ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 76
7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)

Prediction of Airborne Particulate Matter PM 10 Based on Main Curve

Liyun Wang

College of Computer and Information Engineering, Zhengzhou University of Industrial Technology, Zhengzhou Henan 451150,China

912725921@qq.com

Keywords: The principal curve; Suspended particulate matter PM10; Imbalance; Threshold; Model

Abstract. It's very practical significance to predict the density of hazardous substance(such as PM10) in the air, However, in most cases, this kind of data have the characteristics of imbalance and sequential arrived online, It's difficult to realize rapid and effective prediction by traditional supervised learning methods. In order to solve this problem, a PM10 prediction method which based on the principal curve, build a PM10 model of PM10 from 2010 to 2012, received corresponding parameters by fitting, Finally, the main curve is obtained. corresponding threshold values of different density of PM10 respectively by a lot of experiment. The results show that the PM10 prediction model based on the principal curve predicts rapidly and a lower prediction error, meanwhile, the network structure is more compact.

Introduction

In recent years, environmental degradation caused by the concentration of harmful substances in the air increased, the physical health of urban residents have a significant impact. Predicting the concentration of contaminants will help to prepare precautions in advance and alleviate the symptoms of the disease, with significant practical significance. Therefore, this study is the hotspot of current air quality assessment.

At present, a variety of improved algorithms are proposed for the classification of unbalanced data. The improvement direction can be divided into two categories, one is from the perspective of the data set, the other is from the perspective of the algorithm. The processing method of data plane [1] mainly includes over-sampling strategy and under-sampling strategy, which improves the unbalanced data set through some mechanism in order to obtain a balanced data distribution. easily lead to the loss of important sample information. The algorithmic angle [2] mainly includes changing the probability density, single class learning classification, cost-sensitive learning and kernel method of integration algorithm. However, these methods are analyzed for single-core case. However, using single-core mapping method, It is not reasonable to proceed with the process. In addition, in the practical application of all the data is not a visit, and requires online prediction; in order to solve this problem liang [3] and others introduced the ELM algorithm, and put forward the online sequence ELM algorithm.

Principal Curve

Hastie and Stuetzle proposed the concept of the main curve in 1984 [4], the main curve is to find a geometric and intuitive, theoretically complete, the basic idea is to find through the middle of the data distribution, and to meet the self-consistent characteristics of the smooth curve.

The main curve of the algorithm steps:

Step 1: Make the initial curve $f^{(0)}(\lambda)$ is X of the first principal component line, suppose $j=0$;

Step 2:(Projection step) for all $x \in R^d$, request $\lambda_{f^{(j)}}(x)=\max\{t:\|x-f(\lambda)\|=\min_{\tau}\|x-f(\tau)\|\}$;

Step 3: (Expected step) definition $f^{(j+1)}(\lambda)=E\left[X\middle|\lambda_{f^{(j)}}(X=\lambda)\right]$;

Data Collection

Macau is located just 26.8 square miles of land on the southern coast of China [5], by three land areas: Macau Peninsula, Taipa Island and Coloane Island. Similar to many of China's coastal cities. Macau has experienced rapid urban development over the past 10 years, and air pollution has become an important issue. The Macau Meteorological Center has established GM and roadside weather stations in these three areas to collect airborne PM10 data. As the land area of Macau is relatively small, therefore, the air pollution PM10 data used in the Taipa Grande area data.

Input Variable Data is Standardized

Given a training data set $D = (x, y)$ x is the vector of the input variable [6], y is the vector of the output variable. The preprocessing process is described as follows: The required data range is required so that the input variable value x will not control the PM10 forecast level.

Variable Selection

The simple input variables are kept so that the model does not produce the noise due to the noise generated by the training set. Therefore, the daily PM10 content of the suspended particulate matter in the air is measured at 365 days per year [7]. Table 1 is part of the Taipa Grande 2010, 2011, 2012 collected part of the PM10 data.

Table 1 the Taipa Grande region 2010 to 2012 collected PM10 data

2010 data		2011 data		2012 data	
time/day	PM10	time/day	PM10	time/day	PM10
1	72.8	1	57.3	1	61.8
2	75.6	2	46.3	2	80.5
3	60.8	3	148.1	3	128.3
4	66.1	4	97	4	100.2
5	64.9	5	88.4	5	153.2
6	55.1	6	97.5	6	80.7
7	109.2	7	132.1	7	116.7
8	122.7	8	106.7	8	84.8
9	128.5	9	129.9	9	82
10	177.8	10	170.6	12	30.1
11	156.9	11	184.8	13	106.7
12	43.8	12	169.9	14	78.7
13	56.7	13	144.2	15	84.6
14	48.6	14	130.3	16	67
15	63.2	32	174	34	71.9

Simulation

According to the 2010, 2011, 2012 data for the following main curve, the time for the abscissa, the daily air PM10 content for the vertical as shown in Fig. 1,2, 3.

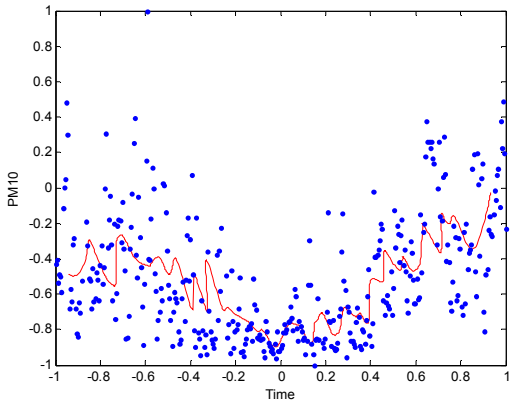

Figure 1. Graph for 2010 data

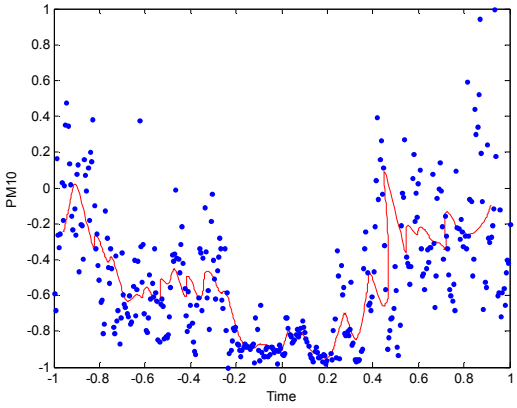

Figure 2. Graph for 2011 data

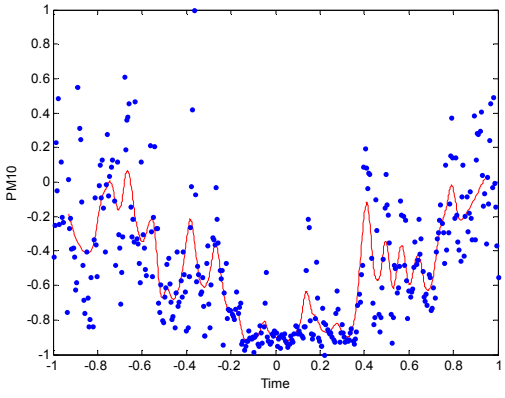

Figure 3. Graph for 2012 data

According to the 2010 to 2012, the main curve of the curve to roughly similar, it can be inferred that the annual air PM10 material content is approximately a parabolic graphics, the first step: greater than the value of this threshold to delete, less than or equal to the value of this threshold to retain. The second step: the reserved threshold value of the anti-normalization [8], delete the same number of days after the anti-normal and PM10 value, to avoid the use of duplicate data. The third step: according to delete the remaining data were drawn after each year after the anti-normalized graphics, and come to the corresponding function expression. Fig.4 shows the main curve y = 0.0035 * x ^ 2-1.2908 * x + 172.0826, respectively, after the backlog of data in 2010. Fig.5 is the main curve y = 0.004 * x ^ 2-1.3843 * x + 167.3183 (the values of the first third order coefficients are 0.0040, -1.3843, 167.3183) plotted after the regression data in 2011. Fig.6 shows the main curve

y = 0.0035 * x ^ 2-1.2908 * x + 172.0826 drawn from the residual data after 2012.

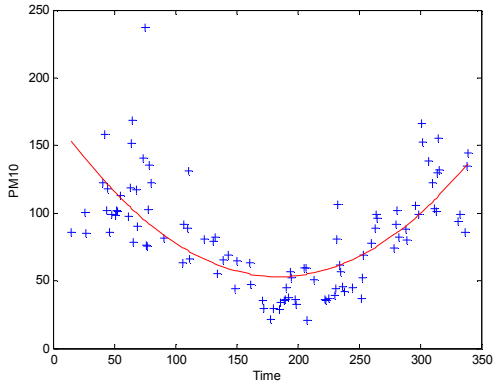

Figure 4. Main curve for 2010

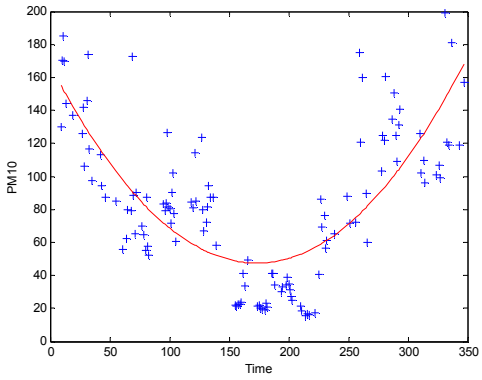

Figure 5. 2011 main curve

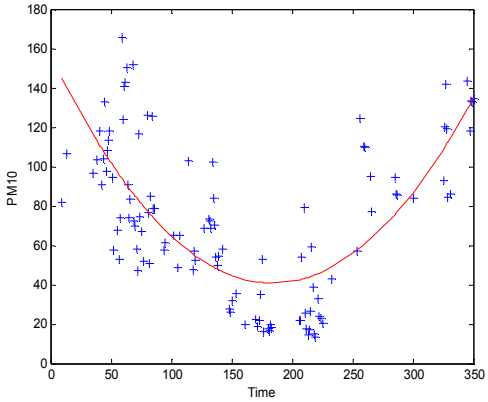

Figure 6. 2012 main curve

Then the main curve of the first three order coefficients from 2010 to 2012 is y = 0.0036 * x ^ 2-1.31 * x + 165.1972.

In order to improve the accuracy of the classification data, it is proposed to establish the main curve model for the unbalanced data, and set a corresponding threshold for the distance between the small sample and the multi-class sample to the main curve. According to the threshold, the PM10 Of the content [9]. The first step: according to the distance to the parabola distance, respectively, from 2010 to 2012 more than 150 data (less) and less than 150 data (multiple) data; the second step: the data into the sub-y = 0.0036 * x ^ 2-1.31 * x + 165.1972. The third step: calculate the three years of data and the number of classes to the main curve of the distance, through a large number of experiments the final multi-class threshold set to 118.0761, less threshold set to 168.2356. The

future forecast of the daily suspended particulate matter PM10 value [10]; only the annual order of time each day can be expected to predict the next few days the amount of PM10 in the air.

Table 2 the prediction of the PM10 content in some of the air in 2013

	daily PM10 content	predicted PM10 content	error/%
Day 1	74.8	76.1	1.7
Day 2	83.6	84.9	1.6
Day 3	88.7	89.9	1.4
Day 4	113.4	114.7	1.2
Day 5	45.1	46.4	2.8
Day22	169.4	170.5	0.6
Day 35	158.1	158.6	0.3

It can be seen from Table 2 that the predicted PM10 content by this main curve is not much different from the true PM10 value in the air; the predicted values are within the set threshold, indicating that the main curve is used to The prediction of PM10 concentration in the air can reach the theoretical requirement in the prediction accuracy.

Conclusion

The problem of unbalanced data set classification is mainly its own characteristics and the limitations of classification. Through a large number of experiments, it can be seen that the predictive model of PM10 based on the main curve is feasible and the prediction accuracy is high, and the effectiveness of the algorithm is verified by simulation experiments.

References

[1] Tao Xin-min, HAO Si-yuan, ZHANG Dong-xue, et al. Study on Unbalanced Data Classification Algorithm [J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2013, 25 (1): 101-110.
[2] Ezawa KJ, Sngh M, Norton S W. Learning goal oriented Bayesian networks for trust management [C]. In Proceedings of the 13th International Conference on Machine Learning San Morgan Kaufmann, 2006: 139-147.
[3] Liang NY, Huang G B. A fast accurate online sequential learning algorithm for feedforword networks [J]. IEEETrans Neural Networks, 2006, 17: 1411-1423.
[4] Zhang jp, Wang J. Summary of the study of the main song [J]. Journal of Computer Science, 2003, 26 (2): 129-146.
[5] Li yl, Guo Wenpu, Xu Donghui. A classification method of unbalanced data [J] .Chinese Journal of Electronics Research, 2012, 3: 246-251.
[6] Wang J, BI hy. A extreme learning machine based on particle swarm optimization [J]. Journal of Zhengzhou University (Natural Science Edition), 2013, 45 (1): 100-104.
[7] Huang GB, Zhou QY, Siew C (2004) Extreme learning maching: a new learning scheme of feedforward neural networks. In: Proceedings of Internation Joint Conference on Neural Networks, 2004, 2:985-990.
[8] Jun Y, Er MJ, An enhanced online sequential extreme learning machine algorithm [J]. Proceedings of Control and Decision Conference, 2008, 2902-2907.
[9] Turney P. Types of Cost in Inductive Concept Learning [J].The computer Research Repository.2002:15-12.
[10]P Jeatrakul, KW Wong, CC Fung, Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm[J],Neural Information Processing. 2010,6444:152-159.