

**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

<b>1</b>	<b>Analýza problému</b>	<b>6</b>
1.1	Časové rady . . . . .	6
1.1.1	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 . . . . .	8
1.2.4	Rozhodovacie stromy . . . . .	8
1.2.5	Random forrest . . . . .	8
1.2.6	Neurónové siete . . . . .	8
1.2.7	Učenie súborov klasifikátorov . . . . .	8
1.2.8	Exponencionálne hľadanie . . . . .	9
1.2.9	Naivné metódy . . . . .	9
1.3	Analýza optimalizačných algoritmov . . . . .	9
1.4	Meranie presnosti predpovedi . . . . .	9
<b>2</b>	<b>Opis riešenia</b>	<b>10</b>
<b>3</b>	<b>Zhodnotenie</b>	<b>11</b>
<b>4</b>	<b>Technická dokumentácia</b>	<b>12</b>

# 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 diskkrétne. Pozorovania pri spojitých časových radoch sú merané v každej jednotke času, zatiaľ čo diskkrétne 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 štvrtú hodinu.

Cieľom predikcií časových radov je predpovedať hodnotu premennej v budúcnosti na základe doteraz nameraných dátových vzoriek. Preto je potrebné nájsť funkciu, 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  $x$ .

### 1.1.1 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 rozdeľovať 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 nepravidelnými udalosťami [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

$$\hat{f} : X \rightarrow \mathbb{R}, \text{ kde, } \hat{f}(x) = f(x), \forall x \in X, \quad (1)$$

Funkcia  $f$  vo vzorci 1 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

$$L(t) = Ln(t) + \sum a_i x_i(t) + e(t) \quad (2)$$

Vo vzorci 2 funkcia  $Ln(t)$  predstavuje odber elektriky v čase  $t$ . Hodnota  $a_i$  je odhadovaný pomaly meniaci sa koeficient. Faktory  $x_i(t)$  nezávisle vplyvajú 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 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

$$Y(t) = V_t a_t + e_t \quad (3)$$

Vo vzorci 3  $t$  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].

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

### 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].

### 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].

#### 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 1.2.9. Keď že 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 predpovedí

## **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.