An Improved Algorithm for Data Filtering Based on Variation for Short Term Air Pollution Prediction in Macau

by

YANG Jing Yi

Master of Science of Software Engineering

2012

Faculty of Science and Technology
University of Macau

An Improved Algorithm for Data Filtering Based on Variation for Short Term Air Pollution Prediction in Macau

by

YANG Jing Yi

(Student No: M-A9-6524-1)

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science of Software Engineering

Faculty of Science and Technology
University of Macau

2012

Approved by _____
Supervisor

_____
Co-Supervisor

Date _____

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Macau, I agree that the Library and the Faculty of Science and Technology shall make its copies freely available for inspection. However, reproduction of this thesis for any purposes or by any means shall not be allowed without my written permission. Authorization is sought by contacting the author at

Telephone: 00853-62169828
E-mail: chris.jyyang@gmail.com

Signature _____

Date _____

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

University of Macau

Abstract

AN IMPROVED ALGORITHM FOR DATA FILTERING BASED ON VARIATION FOR SHORT TERM AIR POLLUTION PREDICTION IN MACAU

by YANG Jing Yi

Thesis Supervisor:

Dr. VONG Chi Man and Dr. IP Weng Fai

Accurate prediction models for air pollution are crucial for forecast and health alarm to local inhabitants. Discrete wavelet transform (DWT) was employed to decompose a series of air pollutant levels, followed by modeling using support vector machines (SVMs) in recent literature. This combination of techniques was reported to produce a more accurate prediction model for air pollutants by investigating different levels of variations. However, DWT demands significant model complexity, namely, the training time and the model size of the prediction model. In this study, a new method called variation-oriented filtering (VF) was proposed to remove the data with low variation which can be considered as noise. Using VF, the size of training data and the number of support vector can be reduced, which can reduce training time, computational complex and model size of prediction model; hence the efficiency of prediction model can be improved. Simultaneously, the noise in the series of air pollutant levels can be filtering by using VF, which can improve the accuracy of prediction model. The SPM (suspended particulate matter) level in Macau was selected as a test case. Experimental results showed that VF can effectively and efficiently reduce the model complexity with improvement in predictive accuracy, as compared with DWT.

Keywords: Air pollution prediction; Data filtering; Support vector machines; Discrete wavelet transform

TABLE OF CONTENTS

LIST OF FIGURES iii
LIST OF TABLES iv
LIST OF ABBREVIATIONS vi
CHAPTER 1: Introduction 1
CHAPTER 2: Methods 7
2.1 Algorithm of Variation-oriented Filtering (VF) 7
2.2 Discrete Wavelet Transform (DWT) 9
2.3 Support Vector Machines (SVMs) 11
2.4 Grid Search (GS) 16
2.5 Genetic Algorithm (GA) 17
CHAPTER 3: Materials and Data Preprocessing 18
3.1 The Environment and Monitoring Sites in Macau 18
3.2 Data Sampling 20
3.3 Data Normalization 20
3.4 Input Variable Selection 21
3.5 Data Grouping 23
3.6 Data Representation 24
CHAPTER 4: Application 26
4.1 Construction of Prediction Models 26
4.2 Kernel Selection for Tested Models 26
4.3 Workflow of Modeling for *SVM* Model 27
4.4 Workflow of Modeling for *SVM-DWT* Model 28
4.4.1 Discrete Wavelet Transform Setup for S*VM-DWT* Model 29
4.5 Workflow of Modeling for *SVM-VF* Model 31
4.5.1 Genetic Algorithm Setup for S*VM-VF* Model 32
CHAPTER 5: Results and Discussion 38
5.1 Evaluation Criterion 38
5.1.1 Mean Absolute Error 38

5.1.2 Root Mean Squared Error 38
5.1.3 Complementary Willmott's Index of Agreement 39
5.1.4 Standard Deviation 40
5.2 Accuracy and Stability of Tested SVMs Models 40
5.3 Seasonal Variation of SPM levels 43
5.4 Model Complexity 47
5.5 Discussion 49
CHAPTER 6: Conclusion 51
6.1 Conclusion 51
6.2 Future Work 51
BIBLIOGRAPHY 53
APPENDIX: Publication 60

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

LIST OF FIGURES

Number *Page*

Figure 1: Variation of SPM Levels in Macau 6
Figure 2: An Example of Variation-oriented Filtering 8
Figure 3: A Multi-resolution Analysis Tree with Three Levels 11
Figure 4: The Principle of Support Vector Machines 12
Figure 5: The Intensive Band $\pm\varepsilon$ and Slack Variables ξ_i and ξ_i^* for ε-SVR 13
Figure 6: The Linear ε-insensitive Loss Function of ε-SVR 14
Figure 7: The location of Macau and The Meteorological Stations 19
Figure 8: Pearson Correlation Coefficient between SPM and Relative Input Variable Candidates in Different Time Periods 24
Figure 9: Data Representation for SVMs Modeling 25
Figure 10: The Workflow of Modeling for *SVM* Model 28
Figure 11: The Workflow of Modeling for *SVM-DWT* Model 29
Figure 12: The Daubechies Wavelets Family 30
Figure 13: The Workflow of Modeling for *SVM-VF* Model 32
Figure 14: The General Workflow of Genetic Algorithm 33
Figure 15: An Example of An Individual Used in Genetic Algorithm 33
Figure 16: An Example of Crossover Used in Genetic Algorithm 34
Figure 17: An Example of Mutation Used in Genetic Algorithm 35
Figure 18: The Specific Workflow of Genetic Algorithm 36
Figure 19: Error Measurements of Tested SVMs Models 42
Figure 20: The Comparison between Predicted and Observed SPM Levels (Year 2006 to Year 2008) 44
Figure 21: The Comparison between Predicted and Observed SPM Levels in Winter (January, year 2006 to year 2008) 45
Figure 22: The Comparison between Predicted and Observed SPM Levels in Summer (July, year 2006 to year 2008) 46
Figure 23: The Size of *TRAIN* of Three Tested SVMs Models 47

Figure 24: The Number of Support Vectors of Three Tested SVMs Models.....48
Figure 25: The Training Time of Three Tested SVMs Models.....48

LIST OF TABLES

Number *Page*

Table 1: Pearson Correlation Coefficients between The Output Variables SPM(d+1) and Input Variables Candidates on Previous Days in The Time Period from Year 2003 to Year 2008 23

Table 2: Grouping of Experimental Data.....24

Table 3: Combinations of Tested SVMs Models.....26

Table 4: CWIA for SVMs Model with Different Kernel Function 27

Table 5: Optimized Hyperparameters c^* and g^* for *SVM* Model in 10 Runs 27

Table 6: Optimized Hyperparameters c^* and g^* for *SVM-DWT* Model in 10 Runs.....29

Table 7: CWIA for *SVM-DWT* Model with Different Mother Wavelet and Decomposition Levels for Different Experimental Data 31

Table 8: Optimized Hyperparameters δ^*, c^* and g^* for *SVM-VF* Model in 10 Runs 37

Table 9: Error Measurements of Tested SVMs Models 41

LIST OF ABBREVIATIONS

ARMA: autoregressive moving average

ANNs: artificial neural networks

AP: atmospheric pressure

CWIA: complementary Willmott's index of agreement

db: Daubechies wavelet

DSMG: Macau government meteorological center

DWT: discrete wavelet transform

FAT: Fourier analysis technique

FFT: fast Fourier transform

GA: genetic algorithm

GS: grid search

HP: high-pass

KKT: Karush-Kuhn-Tucker

LP: low-pass

MAE: mean absolute error

MLP: multilayer perceptron

MRH: mean relative humidity

NO_2: nitrogen dioxide

O_3: ozone

PCA: principle components analysis

PM: particulate matter

PSO: particle swarm optimization

RBF: radial basis functions

RF: rainfall

RMSE: root mean squared error

SHr: sunshine hour

SO_2: sulfur dioxide

SOM: self-organizing maps

SPM: suspended particulate matters

SV: support vector

SVMs: support vector machines

TEMP: temperature

TEST: test data set

TRAIN: training data set

VALID: validation data set

VF: variation-oriented filtering

WMAT: wavelet-based multi-resolution analysis technique

WD: wind direction

WHO: World Health Organization

WIA: Willmott's index of agreement

WS: wind speed

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude and appreciation to my supervisors, Dr. VONG Chi Man and Dr. IP Weng Fai for their enthusiastic guidance, inspiration and insistent support throughout the course of my research. Without their patient instruction, I cannot complete my thesis successfully.

I would like to thank the committee members of my thesis defense, Prof. PUN Chi Man and Prof. WONG Pak Kin, for their attendance and giving me constructive and valuable comments.

Simultaneously, I would like to give my thanks to University of Macau for offering me the studentship.

In addition, special thanks are given to the teachers, Prof. GONG Zhi Guo, Prof. LI Yi Ping, Dr. SI Yain Whar, Prof. TAM Kam Weng, Dr. CHOI Wai Wa, and many others who have gave encouragement and support. I would also like to give thanks to my good friends in University of Macau, WU Yi, ZHAI Yu, SHU Hong, HUANG Jing Ye, LI Sheng, YU Si Ping, SUN Wen Bo, MA Ting, YAN Wei, who have made my life colorful.

Finally, I would like to express the deeply thanks to my family, my boyfriend XU Shao Cai and his family for their love and endless support.