

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta informatiky a informačných technológií

Optimalizácia konfiguračných parametrov predikčných metód

BAKALÁRSKA PRÁCA

2016

Matúš Cuper

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta informatiky a informačných technológií

Optimalizácia konfiguračných parametrov predikčných metód

BAKALÁRSKA PRÁCA

Študijný program: Informatika

Číslo študijného odboru: 9.2.1

Názov študijného odboru: Informatika

Školiace pracovisko: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava

Vedúci záverečnej práce: Ing. Marek Lóderer

Bratislava 2016

Matúš Cuper

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Študijný program: Informatika

Autor: Matúš Cuper

Bakalárska práca: Optimalizácia konfiguračných parametrov predikčných metód

Vedúci práce: Ing. Marek Lóderer

máj 2016

Tu bude text slovenskej anotácie

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree Course: Computer Science

Author: Matúš Cuper

Bachelor thesis: Optimizing configuration parameters of prediction methods

Supervisor: Ing. Marek Lóderer

May 2016

Tu bude text anglickej anotácie

POĎAKOVANIE
Tu bude poďakovanie

ČESTNÉ PREHLÁSENIE

Tu bude prehlásenie

.....
Matúš Cuper

Obsah

1	Analýza problému	6
1.1	Časové rady	6
1.1.1	Analýza časových radov	6
1.1.2	Zložky časových radov	6
1.2	Analýza predičných algoritmov	7
1.2.1	Lineárna regresia	7
1.2.2	Stochastické modely	8
1.2.3	Support vector regression	9
1.2.4	Rozhodovacie stromy	10
1.2.5	Random forrest	10
1.2.6	Neurónové siete	10
1.2.7	Učenie súborov klasifikátorov	10
1.2.8	Exponencionálne hľadanie	11
1.2.9	Naivné metódy	11
1.3	Analýza optimalizačných algoritmov	11
1.3.1	Artificial bee colony	12
1.3.2	Kolónia mravcov	13
1.4	Meranie presnosti predpovedi	13
1.4.1	Stredná chyba predpovede	13
1.4.2	Stredná absolútna chyba	13
1.4.3	Stredná absolútna percentuálna chyba	14
1.4.4	Stredná percentuálna chyba	14
1.4.5	Stredná štvorcová chyba	14
2	Opis riešenia	15
3	Zhodnotenie	16
4	Technická dokumentácia	17

1 Analýza problému

1.1 Časové rady

Časový rad je množina dátových bodov nameraná v čase postupne za sebou. Matematicky je definovaný ako množina vektorov $x(t)$, kde t reprezentuje uplynulý čas. Premenná $x(t)$ je považovaná za náhodnú premennú. Merania v časových radoch sú usporiadané v chronologicky poradí [1].

Časové rady delíme na spojité a diskrétny. Pozorovania pri spojitých časových radoch sú merané v každej jednotke času, zatiaľ čo diskrétny obsahujú iba pozorovania v diskrétnych časových bodoch. Hodnoty toku rieky, teploty či koncentrácie látok pri chemickom procese môžu byť zaznamenané ako spojitý časový rad. Naopak, populácia mesta, produkcia spoločnosti alebo kurzy mien reprezentujú diskrétny časový rad. Vtedy sú pozorovania oddelené rovnakými časovými intervalmi, napr. rokom, mesiacom či dňom [1]. V našom prípade sú namerané dáta dostupné každú celú štvrt' hodinu.

1.1.1 Analýza časových radov

V praxi je vhodný model napasovaný do daného časového radu a zodpovedajúce parametre sú predpovedané na základe známych dát. Pri predpovedaní časových radov sú dáta z uplynulých meraní zhromažďované a analyzované za účelom navrhnutia vhodného matematického modelu, ktorý zachytáva proces generovania dát pre časové rady. Pomocou tohto modelu sú predpovedané hodnoty budúcich meraní. Takýto prístup je užitočný, keď nemáme veľa poznatkov o vzore v meraniach idúcich za sebou alebo máme model, ktorý poskytuje nedostatočne uspokojivé výsledky [1].

Cieľom predikcií časových radov je predpovedať hodnotu premennej v budúcnosti na základe doteraz nameraných dátových vzoriek. Matematicky zapísané ako

$$\hat{x}(t + \Delta_t) = f(x(t - a), x(t - b), x(t - c), \dots) \quad (1)$$

Hodnota \hat{x} je predpovedaná ako hodnota diskrétného časového radu x . Preto je potrebné nájsť funkciu $f(x)$, podobnú funkciu \hat{x} , ktorá predpovedá hodnotu časového radu v budúcnosti konzistentne a objektívne [2].

Časové rady sú najčastejšie vizualizované ako graf, kde pozorovania sú na osy y a plynúci čas na osy x .

1.1.2 Zložky časových radov

Pri predpovedaní časových radov ako napr. meraní odberu elektriky vznikajú 2 typy trendov. Prvým typom je trvalá alebo dočasná zmena spôsobená ekonomickými alebo ekologickými faktormi. Druhým typom je sezónna zmena, spôsobená zmenami ročných období a množstvom denného svetla. Môžeme ju pozorovať na úrovni dní, týždňov alebo rokov. Veličina, ktorú sa snažíme predpovedať postupne mení svoje správanie a model sa tak stáva nepresným. Kvôli tomu je nutné v každom modeli rozdeliť tieto typy tendencií, aby sme vedeli model zmenám prispôbiť [3].

Vo všeobecnosti sú časové rady zložené zo 4 hlavných zložiek, ktoré môžeme odlíšiť od pozorovaných dát. Jedná sa o trendovú, cyklickú, sezónnu a reziduálnu zložku [1].

Trendová zložka V dlhodobom časovom horizonte majú časové rady tendenciu klesať, rásť alebo stagnovať. Príkladom môže byť nárast populácie či klesajúca úmrtnosť [1].

Cyklická zložka V strednodobom časovom horizonte sa vyskytujú okolnosti, ktoré spôsobujú cyklické zmeny v časových radoch. Dĺžka periódy je 2 a viac rokov. Táto zložka je zastúpená najmä pri ekonomických časových radoch napríklad podnikateľský cyklus pozostávajúci zo 4 fáz, ktoré sa stále opakujú [1].

Sezónna zložka Ide o kolísanie počas ročných období. Dôležitými faktormi pri tom sú napr. klimatické podmienky, tradície alebo počasie. Napríklad predaj zmrzliny sa v lete zvyšuje, ale počet predaných lyžiarskych súprav klesá [1].

Reziduálna zložka Jedná sa o veličinu, ktorá nemá žiadny opakovateľný vzor a ani dlhodobý trend. V časových radoch má nepredvídateľný vplyv na pozorovanú veličinu. V štatistike zatiaľ nie je definovaná metóda na jej meranie. Označuje sa aj ako náhodná zložka alebo biely šum. Je spôsobená nepredvídateľnými a nepravidel'nými udalosťami [1].

Vo všeobecnosti sa pre tieto 4 zložky používajú 2 rôzne modely. Je to multiplikatívny model a aditívny model.

$$\begin{aligned} Y(t) &= T(t) \times S(t) \times C(t) \times I(t) \\ Y(t) &= T(t) + S(t) + C(t) + I(t) \end{aligned} \quad (2)$$

Vo vzorci 2 predstavuje $Y(t)$ meranie v čase t . Premenné $T(t)$, $S(t)$, $C(t)$ a $I(t)$ sú zložkami trendu, sezónnosti, cyklu a náhodnosti. Multiplikatívny model je založený na predpoklade, že časové rady môžu byť na sebe závislé a môžu byť ovplyvňované medzi sebou, zatiaľ čo aditívny model predpokladá nezávislosť zložiek [1].

1.2 Analýza predičných algoritmov

Na základe množstva predikčných

1.2.1 Lineárna regresia

Najpoužívanejšia štatistická metóda, ktorá modeluje vzťah závislej premennej a vysvetľujúcej premennej. Závislú premennú predstavuje veličina, ktorú sa snažíme predpovedať, čo je v našom prípade spotreba elektriky. Vysvetľujúca premenná v sebe zahŕňa rôzne faktory, ktoré ovplyvňujú závislú premennú. Môžeme si pod tým predstaviť deň v týždni, počasie, tradície alebo rôzne udalosti, ktoré majú vplyv na predpoveď. [4].

Predpokladajme typický regresný problém. Dáta pozostávajúce z množiny n meraní majú formát $\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$. Úlohou regresie je odvodiť funkciu \hat{f} z dát, kde

$$\hat{f} : X \rightarrow \mathbb{R}, \text{ kde, } \hat{f}(x) = f(x), \forall x \in X, \quad (3)$$

Funkcia f vo vzorci 3 reprezentuje reálnu neznámu funkciu. Algoritmus použitý na odvodenie funkcie \hat{f} sa nazýva indukčný algoritmus alebo žiak. Funkcia \hat{f} sa nazýva model alebo prediktor. Obvykle je úlohou regresie minimalizovať odchýlku funkcie pre štvorcovú chybu, konkrétne strednú štvorcovú chybu MSE [5].

Keďže časový rad pozostáva z viacerých zložiek, môžeme ho zapísať ako funkciu $L(t)$ definovanú ako

$$L(t) = L_n(t) + \sum a_i x_i(t) + e(t) \quad (4)$$

Vo vzorci 4 funkcia $L_n(t)$ predstavuje odber elektriky v čase t . Hodnota a_i je odhadovaný pomaly meniaci sa koeficient. Faktory $x_i(t)$ nezávisle vplývajú na spotrebu elektriky. Môže sa jednáť napríklad o počasie alebo zvyky ľudí. Komponent $e(t)$ je biely šum, ktorý má nulovú strednú hodnotu a pevnú varianciu. Číslo n je počet meraní, obvykle 24 alebo 168, v našom prípade 96 meraní počas jedného dňa [4].

Lineárna regresná analýza Regresná analýza je štatistická metóda používaná na modelovanie vzťahov, ktoré môžu existovať medzi veličinami. Nachádza súvislosti medzi závislou premennou a potenciálnymi vysvetľujúcimi premennými. Používame pri tom vysvetľujúce premenné, ktoré môžu byť namerané súčasne so závislými premennými alebo aj premenné z úplne iných zdrojov. Regresná analýza môže byť tiež použitá na zlúčenie trendu a sezónnych zložiek do modelu. Keď je raz model vytvorený, môže byť použitý na zásah do spomínaných vzťahov alebo, v prípade dostupnosti vysvetľujúcich premenných, na vytvorenie predikcie [6].

Viacnásobná lineárna regresia Viacnásobná regresia sa pokúša modelovať vzťah medzi dvoma alebo viacerými vysvetľujúcimi premennými a závislou premennou vhodnou lineárnou rovnicou pre pozorované dáta. Výsledný model je vyjadrený ako funkcia viacerých vysvetľujúcich premenných [3].

Túto funkciu môžeme zapísať ako

$$Y(t) = V_t a_t + e_t \quad (5)$$

Vo vzorci 5 t označuje čas, kedy bolo meranie uskutočnené. $Y(t)$ predstavuje celkový nameraný odber elektriky. Vektor V_t reprezentuje hodnoty vysvetľujúcich premenných v čase merania. Vysvetľujúce premenné môžu predstavovať meteorologické vplyvy, ekonomický nárast, ceny elektriky či kruhy mien. Chybu modelu v čase t zapíšeme ako e_t [4].

Logistický regresný model Nelineárna diskriminantná štatistická metóda. V **binary response** modeli os y zvyčajne reprezentuje individuálnu alebo experimentálnu jednotku. Y môže nadobúdať hodnoty 0 alebo 1 pre situácie kedy udalosť nastane alebo nenastane. Os x reprezentuje vysvetľujúcu veličinu ako vektor, ktorý môže znázorňovať pravdepodobnosť udalosti ($Y = 1$) [7].

1.2.2 Stochastické modely

Tieto metódy časových radov sú založené na predpoklade, že dáta majú vnútornú štruktúru, ako napr. autokoreláciu, trend či sezónnu variáciu. Najprv sa precízne zostaví vzor zodpovedajúci dostupným dátam a potom sa na jeho základe predpovie budúca hodnota veličiny [4].

Autoregresný model V autoregresívnom modeli je budúca hodnota premennej predpokladaná ako súčet lineárnej kombinácie p predchádzajúcich meraní, náhodnej chyby a konštanty. Matematicky môžeme autoregresný model zapísať ako

$$y_t = c + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t \quad (6)$$

Vo vzorci 6 hodnota y_t predstavuje predpovedanú hodnotu v čase t . Náhodnú chybu v čase t zapíšeme ako ε_t . Hodnoty φ_i sú parametre modelu a c je konštanta. Konštantou p označujeme rad modelu [1].

Model kľzavého priemeru Model kľzavého priemeru na rozdiel od autoregresného modelu používa ako vysvetľujúce premenné chyby predchádzajúcich meraní a nie priamo hodnoty. Matematicky môžeme tento vzťah zapísať ako

$$y_t = \mu + \sum_{j=1}^q \Theta_j \varepsilon_{t-j} + \varepsilon_t \quad (7)$$

Vo vzorci 7 hodnota y_t predstavuje strednú hodnotu postupnosti meraní v čase t . Hodnoty Θ_j sú parametre modelu a konštantou q označujeme rad modelu. Vychádzame z predpokladu, že náhodná zložka ε_t je biely šum, čo je rovnomerne distribuovaná náhodná premenná, ktorá má nulovú strednú hodnotu a konštantnú varianciu σ^2 [1].

Autoregressive Moving-Average model Model reprezentuje súčasnú hodnotu časového rádu lineárne na základe jeho hodnôt a hodnôt bieleho šumu v predchádzajúcich periódach [4].

Ide o kombináciu autoregresie (AR) a kľzavého priemeru (MA), vhodnú pre modelovanie jednorozmerných časových radov. Matematicky môžeme reprezentovať tento model ako súčet predchádzajúcich modelov

$$y_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i y_{t-i} + \sum_{j=1}^q \Theta_j \varepsilon_{t-j} \quad (8)$$

Rad modelu určuje p a q [1].

Autoregressive Integrated Moving-Average model Modely typu ARMA môžu byť použité iba na statické časové rady. Mnoho časových radov v praxi vykazuje nestatické správanie a tiež tie, ktoré obsahujú komponenty trendu a sezónnosti. Kvôli tomu bol navrhnutý model ARIMA, ktorý je generalizáciou modelu ARMA a zahŕňa tak v sebe aj prípady nestatických časových radov. Z nestatických časových radov sa vytvárajú statické pomocou konečného počtu derivovaní dátových bodov. Vzniká tak matematický model, ktorý môžeme zapísať ako

$$\left(1 - \sum_{i=1}^p \varphi_i L^i\right) (1 - L)^d y_t = \left(1 + \sum_{j=1}^q \Theta_j L^j\right) \varepsilon_t \quad (9)$$

Vzorec 9 môžeme zapísať aj jednoduchšie a to

$$\varphi(L)(1 - L)^d y_t = \Theta(L)\varepsilon_t \quad (10)$$

Vo vzorci 9 predstavujú premenné p , d a q rad autoregresného modelu, modelu kľzavého priemeru a integrovaného modelu. Hodnota d zodpovedá stupňu derivovania, zvyčajne je rovná 1. V prípade, že $d = 0$ dostaneme klasický ARMA model. Rovnakým spôsobom vieme dostať modely AR a MA [1].

1.2.3 Support vector regression

Support Vector Machine a Support Vector Regression sú založené na štatistickej teórii učenia, nazývanej aj VC teória, podľa svojich autorov, Vapnik a Chervonenkisa.

Support Vector Machine je použité na množstvo úloh strojového učenia ako je rozoznávanie vzorov, klasifikácia objektov a v prípade predikcií časových radov to je regresná analýza. Support Vector Regression je postup, ktorého funkcia je predpovedaná pomocou nameraných dát, ktorými je Support Vector Machine postupne natrénované. Toto je odklon od tradičných predpovedí časových radov, v zmysle že Support Vector Machine nepoužíva žiadny model, ale predikciu riadia samotné dáta [2].

Táto predikčná metóda nie je závislá na modely ani na žiadnych lineárnych procesoch. Tiež poskytuje malý počet voľných parametrov. Garantuje konvergenciu k ideálnemu riešeniu a môže byť výpočtovo efektívna [2].

1.2.4 Rozhodovacie stromy

Rozhodovacie stromy sú jednou z najrozšírenejších učiacich metód. Používajú sa najmä na klasifikáciu. Rozhodovací strom je reprezentovaný ako množina uzlov a im prislúchajúcich hrán. Uzly reprezentujú atribúty a výstupné hrany sú vždy označené konkrétnou hodnotou pre atribút, z ktorého vychádzajú. Rozhodovanie začína v koreni stromu a končí po dosiahnutí listového uzla. Pre riešenie jedného problému je možné vytvoriť stromy s rôznym počtom a usporiadaním uzlov. Najlepším riešením je strom s najmenším počtom rozhodovacích uzlov [8].

Regresný rozhodovací strom

1.2.5 Random forrest

1.2.6 Neurónové siete

Je veľa typov neurónových sietí napríklad viacvrstvové perceptrónové siete, samoriadiace siete, siete s viacerými skrytými vrstvami atď. V každej skrytej vrstve je množstvo neurónov. Hlavnou výhodou je, že väčšina sietí nepotrebuje model. Na druhej strane, tréning obvykle zaberá veľa času. Výstupom siete je lineárna rovnica váh prepojených so vstupom [4].

Najpoužívanejšou neurónovou sieťou je viacvrstvový perceptrón, ktorý pozostáva z uzlov a im prislúchajúcim hranám. Uzly sú zoskupované do rôznych vrstiev. Prvá vrstva je vstupná vrstva, kde počet d označuje počet vstupných parametrov vstupujúcich do siete. Táto vrstva je následne prepojená hranami so skrytou vrstvou pozostávajúcou z h uzlov. Tá je potom prepojená s výstupnou vrstvou s c uzlami. Kvôli tomu sa tieto siete zvyknú označovať aj ako dopredné siete [8].

Elementy skrytých a výstupných vrstiev sú umelé neuróny pozostávajúce z uzlov, viacerých vstupujúcich a jednou výstupnou hranou. Funkciou neurónu je transformovať lineárnu kombináciu vstupov pomocou nelineárnej aktivačnej funkcie, čiže každú vstupnú hranu prenásobiť jej váhou a výsledok týchto súčinou sčítať. Tak dostaneme pre neurón j vzorec 11 opisujúci lineárnu kombináciu vstupov a_j

$$a_j = \sum_{i=1}^d w_{ji} x_i \quad (11)$$

Príčom váha w_{ji} označuje váhu medzi neurónom i na vstupnej vrstve a neurónom j na skrytej vrstve [8].

Aktivovanie neurónu j závisí od jeho aktivačnej funkcie $g(a_j)$. Jednu z najpoužívanejších aktivačných funkcií, logistickú sigmoidnú funkciu, môžeme matematicky zapísať ako

$$g(a) \equiv \frac{1}{1 + \exp(-a)} \quad (12)$$

Je zrejmé, že funkcia zo vzorca 12 vracia hodnoty v rozmedzí $(0, 1)$ [8].

1.2.7 Učenie súborov klasifikátorov

Používa sa na jednoduchú predikciu. Ak h je počet meraní, ktoré sú denne dostupné, v deň t sa vykoná h predikcií podľa váženého priemeru m modelmi. Nasledujúci deň sa vypočíta chyba predpovede, na základe ktorej sa znova prepočítajú váhy a každý model sa aktualizuje [3].

Učenie súborov klasifikátorov môžeme rozdeliť na homogénne a heterogénne učenie.

Homogénne učenie súborov klasifikátorov Pozostáva z modelov rovnakého typu, ktoré sa učia na rôznych podmnožinách datasetu.

Heterogénne učenie súborov klasifikátorov Aplikuje rôzne typy modelov nad rovnakými dátovými množinami[3].

1.2.8 Exponencionálne hladenie

1.2.9 Naivné metódy

Predpovede sú vytvárané pomocou posledných hodnôt alebo ich priemerov.

Seasonal naïve method Poslednú nameranú hodnotu použijeme ako predpoveď pre nasledujúce obdobie. Ak sú naše dáta vysoko závislé od ročného obdobia, je lepšie použiť na predpoveď hodnotu z rovnakého obdobia, napr. z minulého roka [3].

Naïve average long-term method Predpokladá, že dáta obsahujú vzory, ktoré nie sú závislé od ročných období. Kvôli tomu sú časové rady lokálne stabilné s pomaly meniacim sa priemerom. Hodnotu, ktorú použijeme ako predpoveď je iba priemerom viacerých posledných hodnôt [3].

Naïve In median long-term method Táto metóda je alternatívou k predchádzajúcej metóde. Keďže priemerom nedokáže model dostatočne rýchlo reagovať na rapídne výkyvy a abnormality, lepšie výsledky dosiahneme nahradením priemeru za median posledných n meraní [3].

1.3 Analýza optimalizačných algoritmov

Genetické algoritmy sú biologicky inšpirované algoritmy patriace do triedy evolučných algoritmov. Genetické algoritmy sú stochastické optimalizačné algoritmy s **global search potential**. Od tradičných algoritmov sa líšia hlavne v počte riešení, ktoré sú kandidátmi na najlepšie riešenie. Tradičné vyhľadávacie algoritmy prehľadávajú dôkladne iba jedno riešenie, zatiaľ čo genetické algoritmy hlbšie spracujú viacero kandidátov naraz. Každý kandidát na optimálne riešenie problému je reprezentovaný dátovou štruktúrou, ktorú označujeme pojmom jedinec. Súbor jedincov tvorí populáciu. Začiatok procesu začína náhodnými riešeniami populácie, ktorý sa postupne vylepšuje [9].

Pri genetických algoritmoch sa zavádzajú pojmy ako chromozóm, fitness funkcia, kríženie, elitárstvo, operátor reprodukcie či mutácie [9].

Vytvorenie novej generácie prebieha pomocou genetického operátora, a to selekciou, krížením a mutáciou. Proces selekcie vyberie kvalitnejšie chromozómy, ktoré prežijú a vyskytnú sa tak aj v ďalšej generácii. Tiež sa vyberajú nekvalitné chromozómy, ktoré zahynú a nebudú ďalej brané do úvahy [10].

Chromozóm jedinec sa zvykne nazývať aj kandidát alebo chromozóm [11].

Fitness funkcia je funkcia určujúca efektívnosť chromozómu, pre ktoré je vypočítaná fitness funkcia. Pri hľadaní optimálneho riešenia sa porovnáva hodnota fitness funkcie aktuálneho riešenia s hodnotou funkcie cieľového riešenia [9, 10].

Operátor reprodukcie je obvykle prvý operátor, ktorý sa uplatní na populáciu. Operátor náhodne vyberie reťazce z dvoch chromozómov na párenie [9].

Kríženie je operátor **recombination**. Kríženie vykonáva výmenu blokov chromozónov. Z druhého chromozómu je vybraný reťazec náhodnej veľkosti, ktorý sa vymení s rovnako dlhým reťazcom z prvého chromozómu [9].

Vzniká problém ako, čo najvhodnejšie, určiť bod kríženia a vybrať tak najkvalitnejšie bloky chromozómu. Kvôli tomu je potrebné definovať nový element nazývaný väzba, označovaný ako d_i

$$d_i = \frac{a_i + a_{i+1}}{2} \quad (13)$$

Za predpokladu, že chromozón reprezentujeme ako maticu veľkosti $1 \times N$, potom väzbu definovanú vzorcom 13 môžeme reziprezentovať ako maticu $1 \times (N - 1)$. Vypočítaním priemeru susedných hodnôt aj pre cieľovú maticu a nájdením priemeru matice, určíme body kríženia. Chromozón rozdelíme na miestach, kde je väzba, čo najmenšia [10].

Elitárstvo je to roces pridávania chromozónov s najlepšou hodnotou funkcie fitness priamo do ďalšej populácie. Zaisťuje to, že najlepšie riešenie budúcej generácie bude vždy lepšie alebo pri najhoršom rovnaké ako najlepšie riešenie predchádzajúcej generácie [12].

Operátor mutácie sa vykoná po vykonaní operátora reprodukcie. Mutácia chromozómu invertuje jeden bit s nízkou pravdepodobnosťou [9].

Princíp fungovania genetických algoritmov možno znázorniť v nasledujúcom pseudokóde [9]

Algorithm 1 Pseudokód genetického algoritmu

- 1: Náhodne inicializovanie jedincov Inicializácia
 - 2: Vyhodnotenie fitness funkcie pre každého jedinca
 - 3: Výber jedincov pre ďalšiu populáciu na základe fitness funkcie
 - 4: Kríženie jedincov
 - 5: Mutovanie jedincov
 - 6: Kontrola či nebolo nájdené žiadane optimálne riešenie
 - 7: Ukončenie ak sa našlo takéto riešenie, inak opakovanie krokov 2 až 6
 - 8: Koniec
-

Veľkou výhodou genetických algoritmov je, že mutácia predchádza sklznutiu do lokálnych miním a rekombinácia chromozónov vedie k rýchlemu približovaniu sa k optimálnemu riešeniu. Napriek týmto výhodám, majú genetické algoritmy aj niekoľko nevýhod [12].

- Reprezentovanie kandidátov je príliš obmedzujúce
- Mutácia a kríženie sú v súčasnosti aplikovateľné iba na chromozóny reprezentované bitovým reťazcom alebo číslami
- Definovanie fitness funkcie je často netriviálnou záležitosťou a jej generalizácia je náročná

1.3.1 Artificial bee colony

ABC algoritmus je pomerne nový medzi rojovými algoritmami. Princíp je založený na biologickom procese, správaní medonosných včiel pri hľadaní potravy. Hlavný mechanizmus ktorým včely optimalizujú množstvo procesov je **waggle dance**, ktorým včely lokalizujú zdroje potravy a nachádzajú ďalšie [9].

Každá včela na pracujúca v roji sa spolupodíela na tvorbe celého systému na globálnej úrovni. Správanie systému je určené lokálnym správaním, kde spolupráca a zladenie jedincov vedie k štruktúrovanému kolaboračnému systému [9].

Algoritmus funguje na princípe, že včely nájdu najviac výnosný zdroj s použitím, čo najmenej energie. **Foragers** (zrejme robotnice hľadájúce zdroj jedla) uvažujú presúvanie sa medzi zdrojmi nektárov na základe kvality alebo zisku zdroju. Algoritmus poskytuje samo-manažovateľné a samo-organizované riešenie, vo svojej podstate decentralizované, pre daný problém [13].

1.3.2 Kolónia mravcov

Pri tomto algoritme, mravce tiež opúšťajú mravenisko, kvôli hľadaniu zdrojov potravy náhodne. Potom vyhodnotia kvalitu zdroja potravy a donesú ho naspäť do mraveniska. Zanechávajú pri tom na zemi chemické stopy. Sila týchto stôp závisí od kvality nájdeného zdroja potravy. Mnoho výskumov využíva tento algoritmus na riešenie NP problémov, ako napríklad problém obchodných cestujúcich, vyfarbovanie grafov, smerovanie áut alebo plánovacie problémy. Používa sa aj pri **cloud computing** na nájdenie optimálneho riešenia pri plánovaní úloh pre virtuálne servery [13].

Keď mravce hľadajú potravu prvý krát, hľadajú náhodne až kým nenájdu zdroj potravy. Zanechávajú pri tom za sebou chemickú stopu nazývanú feromón, ktorá tak vedie k zdroju. Tá následne priťahuje ostatné mravce k tomuto zdroju potravy. Tento proces pokračuje pokiaľ mravce nenájdu najkratšiu cestu vedúcu ku konkrétnemu zdroju potravy. Najkratšia cesta je určená naakumulovaným množstvom feromónov na ceste k zdroju potravy [13].

1.4 Meranie presnosti predpovedi

Pre vyhodnotenie efektívnosti a presnosti modelov je potrebné merať ich vlastnosti tak, aby sme ich vedeli medzi sebou porovnávať. V nasledujúcich spôsoboch merania sú použité pojmy ako aktuálna hodnota y_t , predpovedaná hodnota f_t alebo chyba predpovede e_t definovaná ako $e_t = y_t - f_t$. Veľkosť testovacej množiny budeme označovať ako n [1].

1.4.1 Stredná chyba predpovede

V anglickej literatúre označovaná ako MFE. Matematickú funkciu zapísať ako

$$MFE = \frac{1}{n} \sum_{t=1}^n e_t \quad (14)$$

Týmto spôsobom meriame priemernú odchýlku predpovedanej hodnoty od aktuálnej. Zistíme tak smer chyby nazývaný tiež **Forecast bias**. Nevýhodou je, že kladné a záporné chyby sa vynulujú a potom nie je možné zistiť presnú hodnotu chyby. Pri nameraní extrémnych chýb, nie sú nijak špeciálne penalizované. Taktiež hodnota chyby závisí od škály meraní a môže byť ovplyvnená aj transformáciami dát. Dobré predpovede majú hodnotu blízku 0 [1].

1.4.2 Stredná absolútna chyba

V anglickej literatúre označovaná ako MAE. Patrí k jedným z najpoužívanějších. Funkciu môžeme zapísať ako

$$MAE = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (15)$$

Týmto spôsobom meriame priemernú absolútnu odchýlku predpovedanej hodnoty od aktuálnej. Zistíme tak celkový rozsah chyby, ktorá nastala počas predpovede. Narozdiel od merania chyby

pomocou vzorca 14 sa kladné a záporné chyby nevynulujú, čo má však za následok, že nevieme určiť celkový smer chyby. Na druhej strane tiež nenastáva žiadna penalizácia pri extrémnych chybách, hodnota chyby závisí od škály meraní, môže byť ovplyvnená transformáciami dát a dobré predpovede majú hodnotu čo najbližšiu 0 [1, 14].

1.4.3 Stredná absolútna percentuálna chyba

V anglickej literatúre označovaná ako MAPE. Matematicky môžeme túto funkciu zapísať ako

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \times 100 \quad (16)$$

Pomocou tohto merania chyby získavame percentuálny prehľad o priemernej absolútnej chybe, ktorá sa vyskytla počas predpovedi. Veľkosť chyby nezávisí od škály merania, ale je závislá od transformácií dát. Tiež nie je možné zistiť smer chyby a ani nenastáva žiadna penalizácia pri extrémnych chybách [1].

1.4.4 Stredná percentuálna chyba

V anglickej literatúre označovaná ako MPE

1.4.5 Stredná štvorcová chyba

V anglickej literatúre označovaná ako MSE

2 Opis riešenia

3 Zhodnotenie

4 Technická dokumentácia

Literatúra

- [1] R. Adhikari, *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrücken: LAP LAMBERT Academic Publishing, 2013.
- [2] N. I. Sapankevych and R. Sankar, “Time series prediction using support vector machines: A survey,” *IEEE Computational Intelligence Magazine*, vol. 4, pp. 24–38, May 2009.
- [3] G. Grmanová, P. Laurinec, V. Rozinajová, A. Bou Ezzeddine, M. Lucká, P. Lacko, P. Vrablecová, and P. Návrát, “Incremental Ensemble Learning for Electricity Load Forecasting,” *Acta Polytechnica Hungarica*, vol. 13, no. 2, 2016.
- [4] A. Kumar Singh, S. Khatoon, M. Muazzam, and D. K. Chaturvedi, “An Overview of Electricity Demand Forecasting Techniques,” *Network and Complex Systems*, vol. 3, no. 3, 2013.
- [5] J. a. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, “Ensemble approaches for regression: A survey,” *ACM Comput. Surv.*, vol. 45, pp. 10:1–10:40, Dec. 2012.
- [6] L. Liu and G. Hudak, *Forecasting and Time Series Analysis Using the SCA Statistical System*, vol. zv. 1. Scientific computing Associates Corporation, 1992.
- [7] L. Sijia, T. Lan, Z. Yu, and Y. Xiuliang, “Comparison of the prediction effect between the logistic regressive model and svm model,” in *Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on*, pp. 316–318, Sept 2010.
- [8] C. J. Merz, *Classification and Regression by Combining Models*. PhD thesis, University of California, 1998. AAI9821450.
- [9] S. D. Chavan, A. V. Kulkarni, T. Khot, and N. Patil, “Bio inspired algorithm for disaster management,” in *2015 International Conference on Energy Systems and Applications*, pp. 776–781, Oct 2015.
- [10] S. Simonova and J. Panus, “Genetic algorithms for optimization of thematic regional clusters,” in *EUROCON 2007 - The International Conference on Computer as a Tool*, pp. 2061–2066, Sept 2007.
- [11] K. Arun and R. Rejimoan, “A survey on network path identification using bio inspired algorithms,” in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 387–389, Feb 2016.
- [12] R. V. Deolekar, “Defining parameters for examining effectiveness of genetic algorithm for optimization problems,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1061–1064, March 2016.
- [13] A. A. Buhussain, R. E. D. Grande, and A. Boukerche, “Performance analysis of bio-inspired scheduling algorithms for cloud environments,” in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 776–785, May 2016.
- [14] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, “Ordinal regression methods: Survey and experimental study,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, pp. 127–146, Jan 2016.