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 TECHNOLÓGIÍ
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

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 diskrétne. Pozorovania pri spojitých časových radoch sú merané v každej jednotke času, zatiaľ čo diskrétne 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 štvrťhodinu.

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 predopovedaní 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

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

Hodnota $\hat{x}$ je predpovedaná ako hodnota diskrétneho časového radu x. Preto je potrebné nájsť funkciu $f(x)$, podobnú funkciu $\hat{x}$, ktorá predpovedá hodnotu časového radu v budúcnosti konzistentne a objektvíne [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 dňov, týždňov alebo rokov. Veličina, ktorú sa snažíme predpovedať postupne mení svoje správanie a model sa tak stáva nepresným. Kvôli tomu je nutné v každom modely rozdeľovať tieto typy tendencií, aby sme vedeli model zmenám prispôsobiť [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 spôspobená nepredvídateľnými a nepravideľnými udalosťami [1].

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

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

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 predičkných algoritmov

Na základe množstva predikčných

1.2.1 Lineárna regresia

Najpoužívanejšia štatistická metóda, ktorá modeluje vzťah závislej premennej a vysvetľujúcej premmennej. Závislú premmenu predstavuje veličina, ktorú sa snažíme predpoved, 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 $\{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$. Úlohou regresie je odvodiť funkciu $\hat{f}$ z dát, kde

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

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

$$L(t) = Ln(t) + \sum a_i x_i(t) + e(t) \tag{4}$$

Vo vzorci 4 funkcia $Ln(t)$ 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 jednať 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 sezónných 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

$$Y(t) = V_t a_t + e_t \tag{5}$$

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 kruzy mien. Chybu modelu v čase t zapíšeme ako e_t [4].

Logistický regresný model Nelineárna diskriminantná štatistická metóda. V **binary response** modely 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 modely je budúca hodnota premennej predpokladaná ako súčet lineárnej kombinácie p predchdzajúcich meraní, náhodnej chyby a konštanty. Matematicky môžeme autoregresný model zapisať ako

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

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

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

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 linárne na základe jeho hodnôt a hodnôt bieleho šumu v predchádzajúcich periódach [4].

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

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

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 rádov 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 rádov. Z nestatických časových radov sa vytvárajú statické pomocou konečného počtu derivovaní dátových bodov. Vzniká tak matematický model, ktorý môžeme zapísať ako

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

Vzorec 9 môžeme zapísať aj jednoduchšie a to

$$\varphi(L)(1 - L)^d y_t = \Theta(L)\varepsilon_t \tag{10}$$

Vo vzorci 9 predstavujú premenné p, d a q rad autoregresného modelu, modelu kĺzavého 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 teórií 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 ja 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 perceptónové 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, trénovanie obvykle zaberá veľa času. Výstupom siete je lineárna rovnica váh prepojených so vstupom [4].

1.2.7 Učenie súborov klasifikátorov

Používa sa na jednodňovú 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 závisle 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 priemorom viacerých posledných hodnôt [3].

Naïve In median long-term method Táto metóda je alternativou 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

1.4 Meranie presnosti predpovedi

Pre vyhodnotenie efektívnosti a presnosti modelov je potrebné merať ich vlastnosti tak, aby sme ich vedeli medzi sebou porovnáva. 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

$$MFE = \frac{1}{n}\sum_{t=1}^{n} e_t \tag{11}$$

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 namerani 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 najpoužívanejších. Funkciu môžeme zapísať ako

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{12}$$

Týmto spôsobom meriame priemernú absolútnu odchýlku predpovedanej hodnoty od aktuálne. Zistíme tak celkový rozsah chyby, ktorá nastala počas predpovede. Narozdiel od merania chyby pomocou vzorca 11 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 ovplynená transformáciami dát a dobré predpovede majú hodnotu čo najbližšiu 0 [1, 9].

1.4.3 Stredná aboslútna percentuálna chyba

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

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \times 100 \tag{13}$$

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*. Saarbrucken: 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. Návrat, "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] 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.