2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Sparse Bayesian Flood Forecasting Model Based on SMOTEBoost

1st Yirui Wu
College of Computer and Information
Hohai University
Nanjing, China
wuyirui@hhu.edu.cn

2nd Yukai Ding
College of Computer and Information
Hohai University
Nanjing, China
dingyukai@hhu.edu.cn

3rd Jun Feng*
College of Computer and Information
Hohai University
Nanjing, China
fengjun@hhu.edu.cn

***Abstract*—Flood is a common disaster in our daily life. It's of great significance to improve the accuracy of flood forecasting, in order to help get rid of loss in both lives and property. However, there exists a uneven distribution of samples in factors of flood forecasting. Therefore, it's difficult to train a single data-driven model to describe the entire complex process of flood generation. In this paper, we propose a novel SMOTEBoost algorithm to perform flood forecasting with both high accuracy and robustness. Specifically, we firstly adopt a SMOTE algorithm to generate virtual samples, which greatly alleviates the problem of uneven sample distribution. Afterwards, we propose a sparse Bayesian model, which is trained with AdaBoost training strategy by improving its performance in over-fitting. At last, we carry out experiments on flood foretasting in Changhua river, which shows that the proposed method achieves high accuracy in prediction, thus owing practical usage.**

***Index Terms*—Flood forecasting, SMOTE, Adaboost, Sparse Bayes Model**

I. Introduction

Flood is one of the most distributed natural disasters. To improve the ability of emergency response, post-disaster reconstruction, it's very important to improve the accuracy and robustness of flood forecasting. With the establishment of intelligent hydrological monitoring stations, more and more hydrological data (rainfall, runoff, soil moisture, evaporation, etc.) are acquired and stored in the database. Mining the patterns contained in historical hydrological data with data mining methods improves the accuracy of prediction and achieve early warning.

As a regression task, flood forecasting is of great significance and difficulty. From the perspective of data analysis and research, the challenges facing researchers are mainly reflected in the following aspects: i) the number of positive and negative samples is unbalanced. Although the total number of flood samples is large, the number of samples in a single area is small, and it will be difficult to directly explore the regularity if not pre-processed. ii) The dominant factors of flood (precipitation) are clear, but it is difficult to collect all the induced factors (such as soil water content, topography, vegetation type and coverage), and the accuracy of the final forecast results is affected.

In this paper, SMOTE method [1] is used to generate virtual samples, and Sparse Bayesian flood forecasting model is trained by AdaBoost strategy. In order to cope with the imbalance between positive and negative samples of data samples (especially in regression problems), many researchers adopt resampling method. By selecting more small samples and fewer large samples, the proportion of positive and negative samples in training sample set tends to be balanced. Moniz et.al. [2] proposed a SMOTEBoost method to improve the prediction of extreme values. Wang et.al. [3] utilized SMOTE method and proposed a novel ensemble method for imbalanced data learning.

After optimizing the data by SMOTE method, this study uses AdaBoost strategy to train multiple Bayesian models to obtain the integrated model. AdaBoost algorithm [4] is a typical learning startegy based on resampling technology. By dynamically changing sample weight and model weight, the trained weak prediction models are combined into strong prediction models to improve classification accuracy. Liu et.al. [5] compare four AdaBoost method ,which helps to understand and use AdaBoost. Wen et.al. [6] use AdaBoost algorithm for vehicle classification, which indeed gain a good results.

The original Byesian model assumes that the sample obeys the probability distribution, and calculates the weight of the approximation function through the maximum likelihood criterion from the probability point of view, so as to realize the regression analysis. Sparse Bayesian [7] model adds a probability distribution constraint to the weight value of iteration training model in order to avoid over-fitting , which leads to the sparseness of the model parameters to some extent. Zhang et.al [8] design a sparse Bayesian model for classification, and the results proof the advantages of sparse Bayes.

Details of the above methods is given in Methodology, and the whole building process of model will also be introduced in methodology. Experiment part introduces the basic information of experiment basin and data while experiment evaluation criteria and results are also given in this part. Finally, conclusion part combs and analyses the whole research process, and puts forward the prospect of future research.

II. Related Work

The existing methods related to our work can be categorized into the following three types: SMOTE related method, Ad-

978-1-7281-2980-8/19/$31.00 ©2019 IEEE
DOI 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00067

aBoost related algorithm and sparse Beysian model. We thus introduce them one by one in this section.

SMOTE method is good at tackling imbalanced data problems, such classification and prediction. With development of IoT technologies [9], [10], the imbalanced data problem has become more challenging. Maldonado et.al. [11] proposed a SMOTE based method to deal with high-dimensional binary data while a novel distance metric is also proposed for the computation of the neighborhood for each minority sample. THE proposal was compared with various oversampling techniques on low- and high-dimensional datasets with the presence of class-imbalance, including a case study on Natural Language Processing (NLP). Maria et.al. [12] proposed SMOTE-BD method, which is based on SMOTE and is used to takle imbalanced classification in big data. The fully scalable preprocessing approach for imbalanced classification in Big Data is based on one of the most widespread preprocessing solutions for imbalanced classification, namely the SMOTE algorithm, which creates new synthetic instances according to the neighborhood of each example of the minority class. The novel development is made to be independent of the number of partitions or processes created to achieve a higher degree of efficiency. Experiments conducted on different standard and Big Data datasets show the good quality of the proposed design and implementation.

Most recently, Weng et.al. [13] utilized SMOTE method and random forests to improve the accuracy of student weariness prediction in education. Mohasseb et.al. [14] used a hierarchical SMOTE algorithm for balancing different types of questions. The proposed framework is grammar-based, which involves using the grammatical pattern for each question and using machine learning algorithms to classify them. Experimental results implied that the proposed framework demonstrates a good level of accuracy in identifying different question types and handling class imbalance.

AdaBoost algorithm is a efficient learning strategy for prediction especially in a big data environment [15]. It can make full use of the advantages of weak predictors while not be prone to overfitting. Chen et.al [16] proposed a novel model to classify five distinct groups of vehicle images from actual life based on AdaBoost algorithm and deep convolutional neural networks (CNNs). The experimental results demonstrated that the proposed model attains the highest classification accuracy of 99.50% on the test data set, while it takes only 28 ms to identify a vehicle image. This performance significantly outperforms the traditional algorithms. Wu et.al [17] proposed a video based fire smoke detection model using robust AdaBoost algorithm. Extensive experiments on well known challenging datasets and application for fire smoke detection demonstrate that the proposed fire smoke detector leads to a satisfactory performance. Sun et.al. [18] employ AdaBoost-LSTM ensembled learning for financial time series forecasting, and the results showed the good performance of the model. Mao et.al [19] made an Adaboost based model to generate super-resolution face image, which gained a good performance in experiment.

Sparse Beysian model has been successfully applied in many domains and have achieved desirable classification and prediction results. Mishra et.al. [20] used sparse Beysian model to fulfill target imaging and parameter estimation for monostatic MIMO radar systems, and Simulation results demonstrate enhanced imaging and estimation accuracy of the proposed sparse Beysian learning schemes in comparison with the existing techniques for MIMO radar systems. Qiao et.al. [21] studied the sparse Bayesian learning (SBL) framework for channel estimation in underwater acoustic orthogonal frequency-division multiplexing (OFDM) communication systems, which provides a desirable property of preventing structural error with fewer convergence errors for sparse signal reconstruction compared with the compress sensing-based methods. Dai et.al [22] from the perspective of sparse Bayesian learning provided a Bayes-optimal algorithm for robust DOA estimation, which can achieve excellent performance in terms of resolution and accuracy. Zheng et.al. [23] proposed an improvement of Bayesian Classifier with the sparse regression technology, which is the first attempt to extend sparse regression for directly process of categorical variables. And that method was implemented for the case of weighted naive Bayes classifier. Chen et.al. [24] proposed 2D DOA estimation algorithm based on sparse Bayesian ,which was used on a two-parallel nested arrays, which consist of two subarrays with sensors, and can estimate the two-dimensional direction of arrival (DOA) of signal sources.

III. Methodology

This section firstly introduces details of SMOTE, AdaBoost and Sparse Byesian respectively, then the framework and principle of whole model.

A. SMOTE Algorithm

For the regression problem of unbalanced data, many people choose resampling to deal with unbalanced data, that is, to select more small samples and fewer large samples. By this way, the number of positive and negative samples in training samples tends to be balanced. In fact, in addition to changing the sample distribution by re-sampling, another idea is to generate a small number of samples through a certain algorithm, called virtual samples, to increase the number of minority samples to approximate the sample equilibrium. The classical algorithm for increasing minority class samples is Synthetic Minority Over-sampling Technique(SMOTE) method, which synthesizes new sample data on the basis of original training samples through certain algorithm steps. SMOTE method generates synthetic samples in the feature space rather than in the data space of samples. For minority class samples, K-nearest neighbor method is used to select k-nearest neighbor samples, and then the k-nearest neighbor samples are used to generate virtual samples. It's noted that we usually set k as 5 in experiments.

Based on above discussion, we thus propose a specially designed SMOTE for flood prediction task, steps of which are shown in Algorithm. 1. Let's suppose the sampling training set

Algorithm 1 Steps for the proposed SMOTE algorithm.

Step1. Set number of minority class samples as n and the clustering number for K-nearest-neighbor algorithm as k.

Step2. Take one sample $S[i]$ and search in minority class samples set S for k nearest neighbors according to Euclidean distance, then save index of them into N.

Step3. To generate feature for each sample, we first pick one sample from k nearest neighbors randomly and then calculate difference value $diff$ between its real value $S[i]$ and its neighbor mean value $S[N[\alpha], j]$. It's noted that we take the index of k nearest neighbors as α.

$$diff = S[N[\alpha]], j] - S[i, j] \tag{1}$$

Step4. We first generate gap, a random number between 0 and 1 and then generate a synthetic sample (represented in a array Sy) for each feature based on $diff$ and gap.

$$Sy[ni, j] = s[i, j] + gap * diff \tag{2}$$

Step5. We construct the synthetic set M based on output of Sy.

as $s_o = \{x_i, t_i\}_{i=1}^N, x_i \in R^d, t_i \in R^1$. Specifically, we only consider minority class samples, whose statistical flow data is higher than $400m^3/s$ (alert flow). Afterwards, we could achieve a minority class samples set s, where n denotes the feature dimension of the sample. We define array operation $S[,]$, which shows that set S can be constructed as an original minority sample array for storing. Furthermore, ni represents the number of records in the synthetic class. Meanwhile, $Sy[,]$ indicates array operation on synthetic array Sy, which stores synthetic class samples.

B. AdaBoost Algorithm

Boosting algorithm [25] is a typical integration method based on resampling technology. Boosting algorithm has formed its own theoretical system and is widely used in the field of data mining to deal with unbalanced samples or integration models. It is the focus of integrated learning researchers. Some use resampling method to change the distribution of samples, and train multiple models with the original data with multiple sample distributions through multiple sampling. Samples with inaccurate prediction from the previous prediction model are taken as training samples for the next prediction model. In this way, we can take into account almost all sample distributions, and multiple models can depict multiple data with different distribution patterns.

Generally speaking, the prediction of small samples is poor. Therefore, this will realize the situation of "over-sampling minority class samples, under-sampling majority class samples". Finally, through certain strategies, the trained weak prediction models are combined into strong prediction models. Boosting algorithm was originally proposed to improve the accuracy of classification. This section applies it to flood forecasting. Samples near flood peaks can be learned many times through multiple sampling, which can effectively improve the accuracy of flood forecasting near flood peaks.

Algorithm 2 The proposed Adaboost algorithm to improve flood prediction performance.

Input: $\{x_i, t_i\}_{i=1}^N$ and sample weight $W_0(i)$ which are initialized with value $1/N$.

Output: Extract S samples from training set and then train a weak predictor $h_t(x)$.

Step1. Calculate the average error ε_t for the t-th predictor with

$$\varepsilon_t = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y_i - t_i)^2} \tag{3}$$

Then, calculate the corresponding error value $\varepsilon_{t,i}$ for each sample with following formulas:

$$\varepsilon_{t,i} = \sqrt{(y_i - t_i)^2} \tag{4}$$

Step2. Adjust weights for each sample and predictor. After that, perform retraining and verifying operation on the resulting model. Afterwards, we can achieve updated average error value β_t and sample corresponding error value $\beta_{t,i}$ with the following equations:

$$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t} \tag{5}$$

$$\beta_{t,i} = \frac{\varepsilon_{t,i}}{1 - \varepsilon_{t,i}} \tag{6}$$

where Z_t is normalization coefficient.

With updated β_t and $\beta_{t,i}$, we compute W_t and D_t with the following equations:

$$W_t(i) = \frac{W_{t-1}(i)\beta_{t,i}{}^{-\varepsilon_{t,i}}}{Z_t} \tag{7}$$

$$D_t = \frac{1}{2}ln(\frac{1}{\beta_t}) \tag{8}$$

Step3. Combine t weak predictors to construct the final and strong predictor with the following equation:

$$H(x) = \sum_{i=1}^{T} D_t h_t(x) \tag{9}$$

Based on above discussion, we thus design an adaboost algorithm for the goal of improvement of flood prediction performance in Algorithm. 2. It's noted that samples set is defined as $\{x_i, t_i\}_{i=1}^N, x_i \in R^d, t_i \in R^1$, where $W_t(i)$ denotes the weight of x_i in t-th iteration, S represents sampling capacity, D_t is the weight of t-th predictor.

C. Sparse Bayes Model

Suppose the training sample set is $\{x_i, t_i\}_{i=1}^N, x_i \in R^d, t_i \in R^1$, where t_i is the target value, x_i is the input sample, and d is the dimension of the sample. According to the characteristics of the probability distribution of the sample, it is assumed that

the target value t_i is obtained with the noise data ε_i, and the objective function is defined as follows:

$$t_i = y(x_i, \omega) + \varepsilon_i \tag{10}$$

where $\varepsilon_i \sim N(0, \sigma^2)$. The task is finding the approximation function $\hat{y}$.

Meanwhile, with the assumption that training samples conform to same distribution and are independent,the likelihood function is defined as follows:

$$p(t \mid \omega, \sigma^2) = (2\pi\sigma^2)^{-N/2} exp\{-\frac{1}{2\sigma^2} \parallel t - \Phi\omega \parallel^2\} \tag{11}$$

where $t = (t_1, t_2, \cdots, t_N)^T$, $\omega = (\omega_1, \omega_2, \cdots, \omega_N)^T$, $\Phi \in R^{N\times(N+1)}$, $\Phi = [\phi(x_1), \phi(x_2), \cdots, \phi(x_N)]^T$, $\phi(x_i) = [1, K(x_i, x_1), \cdots, K(x_i, x_N)]^T$. $K(x_i, x_N)$ is a certain kernel function. With the increase of parameters, most regression models are prone to over-fit.

In order to tackle that problem, sparse byesian method adds a constraint to the weight satisfies the conditional probability distribution, assuming that the parameter ω obeys a Gaussian distribution with a mean of 0 as shown in the following formula

$$p(\omega \mid \alpha) = \prod_{i=1}^{N} N(\omega_i \mid 0, \alpha_i^{-1}) = \prod_{i=1}^{N} \frac{\sqrt{\alpha_i}}{\sqrt{2}} e^{\frac{-\alpha_i\omega_i^2}{2}} \tag{12}$$

where $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_N\}$ is a hyperparameter that determines the prior distribution of the weight ω, which is the main cause of sparse model.

D. Total Pipeline for flood prediction

The above three subsections introduce the basic principles of SMOTE, AdaBoost and Sparse Beyes respectively. In this subsection, we will show the pipeline of how the whole model is built in Algorithm. 3.

IV. Experiment

In this paper, we take Changhua River as the study object, which is in the upper reaches of Fenshui River ,Zhejiang Province. The basin is high in the northwest and low in the southeast. It belongs to the hilly area of Western Zhejiang Province and has the characteristics of typical small and medium-sized basin. The total area of the basin is $3444km^2$, and the main stream is $1624km$ in length, with a total drop of $965m$. The tropical seasonal climatic conditions in the basin determine that the precipitation is concentrated in summer and less in winter. May to July each year is rainy season (Meiyu Season in Chinese) in which rainfall intensity is high and lasts for a long time.

There are 6 rainfall stations upstream and midstream to provide rainfall information, namely Daoshiwu, Taohuacun, Longmensi, Shuangshi, Lingxia, Yulingguan. The downstream Changhua Station is a hydrological station which can provide rainfall and flow information. Changhua River is an important tributary of Qingshandian Reservoir. Rainfall in the area converge to Changhua River and flow into Qingshandian Reservoir. Therefore, it is of great significance for flood control operation of Qingshandian Reservoir to forecast the flow of Changhua Hydrological Station. Fig. 1 is the schematic map of the Changhua Hydrological Station and its surrounding geographic location.

The annual summer flood data of Changhua River Basin from 1998 to 2010 are selected as dataset, and one data was recorded every 1 hour. The data elements include the Changhua flow and rainfall, and the rainfall of the stations in the upper reaches of Changhua. A total of 6552 samples from 1998 to 2008 are selected as training samples, and 1688

Algorithm 3 The process to construct the proposed model for flood prediction.

Step1. For original sample set O. Count the number of minority samples in O, save them in the sample set S, and save the synthesized minority samples to U, assuming that T sample are synthesized.

Step2. Determine the number of samples synthesized for each minority sample. In order to achieve as much as possible equalization, our idea is to synthesize fewer samples in dense regions and more samples in sparse regions. That is, the sample near the flood peak generates more samples, calculates the weight of each sample, and multiplies the weight by the total number of synthesized samples to obtain the virtual sample number.

$$T_i = [T \times \omega_i] = [T \times \frac{t_i}{\sum_{i=1}^{M} t_i}] \tag{13}$$

where $[\cdot]$ denotes a rounding function that chose the closest integer,M is the number of minority samples, T_i is the number of synthetic samples from t_i.

Step3. For each minority sample $x_i \in S$, find the k-nearst samples $x_{i1}, x_{i2}, \cdots, x_{ik}$ according to Euclidean distance.

Step4. Calculate the mean and variance of the features of each dimension of k neighbor samples.

$$\mu^l = \frac{1}{k}(x_{i1}^l + \cdots + x_{ik}^l) \tag{14}$$

$$\sigma^{2l} = \frac{1}{k-1}\sum_{j=1}^{k}(x_{ij}^l - \mu^l)^2 \tag{15}$$

Step5. We synthetic the target value as the following equation computes:

$$t_{ik} = x_{ik}\omega_{ik} + b_{ik} \tag{16}$$

Step6. Set the AdaBoost hyper-parameters, including the number of models in t, the single model iteration number $iter$. Set the sample and model initial weight $\omega_0(i)$ and D_0. Initialize a single sparse Bayesian model weight parameter distribution.

$$\omega_0(i) = \frac{1}{N} \tag{17}$$

$$D_0 = \frac{1}{k} \tag{18}$$

Step7. Iteratively training model. The Adam algorithm is used to implement the weight parameter update of the model.

samples from 2009 to 2010 are selected as test samples.

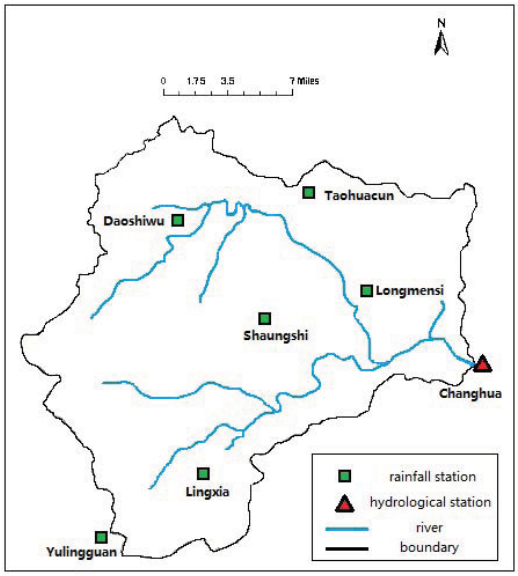

Fig. 1. Map of Changhua Basin.

A. Implementation

Python language is used as the actual coding language in the design of this system. The time accuracy is 3 hours. The length of data input time k is 6 (the actual length is 6 hours), and the prediction period is 9 hours. All experiments are carried out on Linux servers equipped with 2.4GHz 6-core Xeon CPU, 60GB RAM and Nvidia GeForce GTX 1080 Ti.

B. Evaluation

In the experiment, root mean square error (RMSE), deterministic coefficient(DC) and flood peak error(FPE) are used to evaluate the performance of the model comprehensively. Among them, deterministic coefficients and flood peak error are usually used in hydrological forecasting.

1) RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{19}$$

where, n is the number of test samples while y_i is the groundtruth and y_i' is prediction. The smaller the RMSE, the better the performance of the model.

2) DC:

$$DC = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{20}$$

where, n is the number of test samples while y_i is the groundtruth and y_i' is prediction, $\overline{y}$ denotes the mean of all groundtruths. And if DC is large(at least 0.8)then the model is good for prediction.

3) FPE:

$$FPE = \frac{1}{n}\sum_{i=1}^{n}(y_{p_i} - y_{p_i}') \tag{21}$$

where, n is the number of test samples while y_{p_i} is the groundtruth of flood peak and y_{p_i}' is prediction, FPE denotes the mean of all flood peak errors in test dataset.

C. Results

This part compares the performance of single model and Boosting model with sampling capacity of 3000, 4000 and 5000 respectively. Prediction results is listed in the TableI which shows the difference of forecasting ability of different models from three indexes of RMSE, DC and FPE.

TABLE I
DATASET CHARACTERISTICS

Model	Sampling capacity	RMSE	DC	FPE
Single model	3000	99.26	0.79	256
	4000	97.33	0.80	243
	5000	96.43	0.80	251
Ensemble model	3000	75.27	0.82	200
	4000	70.57	0.83	180
	5000	73.96	0.82	196

From the above results, we can see that the overall level of the ensemble model is higher than that of the single model. In the single model and the ensemble model, the sampling capacity has a certain impact on the performance of the model. In this experiment, when the number of sampling is 4000, the performance of the model is the best.

V. Conclusion

This paper proposes a flood forecasting model based on AdaBoost and sparse Bayesian method. Firstly, the SMOTE method is used to correct the flood data imbalance problem. Secondly, by dynamically adjusting the sample and predictor weights, multiple models with weak predictive ability are integrated into a model with strong predictive ability. The predictive performance of the overall model is enhanced while avoiding overfitting. However, the relevant experiments are relatively rough. For example, for the k in the synthetic sample, the parameters of the number of integrated models are only based on experience, and there is no experimental proof. In the next work, we will study the parameters. Based on, to further improve the model and improve model performance.

Acknowledgements

This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Natural Science Foundation of China under Grant 61702160, Grant 61872219 and Grant 61702277, the Science Foundation of Jiangsu under Grant BK20170892, and the open Project of the National Key Lab for Novel Software Technology in NJU under Grant K-FKT2017B05.

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[2] N. Moniz, R. Ribeiro, V. Cerqueira, and N. Chawla, "Smoteboost for regression: Improving the prediction of extreme values," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2018, pp. 150–159.

[3] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning: Bagging of extrapolation-smote svm," *Computational Intelligence & Neuroscience*, vol. 2017, no. 3, p. 1827016, 2017.

[4] H. Schwenk and Y. Bengio, "Adaboosting neural networks: Application to on-line character recognition," in *Artificial Neural Networks - ICANN '97, 7th International Conference, Lausanne, Switzerland, October 8-10, 1997, Proceedings*, 1997, pp. 967–972.

[5] H. Liu, H. qi Tian, Y. fei Li, and L. Zhang, "Comparison of four adaboost algorithm based artificial neural networks in wind speed predictions," *Energy Conversion and Management*, vol. 92, pp. 67 – 81, 2015.

[6] X. Wen, L. Shao, Y. Xue, and W. Fang, "A rapid learning algorithm for vehicle classification," *Information Sciences*, vol. 295, pp. 395 – 406, 2015.

[7] N. Friedman, I. Nachman, and D. Pe'er, "Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm," in *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, 1999, pp. 206–215.

[8] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Sparse bayesian classification of eeg for brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2256–2267, Nov 2016.

[9] X. Wang, L. T. Yang, X. Xie, J. Jin, and M. J. Deen, "A cloud-edge computing framework for cyber-physical-social services," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 80–85, 2017.

[10] L. T. Yang, X. Wang, X. Chen, J. Han, and J. Feng, "A tensor computation and optimization model for cyber-physical-social big data," *IEEE Transactions on Sustainable Computing*, 2017.

[11] S. Maldonado, J. López, and C. Vairetti, "An alternative smote oversampling strategy for high-dimensional datasets," *Applied Soft Computing*, vol. 76, pp. 380 – 389, 2019.

[12] M. Basgall, W. Hasperué, M. Naiouf, A. Fernández, and F. Herrera, "Smote-bd: An exact and scalable oversampling method for imbalanced classification in big data," *Journal of Computer Science and Technology*, vol. 18, p. e23, 12 2018.

[13] Y. Weng, F. Deng, G. Yang, L. Chen, J. Yuan, X. Gui, and J. Wang, "Studying weariness prediction using SMOTE and random forests," in *Smart Computing and Communication - Third International Conference, SmartCom 2018, Tokyo, Japan, December 10-12, 2018, Proceedings*, 2018, pp. 397–406.

[14] A. Mohasseb, M. B. Bader-El-Den, M. Cocea, and H. Liu, "Improving imbalanced question classification using structured smote based approach," in *2018 International Conference on Machine Learning and Cybernetics, ICMLC 2018, Chengdu, China, July 15-18, 2018*, 2018, pp. 593–597.

[15] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, and M. J. Deen, "A tensor-based big-data-driven routing recommendation approach for heterogeneous networks," *IEEE Network*, vol. 33, no. 1, pp. 64–69, 2018.

[16] W. Chen, Q. Sun, J. Wang, J. Dong, and C. Xu, "A novel model based on adaboost and deep CNN for vehicle classification," *IEEE Access*, vol. 6, pp. 60 445–60 455, 2018.

[17] X. Wu, X. Lu, and H. Leung, "A video based fire smoke detection using robust adaboost," *Sensors*, vol. 18, no. 11, p. 3780, 2018.

[18] S. Sun, Y. Wei, and S. Wang, "Adaboost-lstm ensemble learning for financial time series forecasting," in *Computational Science - ICCS 2018 - 18th International Conference, Wuxi, China, June 11-13, 2018 Proceedings, Part III*, 2018, pp. 590–597.

[19] S. Mao, D. Zhou, Y. Zhang, Z. Zhang, and J. Cao, "Weighted patches based face super-resolution via adaboost," in *2018 International Conference on Machine Learning and Cybernetics, ICMLC 2018, Chengdu, China, July 15-18, 2018*, 2018, pp. 234–239.

[20] A. Mishra, V. Gupta, S. Dwivedi, A. K. Jagannatham, and P. K. Varshney, "Sparse bayesian learning-based target imaging and parameter estimation for monostatic MIMO radar systems," *IEEE Access*, vol. 6, pp. 68 545–68 559, 2018.

[21] G. Qiao, Q. Song, L. Ma, S. Liu, Z. Sun, and S. Gan, "Sparse bayesian learning for channel estimation in time-varying underwater acoustic OFDM communication," *IEEE Access*, vol. 6, pp. 56 675–56 684, 2018.

[22] J. Dai and H. So, "Sparse bayesian learning approach for outlier-resistant direction-of-arrival estimation," *IEEE Trans. Signal Processing*, vol. 66, no. 3, pp. 744–756, 2018.

[23] Z. Zheng, Y. Cai, Y. Yang, and Y. Li, "Sparse weighted naive bayes classifier for efficient classification of categorical data," in *Third IEEE International Conference on Data Science in Cyberspace, DSC 2018, Guangzhou, China, June 18-21, 2018*, 2018, pp. 691–696.

[24] L. Chen, D. Bi, and J. Pan, "Two-dimensional angle estimation of two-parallel nested arrays based on sparse bayesian estimation," *Sensors*, vol. 18, no. 10, p. 3553, 2018.

[25] N. Duffy and D. P. Helmbold, "Boosting methods for regression," *Machine Learning*, vol. 47, no. 2-3, pp. 153–200, 2002.