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. V našom prípade sú namerané dáta dostupné každú celú štvrťhodinu [1].

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 objektvíne [2].

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

1.2 Komponenty časových radov

Vo všeobecnosti sú časové rady ovplyvnené 4 hlavnými komponentmi, ktoré môžu byť oddelené od pozorovaných dát. Jedná sa o trendovú, cyklickú, sezónnu a náhodnú zložku [1].

1.2.1 Trendová zložka

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

1.2.2 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 napr. podnikateľský cyklus pozostávajúci zo 4 fáz, ktoré sa stále opakujú [1].

1.2.3 Sezónna zložka

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

1.2.4 Náhodná 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. Je spôspobená nepredvídateľnými a nepravideľnými udalosťami [1].

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].

1.3 Analýza predičkných algoritmov

Na základe množstva predikčných

1.4 Regresia

Predpokladajme typický regresný problém. Dáta pozostávajúce z množiny n meraní vo formáte **vloz vzorec**. Úlohou je odvodiť funkciu **f** z dát, kde **vloz vzorec** kde f reprezentuje reálnu neznámu funkciu. Algoritmus použitý na odvodenie funkcie **f** sa nazýva indukčný algoritmus alebo žiak. Funkcia **f** sa nazýva model alebo prediktor. Obvykle je úlohou regresie minimalizovať odchýlku funkcie pre **štvorcovú chybu**, konkrétne priemernú štvorcovú chybu **mse** [4].

1.4.1 Lineárna regresná analýza

Regresná analýza je štatistická metóda používaná na modelovanie vzťahov, ktoré môžu existovať medzi veličinami. Nachádza súvislosti medzi závislou veličinou a potenciálnymi nezávislými veličinami. Používame pri tom nezávislé veličiny, ktoré môžu byť namerané súčasne so závislými veličinami alebo aj veličiny z úplne iných zdrojov. Regresná analýza môže byť tiež použitá na zlúčenie trendu a sezónných indikátorov do modelu. Keď je raz model vytvorený, môže byť použitý na zásah do spomínaných vzťahov alebo, v prípade dostupnosti nezávislých veličín, na vytvorenie predikcie [5].

1.4.2 Viacnásobná model

Viacnásobná regresia sa pokúša modelovať vzťah medzi dvoma alebo viacerými nezávislými premennými a závislou premennou vhodnou lineárnou rovnicou pre pozorované dáta. Výsledný model je vyjadrený ako funkcia viacerých nezávislých premenných a predpoveď ou nie je priamka ako je to pri lineárnej regresii [3]. Nezávislé premenné môžu predstavovať meteorologické vplyvy, ekonomický rast, ceny elektriky či kruzy mien [6].

1.4.3 Logistic regression 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 nezávislú veličinu ako vektor, ktorý môže znázorňovať pravdepodobnosť udalosti (Y = 1) [7].

1.5 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].

1.6 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].

1.6.1 Regresný rozhodovací strom

1.7 Random forrest

2 Opis riešenia

3 Zhodnotenie

4 Technická dokumentácia

Literatúra

[1] R. R. A. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," *arXiv preprint arXiv:1302.6613*, vol. 1302.6613, pp. 1–68, 2013.

[2] N. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24–38, 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] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression," *ACM Computing Surveys*, vol. 45, no. 1, pp. 1–40, 2012.

[5] L.-M. Liu, G. B. Hudak, G. E. P. Box, M. E. Muller, and G. C. Tiao, "Forecasting and time series analysis using the SCA statistical system," *NAJDI JOURNAL*, vol. 1, p. 1992, 1992.

[6] A. Kumar Singh, S. Khatoon, M. Muazzam, and D. K. Chaturvedi, "An Overview of Electricity Demand Forecasting Techniques," *NAJDI JOURNAL*, vol. 3, no. 3, 2013.

[7] S. Li, L. Tan, Z. Yu, and X. Yu, "Comparison of the prediction effect between the Logistic Regressive model and SVM model," *Proceedings - 2010 2nd IEEE International Conference on Information and Financial Engineering, ICIFE 2010*, pp. 316–318, 2010.

[8] C. J. Merz, *Classification and regression by combining models*. PhD thesis, University of California, 1998.