých aktivít odberateľov úspešne aplikované rozhodovacie stromy. Po vytvorení trénovacej a testovacej množiny boli vygenerované rozhodovacie pravidlá reprezentujúce model normálnej spotreby elektrickej energie. Po predikcii boli porovnané predikované a testovacie dáta pomocou štatistickej metódy RMSE. Výsledkom experimentov je dostatočne presná predikcia spotreby energie vypočítaná iba na základe atribútov extrahovaných z časovej známky. Prekročením stanovej hranice boli inštancie považované za anomálne. Počas experimentov boli použité M5P rozhodovacie učiace stromy.

Predmetom článku [12] bolo navrhnúť novú vlnovú techniku na reprezentovanie viacerých vlastností meraných dát. Tiež vytvorili nový model, ktorý v sebe zahŕňa viacero modelov, čím je pridávanie ďalších komponentov do detekčného systému jednoduché. Navrhovaná metóda je citlivá na lokálne zmeny vo vzore dát. Taktiež dosiahli s relatívne malým množstvom meraní presnosť až 78% na trénovacej množine a 70% na testovacej množine. Metóda je citlivá na zmeny amplitúd a frekvencií v dátach z meračov. Nevýhodou je, že model nie je citlivý na nevýrazne zmeny a trendy v dátach.

3 Špecifikácia požiadaviek

4 Rozpracovanie problému

Pomocou metód strojového učenia a dátovej analytiky 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] Cody, C.; Ford, V.; Siraj, A.: Decision Tree Learning for Fraud Detection in Consumer Energy Consumption. *the 14th IEEE International Conference on Machine Learning and Applications*, 2015, doi:10.1109/ICMLA.2015.80.

[3] 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.

[4] 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.

[5] Dzeroski, S.; Gjorgjioski, V.; Slavkov, I.; aj.: Analysis of time series data with predictive clustering trees. *Knowledge Discovery in Inductive Databases*, 2007: s. 47–58, ISSN 03029743, doi:10.1007/978-3-540-75549-4˙5.

[6] 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/`

[7] Kohonen, T.; Schroeder, M. R.; Huang, T. S. (editori): *Self-Organizing Maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., tretie vydanie, 2001, ISBN 3540679219.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Tagaris, H.; Lachsz, A.; Jeffrey, M.: Wavelet based feature extraction and multiple classifiers for electricity fraud detection. *IEEE/PES Transmission and Distribution Conference and Exhibition*, ročník 3, č. November 2002, 2002: s. 2251–2256, doi:10.1109/TDC.2002.1177814.
URL `http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1177814'escapeXml='false'/{%}3E`

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

[14] Teng, M.: Anomaly detection on time series. *2010 IEEE International Conference on Progress in Informatics and Computing*, ročník 1, 2010: s. 603–608, doi:10.1109/PIC.2010.5687485, 1708.02975.
URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5687485

[15] 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.

[16] Warren Liao, T.: Clustering of time series data - A survey. *Pattern Recognition*, ročník 38, č. 11, 2005: s. 1857–1874, ISSN 00313203, doi:10.1016/j.patcog.2005.01.025.