Neurocomputing 128 (2014) 136–144

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Predicting minority class for suspended particulate matters level by extreme learning machine

Chi-Man Vong [a,*], Weng-Fai Ip [b], Pak-Kin Wong [c], Chi-Chong Chiu [a]

[a] Department of Computer and Information Science, University of Macau, Macau, China
[b] Faculty of Science and Technology, University of Macau, Macau, China
[c] Department of Electromechanical Engineering, University of Macau, Macau, China

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form
21 November 2012
Accepted 25 November 2012
Available online 25 October 2013

Keywords:
PM_{10}
Extreme learning machine (ELM)
Support vector machine (SVM)
Imbalance problem
Prior duplication

ABSTRACT

Suspended particulate matters (PM_{10}) is considered as a harmful air pollutant. Many models attempt to predict numerical levels of PM_{10} but a simple, clearly defined classification of PM_{10} levels is more readily comprehensible to the general public rather than a numerical value. However, the PM_{10} prediction model often suffers from data imbalance problem in the training dataset that results in failure to forecast the minority class of severe cases. In this study, a warning system using extreme learning machine (ELM), compared with support vector machine (SVM), was constructed to forecast the class of PM_{10} level: *Good*, *Moderate*, and *Severe*. An imbalance strategy called prior duplication was also applied to improve the forecast of minority class. The experimental comparisons between ELM and SVM demonstrate that ELM produces superior accuracy relative to SVM in forecasting minority class (*Severe*) of PM_{10} level with or without the imbalance strategy. Furthermore, our results show that the required training time and model size in the ELM model are much shorter and smaller than those of SVM respectively, leading to a more efficient and practical implementation of prediction model for large dataset. The performance superiority of ELM is also discussed in this paper.

1. Introduction

One serious concern about prolong exposure at high levels of air pollution is the potential to cause respiratory diseases. Apart from the various reduction regulations and technological advancement, an early warning system also serves as a preventive measure. Compared to the forecasting of gaseous pollutants, such as SO_2 and NO_x, modeling suspended particulate matters (PM_{10}) concentration was considered more challenging because of the complexity of the processes on their formation, transportation and removal of aerosol in the atmosphere [1]. Artificial neural networks (ANNs) have been used as a cost effective and dependable method for pollution level forecast at various time scales with very good results [1–11]. The findings of numerous works reported that the performance of ANNs is generally superior in comparison to traditional statistical methods, such as multiple regression, classification and regression trees. The multi-layer perceptron (MLP), a particular kind of ANN, can be trained to approximate virtually any smooth, measurable function [3,5,6]. These studies have shown that the ANN approach is effective in modeling the dynamics of non-stationary time series owing to its non-parametric, non-assumable, and high-adaptive properties enabling it to capture the highly non-linear character of those processes.

In the last decade, a kind of ANN called support vector machine (SVM) was reported to be an effective approach to improve generalization performance of MLP and achieve global solutions simultaneously. Based on structural risk minimization principle, SVM attempts to minimize an upper bound of generalization error rather than to minimize training error. Although good forecasting performance using SVM was reported in many applications including air pollutants forecasting [7], similar to conventional ANN approaches, SVM suffers from the data imbalance problem that has an "undo" effect to minority class [12]. For example, when there are 80% of training cases belonging to class 1 while 5% to class 3, the prediction model is likely dominated by the majority class, and subsequently always fails to predict the minority class. This is a general problem in the application of statistical approach, leading to worse generalization in the minority class [9]. Furthermore, most of these approaches target to provide a numerical forecast of the pollutant levels. However, a simple, clearly defined classification of pollutant levels is more readily comprehensible to the general public rather than a numerical value. Therefore the traditional numerical forecasting problem is transformed as a multiclass classification problem in this study.

Recently, extreme learning machine (ELM) [13], another kind of ANN, has attracted the attention of many researchers in different

* Corresponding author. Tel.: +86 3839 74357; fax: +86 3288 38314.
E-mail address: cmvong@umac.mo (C.-M. Vong).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.056

applications [14–19]. ELM is an improved MLP using a more efficient training algorithm. The advantage of ELM is obvious in shorter training time and in compact model size (i.e., computer memory to store the trained model) while the generalization of ELM is comparable to that of SVM. Huang [20] indicated that ELM tends to have better generalization performance for multiclass problems and the experimental results showed that ELM outperforms SVM in the case of multiclass classification. In addition, a similar data imbalance problem of cancer classification was studied using ELM [17]. The experimental results showed that ELM has higher generalization than SVM for treating data imbalance problem.

The generalization for data imbalance problem can be further improved by applying an imbalance strategy which can be categorized into data-based strategy (DS) [21–23] or algorithm-based strategy (AS) [24–26]. DS is also known as sampling strategy that modifies the size of the training dataset without modification of the classification algorithm. Thus, DS can be easily applied to any classification algorithm. Conversely, AS involves modification of the cost-sensitive information in a classification algorithm to account for data imbalance. However, AS suffers from several drawbacks: (i) it is difficult to modify the classification algorithm and precautious measure must be taken or else classification deterioration may be incurred; (ii) different cost-sensitive information is necessary for different applications; (iii) almost all ASs were developed only for binary classification because it is nontrivial to determine the cost-sensitive information among multiple classes [27]. Therefore, AS is difficult to operate for multiclass classification. From this viewpoint, a simple strategy called *prior duplication* [12,28], one approach in DS, was employed as a test case in this study for minority class prediction under multiple classes. Further investigation on other data imbalance strategies for PM_{10} forecasting will be left as a future work. In this study, ELM along with prior duplication was therefore applied to forecast the minority class of PM_{10} level. The performances of ELM (with or without prior duplication) in different aspects were evaluated by comparing the results with SVM. Finally, the performance superiority of ELM over SVM in predicting minority class is also discussed in this paper.

The subsequent sections are organized as follows: a brief introduction of ELM and prior duplication is given in Section 2, followed by the sampling of the experiment in Section 3. The experiment details are presented in Section 4. The results and discussion are given in Section 5. Finally, a conclusion is drawn in Section 6.

2. Employed techniques

2.1. Extreme learning machine (ELM)

Extreme learning machine (ELM) proposed by Huang [13] is a learning method for generalized single-hidden layer feedforward neural networks (SLFNs) where all the hidden node parameters are randomly generated and the output weights of SLFNs are analytically determined. Different from traditional learning algorithms for ANN, ELM not only tends to reach the smallest training error but also obtain the smallest norm of output weights. This leads to the relatively better generalization as shown in Bartlett's theory [29]. Furthermore, ELM tends to provide good generalization at extremely fast learning speed because of its simple and efficient learning algorithm in which iterative tuning is not required in the hidden layers. This is a significant advantage over conventional MLP approach.

Given a set of N training data $D=(\mathbf{x}_i, t_i)$, $i=1$ to N. Assume sigmoid activation function is used for hidden nodes, the classification of ELM with L hidden nodes is

$$f_{ELM}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{L} \beta_i G(\mathbf{w}_i \mathbf{x} + b_i)\right) \tag{1}$$

$$G(y) = \frac{1}{1+\exp(-y)} \tag{2}$$

where $\mathbf{w}_i$ and b_i are the learning parameters of hidden nodes and β_i is the output weight connecting the i-th hidden node and the output nodes. If an ELM with L hidden nodes can approximate N training data with zero error, it implies that there exist β_i, $\mathbf{w}_i$ and b_i such that

$$f_{ELM}(\mathbf{x}_j) = \sum_{i=1}^{L} \beta_i G(\mathbf{w}_i \mathbf{x_j} + b_i) = t_j, j = 1, \ldots, N \tag{3}$$

Then Eq. (3) can be written compactly as

$$\mathbf{H\beta} = \mathbf{T} \tag{4}$$

where $\mathbf{H}$ is the hidden layer output matrix denoted by

$$\mathbf{H} = \begin{bmatrix} G(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & G(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L) \\ \vdots & \vdots & \vdots \\ G(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & G(\mathbf{w}_L \cdot \mathbf{x}_N + b_L) \end{bmatrix} \tag{5}$$

$$\boldsymbol{\beta} = [\beta_1 \quad \cdots \quad \beta_L]^T \text{ and } \mathbf{T} = [t_1 \quad \cdots \quad t_L]^T \tag{6}$$

The i-th column of $\mathbf{H}$ is the i-th hidden node's output vector with respect to inputs $\mathbf{x}_1$, $\mathbf{x}_2$,…, $\mathbf{x}_N$ and the j-th row of $\mathbf{H}$ is the output vector of the hidden layer with respect to input $\mathbf{x}_j$. Eq. (4) is a linear system and the output weight vector $\boldsymbol{\beta}$ can be calculated by

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{7}$$

where $\mathbf{H}^{\dagger}$ is the Moore–Penrose generalized inverse of the hidden layer output matrix $\mathbf{H}$. thus, the ELM algorithm can be summarized as follows:

1. Randomly generate input weight $\mathbf{w}_i$ and bias b_i, $i=1, \ldots, L$.
2. Calculate the hidden layer output matrix $\mathbf{H}^{\dagger}$.
3. Calculate the output weight vector $\boldsymbol{\beta}$. Using known $\boldsymbol{\beta}$, $\mathbf{w}_i$ and b_i, Eq. (1) can be easily calculated.

ELM has a unified framework [13,20] for multiclass classification where K classes simply require K corresponding output nodes. For multiclass classifications, the predicted class of a given test case is closest to the output node of ELM classifier.

2.2. Prior duplication

In this study, prior duplication [12] from DS was employed because it can easily suit multiclass classification. This method simply modifies the size of training dataset by duplicating the data of minority classes by a fixed number of times. After the duplication, the new training dataset has about the same number of data examples in all classes. Further operation and setup of prior duplication is shown in Section 4.3.

3. Application

In this section, the environment and monitoring sites in Macau are briefly introduced. Data sampling, data preprocessing, and variable selection in this study are also described. Although there are many kinds of air pollutants, for illustration purpose, only the prediction of PM_{10} level is selected. The predictions of other air pollutant levels can be conducted in a similar way.