EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

SMOTE-Boost-based sparse Bayesian model for flood prediction

Yirui Wu, Yukai Ding and Jun Feng*

Abstract

With a significant development of big data analysis and cloud-fog-edge computing, human-centered computing (HCC) has been a hot research topic worldwide. Essentially, HCC is a cross-disciplinary research domain, in which the core idea is to build an efficient interaction among persons, cyber space, and real world. Inspired by the improvement of HCC on big data analysis, we intend to involve related core and technologies to help solve one of the most important issues in the real world, i.e., flood prediction. To minimize the negative impacts brought by floods, researchers pay special attention to improve the accuracy of flood forecasting with quantity of technologies including HCC. However, historical flood data is essentially imbalanced. Imbalanced data causes machine learning classifiers to be more biased towards patterns with majority samples, resulting in poor classification of pattern with minority samples. In this paper, we propose a novel Synthetic Minority Over-sampling Technique (SMOTE)-Boost-based sparse Bayesian model to perform flood prediction with both high accuracy and robustness. The proposed model consists of three modules, namely, SMOTE-based data enhancement, AdaBoost training strategy, and sparse Bayes model construction. In SMOTE-based data enhancement, we adopt a SMOTE algorithm to effectively cover diverse data modes and generate more samples for prediction pattern with minority samples, which greatly alleviates the problem of imbalanced data by involving experts' analysis and users' intentions. During AdaBoost training strategy, we propose a specifically designed AdaBoost training strategy for sparse Bayesian model, which not only adaptively and inclemently increases prediction ability of Bayesian model, but also prevents its over-fitting performance. Essentially, the design of AdaBoost strategy helps keep balance between prediction ability and model complexity, which offers different but effective models over diverse rivers and users. Finally, we construct a sparse Bayesian model based on AdaBoost training strategy, which could offer flood prediction results with high rationality and robustness. We demonstrate the accuracy and effectiveness of the proposed model for flood prediction by conducting experiments on a collected dataset with several comparative methods.

Keywords: Human-centered intelligent modeling, Intelligent human-centered computing technique, Big data analyses in HCC, Flood prediction, SMOTE algorithm

1 Introduction

Human-centered computing (HCC) is a key part to interact and collaborate among persons, cyber space, and real world, which develops various human-computer applications to economically and conveniently satisfy the complex non-functional computational requirements from diverse users. Therefore, how to apply HCC in real world to effectively and efficiently solve complex problem has attracted quantity of attentions from researchers.

In this paper, we follow the idea of applying HCC on real-world problem to pursue more intelligent and efficient applications, which could meet demands from different users. Essentially, we intend to solve flood prediction problem with high accuracy under implying HCC technologies. Flood, as one of the most common and largely distributed natural diasters, happens occasionally and brings large damages to life and property. If we could

*Correspondence: fengjun@hhu.edu.cn
College of Computer and Information, Hohai University, Nanjing, China

accurately forecast flood by predicting its time-varying flow rate values in advance, hundreds of lives and quantity of property could be saved. In the past decades, researchers have proposed a quantity of models for accurate and robust flood forecast. We generally categorize them into two types, namely, mathematic models [1, 2] and data-driven models [3, 4].

Mathematic models generally describe formation of floods by function systems, representing flood processes from clues to results. Mathematic models have been successfully applied in flood forecasting systems of large watershed. However, such models are sensitive to parameters [5] and require large research efforts to adjust parameters, which prevents its massive usage for quantity of small watersheds.

Data-driven models construct forecast systems based on historical observations, directly exploring relations between river flow and flood factors without considering physical processes. Due to the developments of internet of things and sensor technologies, researchers can gather and store a quantity of hydrological data (like rainfall, runoff, soil moisture, evaporation) from different locations. Extract patterns from large historical hydrological data with intelligent methods help improve accuracy of flood prediction and could benefit from further development of the latest techniques, like deep learning human-centric representation [6–8] and intelligent human-centered computing [9–15]. Specifically, we refer to patterns as inherent non-linear functional relationship between hydrological data and flood generation, which is too complex to explain with functional system other than implicit description by machine learning models.

With an optimized future in predicting floods with artificial intelligence techniques, there exist two main challenges in applying data-driven models for practical usage. First, researchers must handle the problem of imbalanced data. Although the total number of flood samples acquired by sensors is large, some patterns with less samples can be hard to explore without suitable data argumentation methods. Second, researchers are clear about dominant factors of floods, which should be input of the constructed data-driven model. However, it is difficult to collect and use all induced factors in a single prediction model, such as soil moisture, vegetation type, and vegetation coverage. How to adaptively and incrementally use all these factors in a single model based on experts' knowledge and users' intention thus becomes a major challenge. Recently, there has been a significant progress in intelligent human-centered computing techniques, which offers possible solutions to improve classification by extracting expert knowledge from original data and appropriately modeling human intentions.

To solve these two problems, we propose a novel model to predict river runoff values. Figure 1 shows the workflow of the proposed model, which consists of three modules, namely, *Synthetic Minority Oversampling Technique (SMOTE) method*, *AdaBoost Strategy*, and *Sparse Bayesian Flood Prediction Model*. A SMOTE method in Fig. 1a is used to generate virtual samples for data augmentation, which solves the problem of imbalanced flood data to a certain extent. After preprocessing original data by SMOTE method, we adopt a novel AdaBoost strategy (represented in Fig. 1b) to train multiple Bayesian models, achieving an improved and integrated model after boosting. After integration,

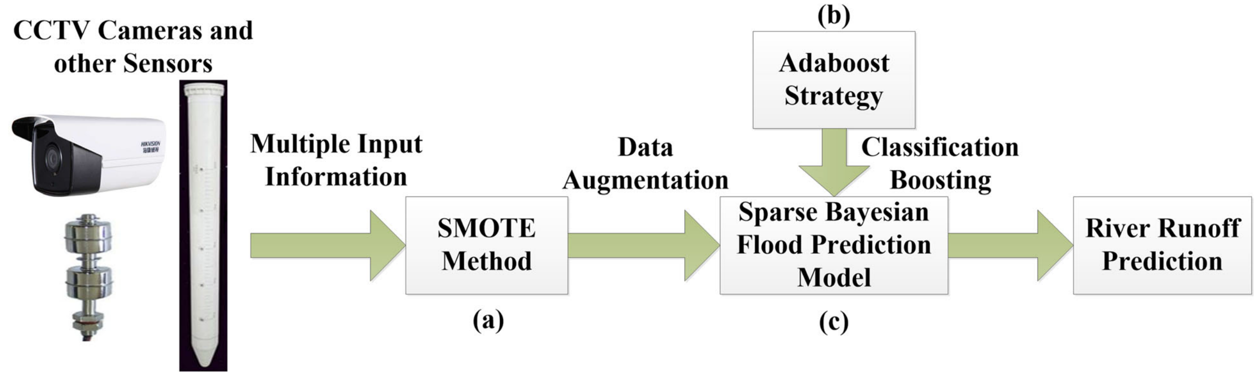

Fig. 1 Workflow of the proposed model for river runoff prediction. Detailed legend: Fig. 1 shows the workflow of the proposed model, which consists of three modules, namely, SMOTE method, AdaBoost Strategy, and Sparse Bayesian Flood Prediction Model. A SMOTE method (**a**) is used to generate virtual samples for data augmentation, which solves the problem of imbalanced flood data to a certain extent. After pre-processing original data by SMOTE method, we adopt a novel AdaBoost strategy (**b**) to train multiple Bayesian models to obtain an improved and integrated model. Afterwards, the Sparse Bayesian Flood Prediction Model (**c**), where the proposed sparse Bayesian model improves the original Bayesian model by offering a probability distribution constraint to weights of iteration training model. With all these steps, we build a complete workflow of a data-driven model to predict river runoffs with high accuracy and robustness

the boosted model could help adaptively increase prediction ability of Bayesian model and prevent the trained model to be over-fitting at the same time. Moreover, integration of different and relatively separated classifiers complete fusion process of all accessible flood factors with a single model. By selecting factors and defining parameters during integration of classifier, the proposed model further involves users' intention and experts' knowledge in constructing data-driven model, which helps realization of artificial intelligence model under principles of human-centered computing. The Sparse Bayesian Flood Prediction Model is shown in Fig. 1c, where the proposed sparse Bayesian model improves the original Bayesian model by offering a probability distribution constraint to weights of iteration training model. Such constraints leads to the sparseness of model parameters and thus helps avoid over-fitting. With all these steps, we build a complete workflow of a data-driven model to predict river runoffs with high accuracy and robustness.

The main contribution of the paper is to propose a new SMOTE-Boost-based sparse Bayesian model that supports accurate river runoff value prediction. Facing problem brought by imbalanced dataset, we utilize SMOTE method to efficiently enhance quality of samples in training dataset, which boosts performance of machine learning model, i.e., sparse Bayesian model, built on it. We believe such data enhancement method with SMOTE technology is an appropriate way to solve imbalanced data problem in HCC and big data analysis. Moreover, the proposed model provides users an efficient approach to forecast flood in advance. By involving experts' analysis and users' intentions in designing steps of data augmentation and classifier integration, we focus on implementation of human-centered computing with artificial intelligence technologies, which helps keep a balance between efficiency and model complexity. Our experimental results and the comparison results prove the high effectiveness and low complexity of the proposed model, which could support practical usage on forecasting flood. We believe this is a successful trail on how to combine principles of human-centered computing with artificial intelligence technologies, which offers inspiration for researchers on designing of novel algorithms.

The rest of the paper is organized as follows. Section 1 gives an overview of the related work. The SMOTE method for flood data augmentation is introduced in Section 2. Then, AdaBoost strategy for sparse Bayesian model under users' intention is introduced in Section 3. In Section 4, the details of the whole process for flood prediction are discussed. Section 5 shows our experimental results, and finally, Section 5 concludes the paper.

2 Related work

The existing methods related to our work can be categorized into the following three types: SMOTE-related methods, AdaBoost algorithm, and sparse Bayesian model.

2.1 SMOTE method

With the development of IoT and data computing technologies [16–18], researchers have access to achieve more data with various types and large amount. However, imbalanced data problem leads to artificial intelligence models built on these data which behave extremely poor in performance. Essentially, an imbalanced dataset refers to samples in the dataset which fail to approximately equally represent all patterns. Oversampling is an efficient technique in dealing with class imbalance problem by reduplicating or generating the minority class samples, resulting in balance between the samples of the majority and minority class. With years' development, Synthetic Minority Over-sampling Technique (SMOTE) [19] is proposed and utilized to tackle imbalanced data problem.

For instance, Maldonado et al. [20] developed a SMOTE-based method to deal with imbalanced problem of high-dimensional binary data; meanwhile, a novel distance metric is proposed to compute neighborhood for each minority sample for efficiency. Their work was compared with various oversampling techniques on imbalanced low- and high-dimensional datasets, achieving a promising result to guarantee performance in constructing NLP application. Later, Maria et al. [21] proposed a SMOTE-BD method to tackle the problem of imbalanced classification in big data. Their proposed scalable approach for imbalanced classification in big data is constructed on the basis of SMOTE algorithm, which helps create new synthetic instances according to the neighborhood of minority class sample.

Most recently, Weng et al. [22] utilized SMOTE method and random forests to improve the accuracy of student weariness prediction in education. Mohasseb et al. [23] used a hierarchical SMOTE algorithm for balancing different types of questions. Their proposed framework is grammar-based, which involves grammatical pattern for each question and machine learning algorithms to classify patterns. Experimental results implied their proposed framework demonstrates high accuracy in identifying different question types and handling class imbalance.

2.2 AdaBoost algorithm

Adaptive Boosting (AdaBoost) algorithm [24, 25] is an efficient learning strategy to build accurate classifiers. The core idea of AdaBoost is that samples misclassified by previous classifier should be used to train the next classifier. With such design, weak classifiers, which only perform well in classifying several specific patterns, can be integrated into a strong classifier, which can deal with problem of classifying all patterns. In spite of its sensitive to noise data and abnormal data, AdaBoost-based method could handle their overfit problem with its feature of integrating different classifiers. Above all, AdaBoost algorithm can make full advantages of different weak predictors; meanwhile, it is prone to prevent overfit situation.

To classify five groups of vehicle images from daily life images, Chen et al. [26] proposed a novel AdaBoost-based model with deep convolutional neural networks (CNNs) built. Experimental results demonstrated the proposed model achieves the highest classification accuracy of 99.50% on the test dataset with only 28 ms to process. Later, Wu et al. [27] utilized a robust AdaBoost model to detect fire smoke in video. Static features (including texture, wavelet, color, edge orientation histogram, irregularity) and dynamic features (including motion direction, change of motion direction, and motion speed) are extracted to train with AdaBoost strategy. They got a satisfactory performance on the final enhanced model with users' intention to adjust the weights of strong or weak classifier iteratively.

Most recently, Sun et al. [28] employed AdaBoost-LSTM-ensembled learning for financial time series forecasting. The AdaBoost algorithm is used to integrate all the long short-term memory (LSTM) predictors trained respectively. The empirical results on public datasets demonstrate that the proposed AdaBoost-LSTM ensemble learning approach outperforms some other single forecasting models and ensemble learning approaches. This suggests that the AdaBoost-LSTM ensemble learning approach is a highly promising approach for time-varying data forecasting, especially for the time series data with nonlinearity and irregularity.

2.3 Sparse Bayesian model

Sparse Bayesian learning (SBL) [29] is an important type of Bayesian statistical optimization algorithms, which is developed on the basis of Bayesian theory. Now, sparse Bayesian learning technology has been successfully applied in intelligent information retrieval [30, 31], data mining [32, 33], and other fields.

For instance, Mishra et al. [34] used sparse Bayesian model to perform parameter estimation for monostatic MIMO radar systems, where simulation results demonstrate their proposed methods achieved high estimation accuracy in comparison with the existing techniques. Later, Qiao et al. [35] proposed sparse Bayesian learning (SBL) framework for channel estimation in underwater acoustic orthogonal frequency-division multiplexing (OFDM) communication system. Compared with the compress sensing-based methods, their proposed method provides a desirable property in preventing structural error and reconstructing sparse signal with fewer convergence errors. Dai et al. [36] addressed the problem of DOA estimation in additive outliers on the basis of sparse Bayesian learning framework, which achieves excellent performance in terms of resolution and accuracy.

Most recently, Zheng et al. [37] proposed an improvement of Bayesian classifier with the sparse regression technology, which firstly tries to extend sparse regression for categorical variables and implemented with design of weighted naive Bayes classifier. Salucci et al. [32] adopted a customized multi-task Bayesian compressive sensing (MT-BCS) method to yield regularized solutions of the 3D-IS problem with a low computational complexity. Selected numerical results on representative benchmarks are presented and discussed to assess the effectiveness and the reliability of the proposed MT-BCS strategy in comparison with other competitive state-of-the-art approaches.

3 Methods

In this section, we describe steps of SMOTE method for flood data augmentation, AdaBoost strategy for classifier integration, and sparse Bayesian model for flood prediction, respectively.

3.1 SMOTE method for flood data augmentation

Class imbalance refers to the uneven distribution of training sets used in the process of training classifier. More precisely, it means the number of samples belong to a certain pattern, named as minority pattern, is too small to provide enough information for construction of classifier. If we take average loss as learning criterion on such class-imbalanced dataset, the generated model could be bias to certain patterns with large amount of samples, which could be regarded as majority pattern in our paper.

In order to deal with class imbalance problem in regression cases, resampling method is firstly used by selecting more samples with minority pattern and fewer samples with majority pattern. In that way, the proportion of samples with minority and majority pattern in training dataset tends to be balanced. However, such method can only be applied in cases with enough but imbalanced samples. In flood prediction, sensors acquire multiple variable with different frequencies according to users' intention. For example, river runoff is generally obtained every 1 h; meanwhile, soil evaporation is measured once in a day. With such constraint brought by property of sensors, we can conclude the number of samples for soil evaporation can be too small to impact on the trained classifier.

Therefore, we adopt another idea, i.e., Synthetic Minority Over-sampling Technique (SMOTE) method, to generate a number of virtual samples on the basis of original training samples, which could increase the sample number of minority pattern, thus approximating the sample

Algorithm 1 SMOTE method for flood data augmentation.

```
Input: Set with minority flood samples $S = \{\{x_{i,j}\}_{j=1}^{M}\}_{i=1}^{N}$, where $M$ and $N$ represents total feature numebr for each sample and the number of samples in dataset respectively, number of generated virtual flood samples $v$, parameter $k$
Output: Synthetic flood dataset $T$
Step1. Take $s_i$ from $S$ , and generate set of neighbor samples $S_{knn} = \{\{x_{m,j}\}_{j=1}^{M}\}_{m=1}^{k}$ by selecting $k$ samples, which have smallest Euclidean distance values with $s_i$ in feature space.
Step2. Calculate distance $d_i$ between sample $s_i$ and neighbor sample set $S_{knn}$:
    $$d_i = \sqrt{\sum_{j=1}^{M}(x_{i,j} - \frac{1}{k}\sum_{m=1}^{k} x_{m,j})^2} \qquad (1)$$
Step3. Generate a random number $\xi(0 \leq \xi \leq 1)$ and create a synthetic sample $t$ based on $d_i$:
    $$t = \{\tilde{x}_{i,j} | \tilde{x}_{i,j} = x_{i,j} + \xi * d_i; j = 1, ..., M\} \qquad (2)$$
Step4. Firstly, define one round of generation as repeating steps 1 to 3 with $i = 1, ..., N$. Then, calculate $r = \lceil \frac{v}{N} \rceil$ and perform $r$ rounds of generation, where $\lceil\ \rceil$ means rounding down operation. Finally, calculate $t = v\%N$ and repeat steps 1 to 3 with $t$ randomly chosen sample from $S$. All these generated samples makes up Synthetic flood dataset $T$.
```

equilibrium. It is noted SMOTE method generates synthetic samples in the feature space rather than data space. Under the consideration of efficiency and accuracy, we adopt k-nearest neighbor SMOTE method for generation of virtual samples. The core idea of such method is neighbor principle, that is nearest samples or samples in a group tend to own nearly same property in feature space. With such idea, we could select k nearest neighbor samples to generate virtual samples.

Under the guidance of k-nearest neighbor SMOTE method, we propose a specially designed SMOTE method for flood data augmentation. The core idea to generate virtual samples is shown in Fig. 2, where we generate a new sample in minority flood pattern s_i, named as minority sample, with synthetic sample t, which is created based on feature values of the k nearest neighbor samples forming a set of samples named as S_{knn}. With such idea, we list all steps of the proposed SMOTE method in Algorithm 1. In the **Input** line of Algorithm 1, we only perform data augmentation on minority pattern, which is defined as samples with statistical flow data higher than alert runoff value (defined as $400m^3/s$ in experiments according to China's law). The reason to adopt such definition lies in the fact that the minority pattern in flood prediction is cases of flood happening, since the majority pattern for a river are cases without floods. Specifically, we define feature value $x_i \in R^d$, where d represents the feature dimension.

It is not easy to determine the value of k and v, since too large value leads to produce similar synthetic samples and too small value may introduce too much noise into synthetic sample set. In this paper, we consulted hydrology researchers and users to determine initial value of k and v. Afterwards, we manually adjust both numbers iteratively to achieve the most robust and appropriate generation set.

3.2 AdaBoost strategy for classifier integration

AdaBoost algorithm [38] is a typical learning strategy 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. The basis of such strategy lies in the fact that objective goal is most likely sparse event in dataset of task. By involving sequences of weak classifiers, we can iteratively eliminate wrong-labeled samples to improve efficiency. Moreover, different weak classifier could be fit to handle with different input data distribution, where we could assign different weights to weak classifier based on data distribution. With such adaptive weight strategy, AdaBoost algorithm could have a consistent performance facing different dataset or application scenarios. Due to its significant ability to handle with imbalanced data by integrating various types of weak classifiers, AdaBoost algorithm has been widely used in the field of data mining and machine learning.

The most common usage to deal with imbalanced data by AdaBoost algorithm is to first resample for modification of sample distribution and then train multiple classifiers based on the modified data with multiple sample distributions, which could be achieved by multiple sampling technology. Afterwards, samples with inaccurate prediction after first round of training are taken as input of classifiers, which are built during the second round of construction. Finally, such iteration training strategy would result in a strong classifier, which is able to depict all distribution patterns inherently represented by imbalanced data.

In the case of flood prediction, constructed classifiers following common procedures would result in poor accuracy due to highly imbalanced property and shortage of enough flood data. Therefore, we propose a novel and scenically designed AdaBoost training strategy to handle case of flood prediction. The core idea of such strategy lies in the principle that we should pay more attention on sample near flood peaks, i.e., minority samples in flood prediction, which should be utilized multiple times to effectively improve the accuracy of flood forecasting near flood peaks.

Algorithm 2 AdaBoost strategy with active learning to improve performance of flood prediction.

Input: Training set $S = \{x_i, y_i\}_{i=1}^{N}$, sample size m, weight D_t for classier and $W_{i,t}$ for each sample, where i and t represents index of sample and iteration time.

Output: Integrated Classifier $H(x)$.

Step1. Initiate $W_{i,1} = 1/N$. Randomly extract m samples from S to form sample set S_t, and train a weak classifier $h_t()$ based on S_t. In our algorithm, such weak classifier refers to sparse Bayesian classifier.

Step2. Calculate the average error ε_t for the t-th classifier $h_t()$

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

Step3. Update weights for each sample $W_{i,t}$ and weights for classifier D_t with

$$W_{i,t} = \frac{W_{i,t-1}{\beta_t}^{-\varepsilon_t}}{Z_t} \tag{4}$$

$$D_t = \frac{1}{2}ln(\frac{1}{\beta_t}) \tag{5}$$

where $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ and Z_t is the normalization factor to make $\sum_{i=1}^{N} W_{i,t} = 1$.

Step4. Define uncertainty value for each sample as

$$\mu = \beta_i(h_t(x_i, l_1) - h_t(x_i, l_2)) \tag{6}$$

where β_i is the balance factor to ensure balance property between different patterns, l_1 and l_2 are the confidence output values with the largest and second largest values, respectively. In other words, smaller μ is, larger uncertainty with classifier $h_t()$ and we should use such sample for next iteration of training.

Step5. Count the number of samples in each pattern. Define number of minority pattern with smallest samples as c_1 and number of majority pattern with the largest samples as c_2. Judge whether $\frac{c_2}{c_1} > thresh$. If so, define $\beta_i = \varepsilon_t$. If not, define $\beta_i = 1$.

Step6. Calculate μ for each sample and find samples with most smallest μ to form set Φ. Finally, we achieve dataset $S_{t+1} = S_t \cup \Phi$ for next round of training.

Step7. Repeat Step1 to 6 with T iterations, and integrate T weak classifiers to construct the final classifier with

$$H(x) = \sum_{i=1}^{T} D_t h_t(x) \tag{7}$$

Considering the fact that size of flood data is growing every day, we should take more data into account for higher prediction accuracy. Therefore, new samples are involved to participate in the training for the purpose of updating classifiers. However, imbalanced new data leads to unsatisfied classifiers, where majority patterns are always updated and minority patterns are never updated. In other words, there exists an imbalanced classifier problem during the iteratively evolving of AdaBoost framework. To solve this problem, we thus propose a new sample selection strategy based on active learning technology, which chooses the most informative sample from dataset to form the training dataset, especially for majority samples.

Above all, the proposed AbaBoost training strategy with active learning technology helps relieve the burden of users on how to build accurate and strong classifiers with imbalanced data at first and then improve the constructed classifier with more data. Essentially, such method is designed under the guidance of human-centered computing, which appropriately involves more data for the improvement of constructed model without additional work of users.

Under guidance of AbaBoost training strategy with active learning technology, we design an algorithm to improve flood prediction as shown in Algorithm. 2. It is noted that steps 1 to 3 refer to the steps of an AdaBoost training strategy with sparse Bayesian classifier; meanwhile, steps 4 to 6 represent the active learning algorithm on selecting informative samples to form dataset for the next iteration of training.

3.3 Sparse Bayesian model for flood prediction

Sparse Bayes model (short for SBL) [39] assumes that sample obeys the probability distribution and calculates the weight of the approximation function through the maximum likelihood criterion. Afterwards, the posterior probability distribution is calculated with Bayesian rule. Finally, the inference of unknown parameters is made based on prior information and posterior probability.

Define training sample set as $\{x_i, y_i\}_{i=1}^{N}$. With the assumption that training samples obey the same distribution and are independent, we can define the likelihood function as:

$$p(y \mid \omega, \sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2} \parallel t - \Phi\omega \parallel^2\right\} \tag{8}$$

where $y = (y_1, y_2, \cdots, y_N)^T$, $\omega = (\omega_1, \omega_2, \cdots, \omega_N)^T$, $\Phi \in R^{N\times(N+1)}$, $\Phi = [\phi(x_1), \phi(x_2), \cdots, \phi(x_N)]^T$, $\phi(x_i) = [1, K(x_i, x_1), \cdots, K(x_i, x_N)]^T$, and $K(x_i, x_N)$ is a certain kernel function. It is noted that most regression models are prone to over-fit with the increase of number of parameters.

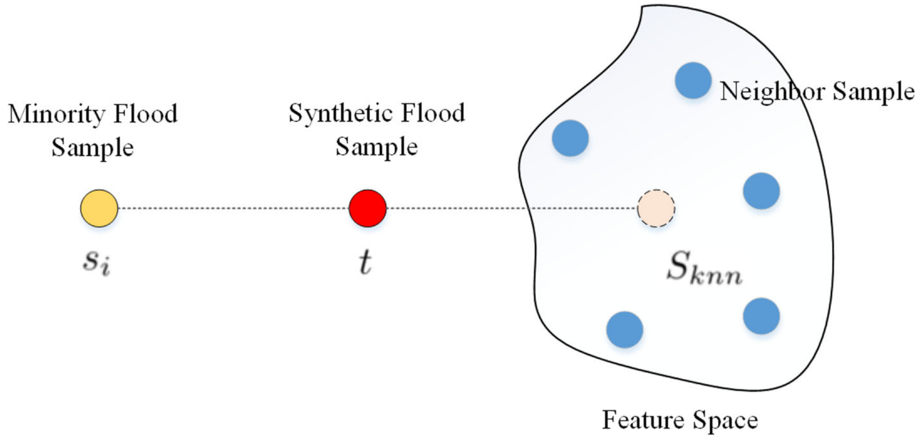

Fig. 2 Illustration of our proposed SMOTE algorithm. Detailed legend: The core idea to generate virtual samples is shown in Fig. 2, where we generate new sample in minority flood pattern s_i, named as minority sample, with synthetic sample t, which is created based on feature values of the k nearest neighbor samples forming a set of samples named as S_{knn}

In order to tackle that problem, SBL adds a constraint to the weight that the parameter ω obeys a Gaussian distribution with its mean value equals 0. With such constraint, Eq. 8 can be rewritten as

$$p(\omega \mid \alpha) = \prod_{i=1}^{N} N(\omega_i \mid 0, \alpha_i^{-1}) = \prod_{i=1}^{N} \frac{\sqrt{\alpha_i}}{\sqrt{2}} e^{\frac{-\alpha_i \omega_i^2}{2}} \tag{9}$$

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 idea to construct a sparse model.

After defining the sparse Bayesian model, we could facilitate the pipeline of the whole proposed model with the parts of SMOTE, AdaBoost, and sparse Bayes model, where SMOTE is designed to generate virtual sample, sparse Bayes model is defined as the weak classifier, and AdaBoost training with active learning technology is to integrate all weak classifiers constructed during the training iterations, which finally form a strong and accurate classifier for flood prediction.

4 Results and discussion

In this section, we show the effectiveness of the proposed method in predicting runoff values. We would describe dataset, quality measures, and experimental results, respectively.

4.1 Dataset

In this study, we apply the proposed method to predict daily flow rate of Changhua Gage Station, based on the 1998–2010 historical flood data of 7 rainfall stations, 1 evaporation station and 1 gaging station in Changhua watershed, a watershed in Xinanjiang River basin in China. We show the map of the Changhua watershed with various kinds of stations in Fig. 3. Note that we need to predict the flow rate values of river gaging station Changhua and station Shuangshi functions as an evaporation station to offer evaporation values. We collect hourly data of 40 floods happened from 1998 to 2010 and utilize 8-folder cross-validation to evaluate our proposed method. A total of 6552 samples from 1998 to 2008 are selected as training samples, and 1688 samples from 2009 to 2010 are selected as test samples. It is noted that the collected data from Changhua river is an essential imbalanced dataset, where some flood patterns only occur once in all samples. The imbalanced property of Changhua dataset is the major difficulty for accurate flood prediction.

The Changhua River is a tributary of Xinanjiang River, originated from Jixi County, Anhui Province, China. It flows through Jixi County, Lingan County, Changhua County, and eventually into Xinanjiang River. The river is 96-km long and the watershed area is 905 km^2. Changhua gage station is a major gage station in Changhua River, located in 119.212 E, 30.166 N. The daily flow rates from 1998 to 1986 at Changhua gage station as well as other related data during floods are collected for this study. Some descriptive statistics for the flood data is given in Table 1, where E represents evaluation and SD refers to the standard deviation. The daily flow rate varies from 0.58 m$^3/s$ occurring in 2007 to 2100 m^3/s appearing in 1999; the mean daily flow rate is 146.651 m^3/s with a variance of 202.501 m^3/s.

4.2 Quality measures

We use standard quality measures such as root mean square error (RMSE), deterministic coefficient (DC), and flood peak errors (FPE) for measuring the quality of flood forecasting achieved by the proposed method. Note that the latest measurement is specially designed for

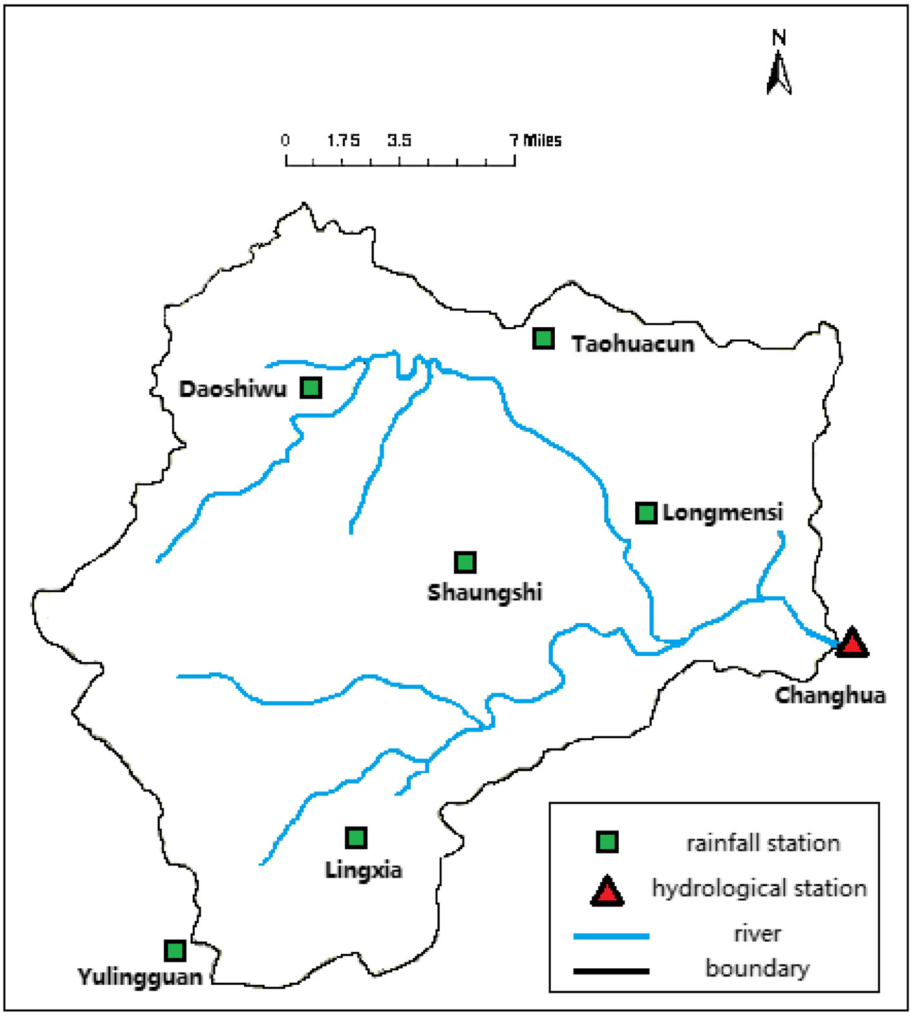

Fig. 3 Map of the Changhua watershed with various kinds of stations. Detailed legend: The sketch map of study area is presented in Fig. 3 and the prediction is performed based on the 1998–2010 historical daily flow rates of 1 gaging station (SS), rainfall values of 7 rainfall stations (including DSW, THC, LMS, SS, LX, YLG, and CH), and evaporation values of 1 evaporation station (SS) in Changhua watershed, a watershed in Xinanjiang River basin in China

flood forecasting by emphasizing the appearance time and values of flood peak, which often brings most serious damage to persons and property. During these three measurements, RMSE could be represented as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_c(i) - y_0(i)\right]^2} \tag{10}$$

where RMSE reflects the degree of deviation between predicted values y_c and true values y_0 during the flood forecasting process. The smaller RMSE has a better performance achieved by the adopted model. Measurement DC could be represented as:

$$\mathrm{DC} = 1 - \frac{\sum_{i=1}^{n}\left[y_c(i) - y_0(i)\right]^2}{\sum_{j=1}^{n}\left(y_0(i) - \bar{y_0}\right)^2} \tag{11}$$

where $y_c(i)$ is the predicted value, $y_0(i)$ is the measured value, $\bar{y_0}$ is the measured value mean, and n is the number of samples. It is noted DC reflects the degree of coincidence between the flood forecasting process and the measured process. The closer the result is to 1, the higher the forecast accuracy rate. The third measurement FPE could be formulated as

$$\mathrm{FPE} = \frac{1}{n}\sum_{i=1}^{n}(y_{p_i} - y'_{p_i}) \tag{12}$$

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

Table 1 Descriptive statistics of daily flow and relevant data from 1998 to 2010 in Changhua dataset, where DSW, THC and other abbreviates represent names of rainfall, evaporation and gaging stations, and *p*, *R* and *E* refer to flow rates observed at CH, rainfall observed at rainfall stations and evaporation observed at SS, respectively

E	p	R(DSW)	R(THC)	R(LMS)	R(SS)	R(LX)	R(YLG)	R(CH)	E(SS)
Mean	146.651	0.596	0.618	0.704	0.786	0.712	0.822	0.631	0.090
SD	202.501	2.411	2.303	2.636	2.553	2.560	2.666	3.405	0.071
Median	80.320	0	0	0	0	0	0	0	0.083
Skewness	3.987	11.026	7.946	10.700	7.819	8.599	7.484	27.617	1.205
kurtosis	24.388	198.996	89.910	190.660	104.110	120.549	100.293	1124.804	7.474

4.3 Results and discussion

We conduct two groups of experiments to show the performance of the proposed model with different parameters and compare with other models for runoff prediction.

In the first group of experiment, we show the prediction results of the proposed model with different parameters, i.e., sampling number of samples in SMOTE method and training iterations in AdaBoost training strategy. Comparison results are shown in Table 2. For convenience of readers, we further show a comparison figure in Fig. 4, where n refers to the number of sampling samples. From Fig. 4, we can clearly see RMSE, DC, and FPE values achieved by the proposed ensemble model which is much higher than that of the single model, which proves the effectiveness of the proposed AdaBoost training strategy with active learning technology. Furthermore, we find that the adopted iteration, i.e., the number of classifiers, is clearly affected by margin effect. In other words, adopting more classifiers does not always improve measurement values. Therefore, we try different iterations and achieve the best performance with 6 iterations.

Sampling number of samples is the most important parameter for SMOTE algorithm and has a great impact on the final prediction results. From Fig. 4, we can find that the best performance is achieved by the model by defining $n = 4000$. Setting either $n = 3000$ or $n = 5000$ would get lower performance, since too small number of virtual samples would not help reduce imbalanced data problem to a large extent and too large number of virtual samples would bring too much noise for accurate prediction. The consistent performance in three measurements, i.e., RMSE, DC, and FPE, further proves the robustness of the proposed model by firstly generating virtual samples and then integrating weak classifiers, where both procedures has high intention for dealing with the problem of imbalanced data.

In the second group of tests, we show the detailed statistics of the proposed method and other data-driven-based methods for the Changhua dataset in Table 3. Among these comparative methods, Han et al. [40] apply SVM in flood forecasting with a special design on optimum selection among a large number of various input combinations and parameters. Note that we apply linear kernel function for [40] during experiments. Wu et al. [41] construct entities and connections of Bayesian network to represent variables and physical processes of a famous physical model, which appropriately embeds hydrology expert knowledge for high rationality and robustness. Dawson et al. [42] develop Artificial Neural Networks (ANNs) for 6 h lead times flow forecasting using real hydrometric data. Chang et al. [43] develop a two-stage rainfall runoff model for 3-h-ahead flood forecasting based on radial basis function (RBF) neural network, which firstly utilize fuzzy min-max clustering to determine the characteristics of the nonlinear RBFs and then adopt multivariate linear regression to determine the weights between the hidden and output layers. Above all, the cores of Han et al. [40], Dawson et al. [42], Chang et al. [43] , Lima et al. [44], and Wu et al. [41] are SVM, Neural Network, Radical Basis Function Network, Extreme Learning Machine, and Bayesian Network, respectively. All these machine learning structures are popular to predict floods in pattern recognition community. We implement these algorithms according to the instructions given in their papers.

From Table 3, we could see the proposed method achieves the best performance in RMSE, DC, and the second best performance in FPE. The small value of RMSE

Table 2 Performance comparison with different sampling number of samples and training iterations

Model	Sampling Number	RMSE	DC	FPE
Single week classifier	3000	99.26	0.79	256
	4000	97.33	0.80	243
	5000	96.43	0.80	251
Ensemble model with 3 iterations	3000	84.58	0.81	218
	4000	80.27	0.82	207
	5000	82.13	0.81	214
Ensemble model with 6 iterations	3000	75.27	0.82	200
	4000	70.57	0.83	180
	5000	73.96	0.82	196

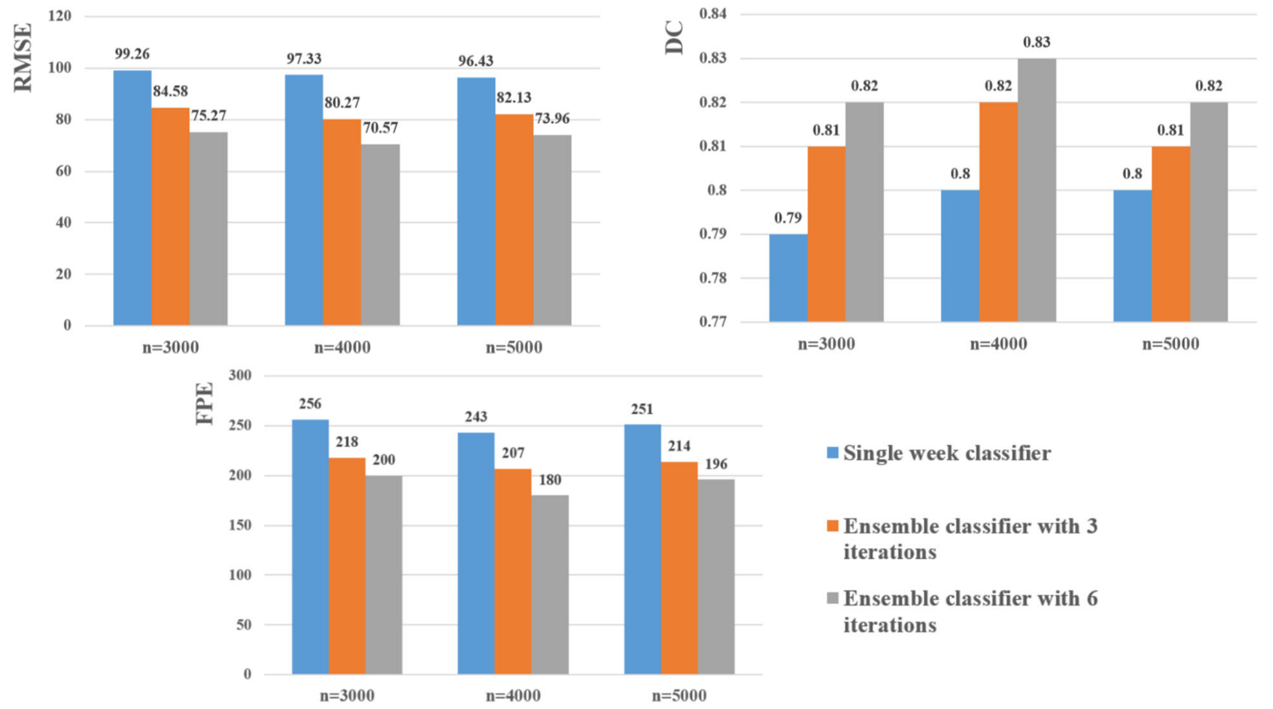

Fig. 4 Comparison of RMSE, DC, and FPE values achieved by the proposed model with different parameters. Detailed legend: We show the prediction results of the proposed model with different parameters, i.e., sampling number of samples in SMOTE method and training iterations in AdaBoost training strategy, in Fig. 4, where *n* refers to the number of sampling samples

by the proposed method implies our method is more accurate and robust to predict runoff values; meanwhile, large value of DC achieved implies our method quantify uncertainty to a certain extent. Wu et al. [41] is more accurate in predicting the appearance time and runoff values of flood peaks than the proposed model, since it contains the embedded hydrology processes and variables to increase prior knowledge for accurate prediction of flood peaks. To sum up, both generating virtual samples and integrating classifiers help accurately predict floods even with imbalanced data. Due to not adopting heavy deep learning architecture, the proposed method could averagely operative one input sample in 3.41s on a PC with 2.4 GHz 2-core i7 CPU, 16G RAM, which is fast enough in time complexity to guarantee instant flood prediction.

Table 3 Performance comparison with comparative data-driven methods on Changhua dataset. It is noted that we adopt 4000 samples and 6 iterations to train the proposed model

Methods	DC	RMSE	FPE
Han et al. [40]	0.79	96.31	210
Dawson et al. [42]	0.76	94.29	194
Chang et al. [43]	0.82	83.59	203
Lima et al. [44]	0.71	85.15	198
Wu et al. [41]	0.80	78.55	**175**
The proposed model	**0.83**	70.57	180

5 Conclusions

This paper proposes SMOTE-Boost-based sparse Bayesian model to perform tasks of accurate flood prediction. During the first step, SMOTE method is used to solve the imbalanced flood data problem by generating more virtual samples. Under a framework of AdaBoost training strategy with property to dynamically adjust sample number and weights for samples and classifiers, multiple sparse Bayesian models with weak predictive ability are integrated into a model with strong predictive ability. We further involve active learning technology to update the model by selecting informative samples for training. Experiments have demonstrated the accuracy and effectiveness of the proposed model for flood prediction on a collected dataset with several comparative methods. In the future work, we will study the parameters based on AdaBoost training strategy to further improve the model and improve model performance.

Abbreviations

SMOTE: Synthetic Minority Over-sampling Technique; IoT: Internet of things; NLP: Natural Language Processing; AdaBoost: Adaptive Boosting; CNN: Convolutional Neural Network; LSTM: Long short-term memory; SBL: Sparse Bayesian learning; MIMO: Multiple-input multiple-output; OFDM: Orthogonal frequency-division multiplexing; DOA: Direction of arrival; SD: Standard deviation; MSE: Mean square error; DC: Deterministic coefficient; FPE: Flood peak errors; SVM: Support vector machine

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions. The authors are grateful to the National Science Foundation of China and Science Foundation of Jiangsu for their support of this research.

Authors' contributions

The algorithms proposed in this paper have been conceived by YW, YD, PS, and JF. YW and YD designed the experiments. YW and YD performed the experiments and analyzed the results. YW is the main writer of this paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key R&D Program of China under Grant 2018YFC0407901; the Natural Science Foundation of China under Grant 61702160, Grant 61672273, and Grant 61832008; the Science Foundation of Jiangsu under Grant BK20170892; and Key R&D Program Projects in Jiangsu Province under Grant BE2015707.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

Received: 2 September 2019 Accepted: 6 March 2020
Published online: 15 April 2020

References

1. E. Paquet, F. Garavaglia, R. Garçon, J. Gailhard, The SCHADEX method: a semi-continuous rainfall–runoff simulation for extreme flood estimation. J. Hydrol. **495**(15), 23–37 (2013)
2. M. Rogger, A. Viglione, J. Derx, G. Blöschl, Quantifying effects of catchments storage thresholds on step changes in the flood frequency curve. Water Resour. Res. **49**(10), 6946–6958 (2013)
3. S. Han, P. Coulibaly, Bayesian flood forecasting methods: a review. J. Hydrol. **551**, 340–351 (2017)
4. D. L. Shrestha, D. P. Solomatine, Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. **19**(2), 225–235 (2006)
5. C. Yao, K. Zhang, Z. Yu, Z. Li, Q. Li, Improving the flood prediction capability of the Xinanjiang model in ungauged nested catchments by coupling it with the geomorphologic instantaneous unit hydrograph. J. Hydrol. **517**, 1035–1048 (2014)
6. X. Xu, Y. Xue, L. Qi, Y. Yuan, X. Zhang, T. Umer, S. Wan, An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles. Futur. Gener. Comp. Syst. **96**, 89–100 (2019)
7. X. Xu, Q. Liu, Y. Luo, K. Peng, X. Zhang, S. Meng, L. Qi, A computation offloading method over big data for IoT-enabled cloud-edge computing. Futur. Gener. Comp. Syst. **95**, 522–533 (2019)
8. X. Xu, Y. Li, T. Huang, Y. Xue, K. Peng, L. Qi, W. Dou, An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks. J. Netw. Comput. Appl. **133**, 75–85 (2019)
9. X. Xu, S. Fu, L. Qi, X. Zhang, Q. Liu, Q. He, S. Li, An IoT-oriented data placement method with privacy preservation in cloud environment. J. Netw. Comput. Appl. **124**, 148–157 (2018)
10. X. Xu, Y. Chen, Y. Yuan, T. Huang, X. Zhang, L. Qi, Blockchain-based cloudlet management for multimedia workflow in mobile edge computing. Multimed. Tools Appl. (2019). https://doi.org/10.1007/s11042-019-07900-x
11. X. Xu, X. Liu, L. Qi, Y. Chen, Z. Ding, J. Shi, Energy-efficient virtual machine scheduling across cloudlets in wireless metropolitan area networks. Mob. Netw. Appl. 1–15 (2019)
12. X. Xu, X. Zhang, M. Khan, W. Dou, S. Xue, S. Yu, A balanced virtual machine scheduling method for energy-performance trade-offs in cyber-physical cloud systems. Futur. Gener. Comput. Syst. **105**, 789–799 (2020)
13. L. Qi, R. Wang, C. Hu, S. Li, Q. He, X. Xu, Time-aware distributed service recommendation with privacy-preservation. Inf. Sci. **480**, 354–364 (2019)
14. L. Qi, Q. He, F. Chen, W. Dou, S. Wan, X. Zhang, X. Xu, Finding all you need: Web APIs recommendation in web of things through keywords search. IEEE Trans. Comput. Soc. Syst. **6**(5), 1063–1072 (2019). https://doi.org/10.1109/tcss.2019.2906925
15. L. Qi, Y. Chen, Y. Yuan, S. Fu, X. Zhang, X. Xu, A QOS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems. World Wide Web. **23**(2), 1275–1297 (2020)
16. X. Wang, L. T. Yang, X. Xie, J. Jin, M. J. Deen, A cloud-edge computing framework for cyber-physical-social services. IEEE Commun. Mag. **55**(11), 80–85 (2017)
17. L. T. Yang, X. Wang, X. Chen, J. Han, J. Feng, A tensor computation and optimization model for cyber-physical-social big data. T-SUSC. **4**(4), 326–339 (2019)
18. X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, M. J. Deen, A tensor-based big-data-driven routing recommendation approach for heterogeneous networks. IEEE Netw. **33**(1), 64–69 (2018)
19. A. Fernández, S. Garcia, F. Herrera, N. V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)
20. S. Maldonado, J. López, C. Vairetti, An alternative smote oversampling strategy for high-dimensional datasets. Appl. Soft Comput. **76**, 380–389 (2019)
21. M. Basgall, W. Hasperué, M. Naiouf, A. Fernández, F. Herrera, SMOTE-bd: an exact and scalable oversampling method for imbalanced classification in big data. J. Comput. Sci. Technol. **18**, 23 (2018)
22. Y. Weng, F. Deng, G. Yang, L. Chen, J. Yuan, X. Gui, J. Wang, in *Proceedings of Third International Conference on Smart Computing and Communication*. Studying weariness prediction using SMOTE and random forests (Springer, Birmingham, 2018), pp. 397–406
23. A. Mohasseb, M. B. Bader-El-Den, M. Cocea, H. Liu, in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. Improving imbalanced question classification using structured smote based approach (IEEE, 2018). https://doi.org/10.1109/icmlc.2018.8527028
24. Y. Freund, R. E. Schapire, in *Proceedings of Second European Conference on Computational Learning Theory*. A decision-theoretic generalization of on-line learning and an application to boosting (Springer, New York, 1995), pp. 23–37
25. Q. Huang, Y. Chen, L. Liu, D. Tao, X. Li, On combining biclustering mining and AdaBoost for breast tumor classification. IEEE Trans. Knowl. Data Eng. **32**(4), 728–738 (2020)
26. W. Chen, Q. Sun, J. Wang, J. Dong, C. Xu, A novel model based on AdaBoost and deep CNN for vehicle classification. IEEE Access. **6**, 60445–60455 (2018)
27. X. Wu, X. Lu, H. Leung, A video based fire smoke detection using robust AdaBoost. Sensors. **18**(11), 3780 (2018)
28. S. Sun, Y. Wei, S. Wang, in *Lecture Notes in Computer Science*. AdaBoost-LSTM ensemble learning for financial time series forecasting (Springer, 2018), pp. 590–597. https://doi.org/10.1007/978-3-319-93713-7_55
29. M. E. Tipping, Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
30. J. Fu, G. Wu, Y. Zhang, L. Deng, S. Fang, Active user identification based on asynchronous sparse Bayesian learning with SVM. IEEE Access. **7**, 108116–108124 (2019)
31. J. Dai, A. Liu, H. C. So, Sparse Bayesian learning approach for discrete signal reconstruction. CoRR. **abs/1906.00309** (2019)
32. M. Salucci, L. Poli, G. Oliveri, Full-vectorial *3D* microwave imaging of sparse scatterers through a multi-task Bayesian compressive sensing approach. J. Imaging. **5**(1), 19 (2019)

33. Y. Yang, Research on the single image super-resolution method based on sparse Bayesian estimation. Clust. Comput. **22**(Suppl 1), 1505–1513 (2019)
34. A. Mishra, V. Gupta, S. Dwivedi, A. K. Jagannatham, P. K. Varshney, Sparse Bayesian learning-based target imaging and parameter estimation for monostatic MIMO radar systems. IEEE Access. **6**, 68545–68559 (2018)
35. G. Qiao, Q. Song, L. Ma, S. Liu, Z. Sun, S. Gan, Sparse Bayesian learning for channel estimation in time-varying underwater acoustic OFDM communication. IEEE Access. **6**, 56675–56684 (2018)
36. J. Dai, H. So, Sparse Bayesian learning approach for outlier-resistant direction-of-arrival estimation. IEEE Trans. Sig. Process. **66**(3), 744–756 (2018)
37. Z. Zheng, Y. Cai, Y. Yang, Y. Li, in *Proceedings of Third IEEE International Conference on Data Science in Cyberspace*. Sparse weighted naive Bayes classifier for efficient classification of categorical data (IEEE, 2018), pp. 691–696. https://doi.org/10.1109/dsc.2018.00110
38. H. Schwenk, Y. Bengio, in *Lecture Notes in Computer Science*. AdaBoosting neural networks: application to on-line character recognition (Springer, 1997), pp. 967–972. https://doi.org/10.1007/bfb0020278
39. N. Friedman, I. Nachman, D. Pe'er, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm (Morgan Kaufmann, Stockholm, 1999), pp. 206–215
40. D. Han, L. Chan, N. Zhu, Flood forecasting using support vector machines. J. Hydroinformatics. **9**(4), 267–276 (2007)
41. Y. Wu, W. Xu, J. Feng, S. Palaiahnakote, T. Lu, in *2018 24th International Conference on Pattern Recognition (ICPR)*. Local and global Bayesian network based model for flood prediction (IEEE, 2018). https://doi.org/10.1109/icpr.2018.8546257
42. C. W. Dawson, R. Wilby, An artificial neural network approach to rainfall-runoff modelling. Hydrol. Sci. J. **43**(1), 47–66 (1998)
43. F.-J. Chang, J.-M. Liang, Y.-C. Chen, Flood forecasting using radial basis function neural networks. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **31**(4), 530–535 (2001)
44. A. R. Lima, A. J. Cannon, W. W. Hsieh, Forecasting daily streamflow using online sequential extreme learning machines. J. Hydrol. **537**, 431–443 (2016)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.