Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2015, Article ID 562724, 8 pages
http://dx.doi.org/10.1155/2015/562724

Research Article

Imbalanced Data Sets Classification Based on SVM for Sand-Dust Storm Warning

Yonghua Xie,[1] Yurong Liu,[2,3] and Qingqiu Fu[1]

[1]*School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*Department of Mathematics, Yangzhou University, Yangzhou 225002, China*
[3]*Communication Systems and Networks (CSN) Research Group, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Yonghua Xie; xyh_76@nuist.edu.cn

Received 6 February 2015; Revised 15 April 2015; Accepted 19 April 2015

Academic Editor: Zidong Wang

In view of the SVM classification for the imbalanced sand-dust storm data sets, this paper proposes a hybrid self-adaptive sampling method named SRU-AIBSMOTE algorithm. This method can adaptively adjust neighboring selection strategy based on the internal distribution of sample sets. It produces virtual minority class instances through randomized interpolation in the spherical space which consists of minority class instances and their neighbors. The random undersampling is also applied to undersample the majority class instances for removal of redundant data in the sample sets. The comparative experimental results on the real data sets from Yanchi and Tongxin districts in Ningxia of China show that the SRU-AIBSMOTE method can obtain better classification performance than some traditional classification methods.

1. Introduction

Sand-dust storm is an important environmental problem (Juaergui, 1989) and is also one of the main causes of desertification. In the last few decades, sand-dust storms have blanketed many regions over the world (e.g., North-East Asia, particularly China) more frequently and tend to be more severe [1]. The frequent occurrences of sand-dust storms disrupt our daily life, and put our environment, economy and health at risk [2], which brings about more and more focuses on the research of sand-dust storm warning.

SVM (support vector machine) is a new generation of machine learning algorithms based on VC (Vapnik-Chervonenkis) dimension theory and structural risk minimization principle of statistical learning theory which were proposed by Vapnik in 1990s. It is a good way to solve practical problems with small samples, high dimensions, and local minimum [3]. Because of these advantages, SVM is widely used in various fields of regression and pattern recognition, and also has been gradually used in the research on sand-dust storm warning in recent years. SVM belongs to the supervised classification method. In order to obtain better generalization ability, the numbers of different types of samples should make no difference. Class imbalance occurs when the number of instances of one class exceeds the number of instances of other classes in samples or training data sets. In imbalanced data sets, the class with the largest number of samples is called majority class, and the other classes are called minority classes [4]. The occurrence probability of sand-dust storm is very small and the final number of sand-dust storm samples is far less than the number of non-sand-dust storm samples, so the sand-dust storm samples belong to minority class and non-sand-dust storm samples belong to majority class. If the SVM is used for imbalanced data of sand-dust storm, the interface of classification will approach the minority class [5] and the predictions based on minority class will generally have poor performance results. In view of the above analysis, the imbalanced distribution of sand-dust storm data has been an important problem in the sand-dust storm warning technology.

Currently, the methods used to solve the problems of imbalanced data mainly focus on two levels: algorithm level

and data level. Research on algorithm level contains modifying traditional algorithms or proposing new algorithms. In terms of the traditional algorithms, many measures are used to make it applicable to the classification of minority class by adjusting the cost function between different types of samples, changing the probability density, adjusting the classification boundary, and so on [6, 7]. These methods on algorithm level are often restricted to apply to one kind of data sets because they do not change the distribution of the samples. In practical applications, data preprocessing methods are more popularly used. So this paper mainly focuses on the methods of data level. The method of data level is one kind of resampling which includes undersampling and oversampling [8]. Generally undersampling method balances the number of different categories by reducing the number of samples in majority class. However, when part of data in majority class is deleted randomly, the potential and useful data will be removed as well; thus, some important information in the majority class will lose. In contrast with the undersampling, oversampling balances the samples with the method of repeating samples in the minority class or creates some samples artificially, which may increase training time and prone to overfitting. Therefore, the two methods of undersampling and oversampling are often mixed to use in practice. Lin et al. [9] proposed an algorithm which combined K-Means clustering undersampling with SMOTE (synthetic minority oversampling technique), and the results on UCI (University of California, Irvine) test data demonstrated the effectiveness of this algorithm. However, SMOTE technology made the possibility of interclass repeating increase which composed new samples using existing minority samples without considering the distribution of neighbor samples. Tao et al. [10] combined BSMOTE (border synthetic minority oversampling technique) which only oversampled on boundary samples with ODR (optimization of decreasing reduction) method. This new algorithm balances training data through not only considering the distribution of majority class on the boundary of minority class, but also removing the noise and redundant information from majority class. Dong and Cai [11] proposed an improved method of SMOTE called space synthetic minority oversampling technique which generated new synthetic samples inside the super geometry based on the minority class and its K-nearest neighbors. This method overcomes the limitation which the new samples generated only on the connection between two samples and performs better than SMOTE for the classification performance of minority class and the whole data set. Xu et al. [12] proposed another space sampling method based on SMOTE called ISMOTE (improved synthetic minority oversampling technique). ISMOTE improves the imbalanced distribution of data through randomizing interpolation in the spherical space constituted of the minority class instances and their nearest neighbors. The ISMOTE was validated to have substantial advantages over SMOTE, but also have some problems. First, the minority class samples need to be oversampled many times when the majority class samples far exceed the minority class samples, which will lead to overlearning. Second, this method does not add new information to minority class, and the number of inserted samples for each minority class samples can only be a unique value instead of a random number based on the distribution of neighbor samples. In recent years, there still exist a lot of research results in imbalanced data sets classification. Cao et al. [13] introduced a new method based on Particle Swarm Optimization. This method optimized sampling rate and selected the feature set simultaneously through particle swarm optimization with the imbalanced data evaluation metric as objective function. The experimental results show this method has substantial advantages. Li et al. [14] presented a support vector machine algorithm based on space spreading through space spreading in multidimensional Euclidean space based on space spreading principle, which added the size of minority class data sets by using upsampling. Although these algorithms have better performance, the shortages which are mentioned above still exist.

As a result, combined with the random undersampling method, this paper proposes a hybrid self-adaptive sampling method named SRU-AIBSMOTE algorithm for sand-dust data classification. This algorithm makes full use of distributed information of different sample classes and imports adaptive degree and space interpolation technology into SMOTE method. The undersampling and oversampling are introduced to improve the classification performance, in which the oversampling randomizes interpolation in the sphere space to overcome the limitations of simple linear interpolation, while the undersampling only samples on the nonboundary samples of majority class for retaining the useful information. The experimental results on the real data sets from Yanchi and Tongxin district in Ningxia validate good classification performance of the proposed method.

The paper is organized as follows. In Section 2 the principle of traditional algorithm and improved algorithm (SRU-AIBSMOTE) is introduced. The steps for describing the improved algorithm are also presented in this section. In Section 3, we present the experimental data sets and results. The experimental conclusion and future research directions are presented in Section 4.

2. Hybrid Self-Adaptive Sampling Algorithm (SRU-AIBSMOTE)

2.1. SMOTE Algorithm. The basic idea of SMOTE is balancing the samples by generating new minority class samples synthetically. Instead of copping minority class samples simply, SMOTE algorithm generates samples through linear interpolation between minority class samples which locate closely each other. The main steps are as follows.

First, based on the rate of oversampling N, this method chooses N samples for each minority class samples from their K neighbor samples, which should also belong to minority class. Second, SMOTE algorithm generates N new samples as formula (1) for each minority class samples. Finally, it combines the new samples with quondam data sets:

$$\text{New Sample} = x + \text{rand}(0,1) \times (x_i - x), \quad (1)$$

where $i = 1, 2, \ldots, N$, rand$(0, 1)$ stands for the random number between 0 and 1, x stands for each minority class

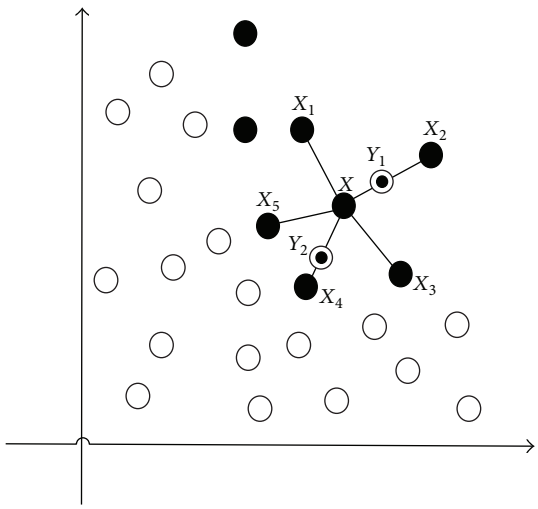

Figure 1: Schematic diagram of SMOTE algorithm.

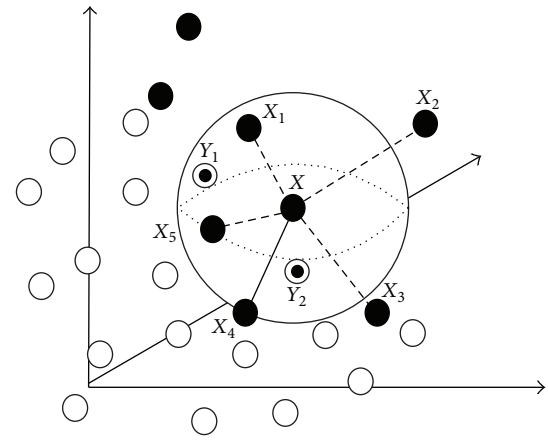

Figure 2: Schematic diagram of AIBSMOTE algorithm.

samples, x_i stands for these chosen neighbor samples and New Sample stands for the new generated samples.

Figure 1 presents a schematic diagram of SMOTE algorithm, in which the white circles represent majority class samples and the black ones represent minority class samples. X is a minority class sample, and X_1, X_2, X_3, X_4, X_5 are five neighbors of X. Y_1, Y_2 stand for the new composed minority class samples, which are on the line between X and its neighbors X_2, X_4, which are also belong to minority class. By increasing the number of minority class samples, this algorithm reduces the degree of imbalance.

2.2. The Improved Oversampling Algorithm AIBSMOTE. The SMOTE algorithm generates new samples by oversampling in the line between minority class sample and its nearest neighbor that belong to the same class [15]. But the proposed SVM model still has following problems:

(1) SMOTE algorithm is based on the assumption that the samples which are close to the minority class also belong to minority class. Actually there exits the situation that some samples closed to minority class belong to majority class such as the boundary samples, and the classification interface created by SVM method only depends on the support vectors which only exit around the boundary samples. In order to improve the classifier performance, these samples at the boundary need to be used to generate new samples. So the traditional SMOTE is not applicable for SVM classification.

(2) Samples generated by traditional SMOTE are always on the line connected by two samples. However, the real distribution of sand-dust storm data is in multidimensional space, a simple linear sampling will have some limitations for classification.

(3) SMOTE algorithm synthesizes the same number of new samples for each minority sample, but the actual situation is that the sample distribution around each minority class samples is different. So using the same synthesis method cannot meet the actual requirements.

In order to solve the above problems and find suitable method to handle imbalanced data for SVM forecasting model, this paper improves the SMOTE algorithm and proposes a space oversampling algorithm called AIBSMOTE.

For the first question, in order to maximize the performance of the classifier, AIBSMOTE algorithm only oversamples the boundary samples because the effect of SVM classification algorithm is mostly determined by its support vectors around boundary samples. In order to solve the second problem, AIBSMOTE algorithm proposes the concept of spatial interpolation. The main idea of this algorithm is taking the boundary minority class sample as the center and taking the Euclidean distance between the center and its nearest neighbor sample as the radius of an n-dimensional sphere. The virtual minority class samples in this sphere will be generated randomly. Figure 2 is a schematic diagram of AIBSMOTE algorithm, X_1, X_2, X_3, X_4, X_5 are neighbors of boundary minority sample X, X_4 is one of neighbors which are selected randomly. The new generated samples Y_1, Y_2 are shown in the figure, which are interpolated randomly in the sphere by taking X as the center and the distance between X and X_4 as the radius.

For the third problem, in AIBSMOTE algorithm the number of the new samples of each minority class instance is based on the majority class samples number of its K nearest neighbors. Due to the boundary minority samples, the closer to majority samples the greater possibility to make a mistake. In order to strengthen the training on error prone samples, AIBSMOTE algorithm synthesizes more virtual samples when the number of majority samples around minority samples is relatively large. This algorithm is described as follows.

Assuming that the minority class sample set is p, $p = \{p_1, p_2, \ldots, p_{p\text{num}}\}$, $p$num stands for the number of minority class instances and the number of attributes of each instance is w, the instance p_i can be represented as $p_i = \{p_{i1}, p_{i2}, \ldots, p_{iw}\}$.

Step 1. Find K number nearest neighbors of each minority class instance p_i by using K-means method which can be described as $Ne_i = \{Ne_{i1}, Ne_{i2}, \ldots, Ne_{iK}\}$, the number of majority class instances in neighbors is n_i. Compared n_i with K, if $n_i = 0$, the samples closed to p_i all belong to minority class and then put p_i into Psafe set; if $0 < n_i < K$, p_i belongs to boundary samples, put p_i into Pboundary set; if $n_i = K$, p_i maybe noise samples, put them into Pleft set.

Step 2. Calculate the number of inserted virtual samples ADn_i for each minority sample in the Pboundary set through formula (2), in which $b$num is the number of boundary minority class instances and Insertnum stands for the total number to be generated:

$$ADn_i = \frac{n_i}{\sum_{i=1}^{b\text{num}} n_i} \times \text{Insertnum}. \tag{2}$$

Step 3. For each minority class instance p_i in Pboundary set, choose ADn_i number of instances in the range of its K nearest neighbors randomly. Generate a virtual instance x_{new} in the sphere combined by p_i and x_{new}, it must also satisfy the following formulas [12]:

$$\left\|x_{\text{new}} - p_i\right\| \le \left\|p_i - x_{\text{near}}\right\| \tag{3}$$

$$x_{\text{new}j} = p_{ij} + \text{rand}_j \cdot \left|x_{\text{near}j} - p_{ij}\right|, \quad 1 \le j \le w, \tag{4}$$

where $\|x_{\text{new}} - p_i\|$ and $\|p_i - x_{\text{near}}\|$ represent the corresponding Euclidean distance, $|x_{\text{near}j} - p_{ij}|$ is absolute value of difference between the corresponding attributes for instance p_i and instance x_{near}, rand_j is a random number which locates interval $(0, 1)$ when x_{near} belongs to minority class or locates interval $(0, 0.5)$ when x_{near} is majority instance. Through this way, the new samples will be generated more close to minority class samples and it is better to avoid samples aliasing.

Step 4. Repeat Step 3 until enough virtual minority instances are obtained and then combine the new set with Psafe set and Pleft set to get the final training data.

AIBSMOTE algorithm uses 1-NN method to test the new virtual samples. The new sample belongs to minority class if its nearest neighbor belongs to minority class and this virtual sample is valid. Virtual sample whose nearest neighbor belongs to majority class is judged as noise sample, which should be discarded. The new virtual samples can be resynthesized in the same way. Thus, the space oversampling algorithm can effectively synthesize valid minority class samples; its effectiveness will be verified further by simulation results.

2.3. Random Under-Sampling Algorithm SRU. Since the noise and redundant information in the majority samples would seriously affect the generation of SVM classifier interface, the samples in minority class need to be oversampled for many times when the number of majority samples is very large. In this way, classification will create a small decision domain after training and overlearning. So it is necessary to undersampling for majority class samples.

RU (random undersampling) algorithm is the most basic undersampling algorithm, which balances data sets by deleting some of the majority class samples. But this method will lead to the missing of some important information because when part of majority class samples is removed randomly, some useful potential data will also be deleted. Considering that boundary-samples are very important to SVM-classification, in order to avoid removing the useful samples by mistake, this paper introduces an undersampling algorithm which only undersampling on safety samples called SRU algorithm. The main idea is using KNN (K-nearest neighbor algorithm) to detect safety-samples and then undersampling randomly on safety-samples, which can choose some majority class samples on safety samples randomly and then remove this samples from majority class samples. The noise and redundant information will be removed at the same time when reducing the quantity gap between different classes.

2.4. SRU-ABSMOTE Algorithm. Oversampling balances the samples through repeating samples in the minority class or creating some samples artificially, which may increase training time and prone to overfitting, while undersampling method balances the number of different categories by reducing the number of samples in majority class, which easily lead to information missing. In order to overcome the shortcomings of the two sampling algorithms, this paper introduces a new kind of SVM algorithm called SRU-AIBSMOTE which combines RU (random undersampling) algorithm and AIBSMOTE algorithm together. This algorithm is divided into two parts: the first part is undersampling by SRU algorithm and the second part is using improved oversampling algorithm-AIBSMOTE to increase the number of minority class samples. As is shown in Figure 3, the algorithm can be described as follows.

(a) Use the K-nearest neighbor (KNN) to detect safety-samples from minority class and take them as Psafe. Similarly, take detected left-samples from minority class as Pleft, take detected boundary-samples from minority class as Pboundary, take detected boundary-samples from majority class as Nboundary and take detected safety-samples from majority class as Nsafe in data set T.

(b) According to the formula (5), the number of minority class samples to be inserted can be calculated as Insertnum, the parameter ∂ denotes the number of the inserted samples compared to the difference in the number of majority class samples and minority class samples. Then, ADn_i can be obtained according to the Insertnum and formula (2):

$$\text{Insertnum} = \partial \times (n\text{num} - p\text{num}). \tag{5}$$

(c) Insert samples for each of minority class samples into Pboundary data set and save it as a new virtual minority class with ABSMOTE oversampling algorithm.

(d) Remove S number of samples from safe samples of majority class-Nsafe with random undersampling

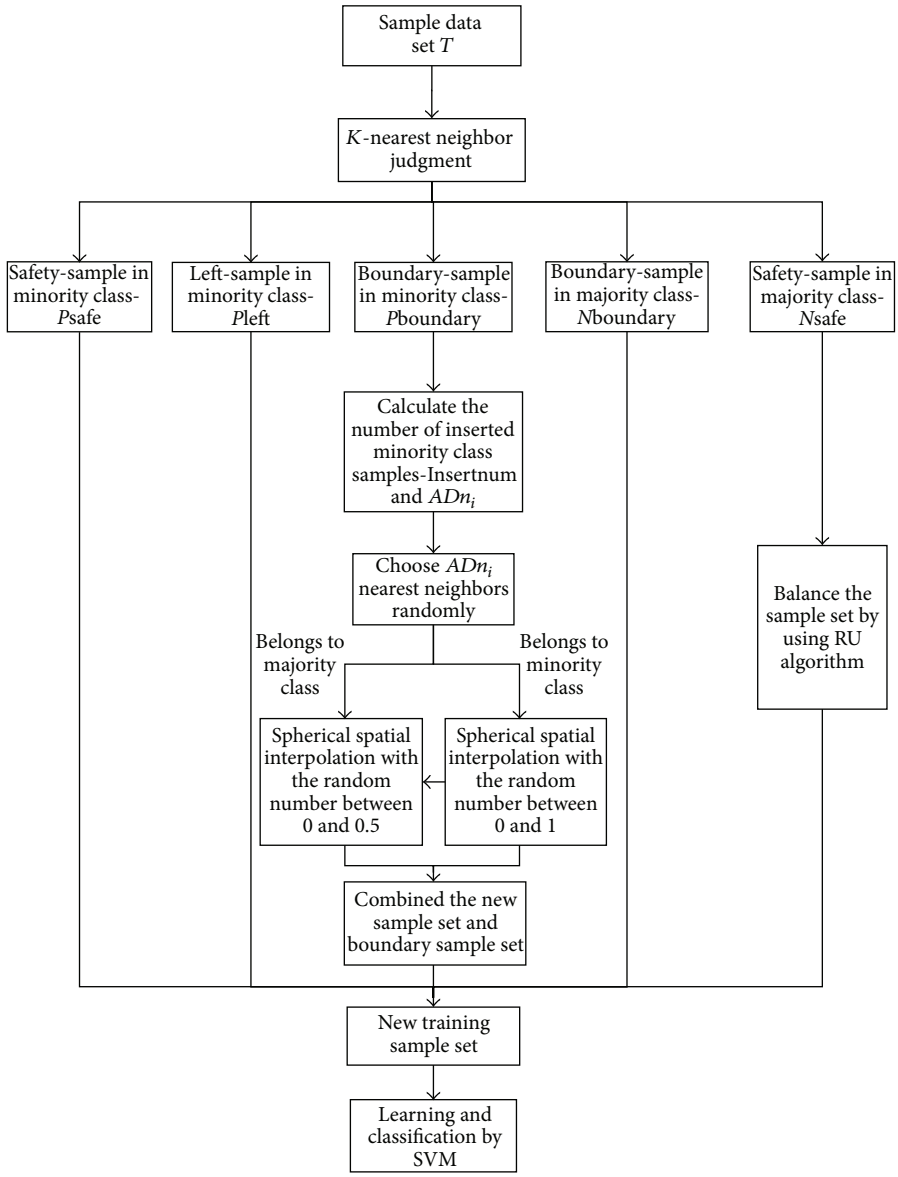

Figure 3: Flow chart of RU-AIBSMOTE algorithm.

(RU) method. S is the difference between the number of majority class samples and the number of minority class samples after oversampling.

(e) Create a new data set for SVM training and classification.

3. Experimental Results and Analysis

3.1. Experimental Data. The physical generating mechanism of sand-dust storm is very complex, which should meet three requirements: source of sand, huge wind, and unstable structure of lower atmosphere. According to the requirements, this paper selected daily average atmospheric pressure, daily average temperature, daily temperature difference (the difference between the highest temperature and lowest temperature in

Table 1: The features for sand-dust storm warning.

Characteristic	Unit
Daily average atmospheric pressure	0.1 hpa
Daily average temperature	0.1°C
Temperature difference	0.1°C
Daily average relative humidity	%
Daily average precipitation data	0.1 mm
Daily average wind velocity	0.1 m/s
Daily hours of sunshine	0.1 h
Geopotential height of 500 hpa	m
Zonal wind-speed of 700 hpa	0.1 m/s
Meridional wind-speed of 700 hpa	0.1 m/s
Potential temperature of 850 hpa	0.1 K

Table 2: Sample data set.

1	2	3	4	5	6	7	8	9	10	11	12
8674	21	79	27	32700	33	42	5636	1.4200	−1.5900	28.5123	0
8710	3	160	25	0	18	102	5645	1.6700	−4.9900	28.5846	0
8602	64	174	28	0	50	106	5648	7.4200	3.5200	29.2738	0
8539	85	154	32	0	75	70	5627	9.2900	−3.4400	29.4885	1
8585	45	173	8	0	40	89	5647	6.6200	12	29.3660	1
8604	63	75	13	0	33	78	5648	1.2200	3.2000	29.4414	1
8618	36	85	24	32700	20	0	5606	−5.1100	1.3500	28.8181	0
8634	34	42	57	30002	20	0	5595	−6.6600	−1.4400	28.1509	0
8622	34	134	71	32700	25	28	5591	3.2000	3.6200	28.5385	0
8619	37	122	71	0	38	89	5622	4.7000	2.5500	28.6872	1

a day), daily average relative humidity, daily average precipitation data, daily average wind velocity, and daily hours of sunshine as forecasting features of this classification system from the daily data of meteorological observation station [16]. According to [1], this paper also selected 4 kinds grid data of NCEP [17] (National Centers for Environmental Prediction) which are the nearest to the observation station: geopotential height of 500 hpa, southeast wind velocity, northwest wind velocity of 700 hpa, and potential temperature of 850 hpa. The total number of forecasting features is 11 as shown in Table 1. The class of sand-dust storm will be marked according to the data set of Chinese strong sand-dust storm [16].

According to the above methods, the data set is collected from Yanchi observation station (station number: 53723) and Tongxin observation station (station number: 53810) in Ningxia from March to May during the period from 1956 to 1965 when sand-dust storm happened frequently. A certain item of data will be removed if the corresponding information in this set is lacked. Ten examples of the data are shown in Table 2. Every column means one kind of feature which is shown in Table 1. The last column shows the situation of sand-dust storm, 1 means a sand-dust storm exists while 0 means not.

3.2. Standard of Performance Evaluation. The traditional performance evaluation is based on the integrated performance of a classification; in other words, it is based on the accuracy of classified results with all samples. However, in the imbalanced data set, the minority class is easy to be classified falsely. Sand-dust storm is an event of small probability. For example, if the number of sample sets is 100 and only one of them is sand-dust storm sample, then after classification all of them were recognized as non-sand-dust storm. Thus, in traditional evaluation method the accuracy is 99% (to majority class), but to minority class (the sand-dust storm) the accuracy is 0%. In sand-dust storm forecast, the accuracy of sand-dust storm is more important, so the traditional performance evaluation is not feasible.

Imbalance data set often uses the following indexes for performance evaluation [13]. Provided that the minority class is defined as P and the majority class is defined as N. FN means the number of samples in minority class which are classified as majority class. FP means the number of samples in majority class which are classified as minority class. TN and TP mean the number of samples in majority class and in minority class separately which are classified correctly. The mixing matrix of sand-dust storm data set is shown in Table 3.

Table 3: Mixing matrix of sand-dust storm data set.

	Recognized as non-sand-dust storm	Recognized as sand-dust storm
Non-sand-dust storm sample	TN	FP
Sand-dust storm sample	FN	TP

Table 4: Number of the samples with different algorithm.

	Yanchi (53723)				Tongxin (53810)			
	N_s	N_{ns}	N_{As}	N_{Asn}	N_s	N_{ns}	N_{As}	N_{Asn}
SVM	127	793	127	793	113	807	113	807
RU-SMOTE	127	793	381	460	113	807	452	455
*K*means-SMOTE	127	793	381	350	113	807	339	350
SRU-AIBSMOTE	127	793	332	330	113	807	323	323

The sensitivity of sand-dust storm sample is

$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{6}$$

The precision of sand-dust storm sample is

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{7}$$

The F-measure of sand-dust storm sample is

$$F = \frac{2 * \text{Sensitivity} * \text{precision}}{\text{Sensitivity} + \text{precision}}. \tag{8}$$

Both sensibility and precision are considered in formula (8). The parameter F will be low if both of them are low or one of them is low and the other is high, while will be high only both of these two values are high. So the classifier will obtain better performance with the higher parameter F.

3.3. Experimental Results and Analysis. The experiments were performed with data from Ningxia Yanchi and Tongxin

Table 5: The comparison performances of different algorithms.

	Yanchi (53723)			Tongxin (53810)		
	Sensitivity	Precision	F-measure	Sensitivity	Precision	F-measure
SVM	43.39%	40.18%	41.72%	42.43%	40.17%	41.27%
RU-SMOTE	80.20%	81.59%	80.89%	80.34%	80.67%	80.50%
Kmeans-SMOTE	81.01%	83.09%	82.04%	80.46%	81.59%	81.02%
SRU-AIBSMOTE	87.54%	85.25%	86.38%	87.91%	86.52%	87.21%

stations, which used libsvm-mat-2.89-3 tool box to train and test SVM model and then achieved simulation experiment with MATLAB R2010a. In order to evaluate the performance of SRU-AIBSMOTE algorithm, the results were compared with those of the representational algorithms referred above, including the classical SVM algorithm, the imbalance SVM algorithm based on random under-sampling and SMOTE [18] (RU-SMOTE), and Kmeans-SMOTE algorithm [9] with cluster undersampling and SMOTE. Each data set uses tenfold cross validation method to reducing the impact of random. The kernel function of SVM classification is Gaussian kernel, in KNN (K-nearest neighbor) method, parameter K is set as 5, in RU-SMOTE parameter ∂ is set as 0.3. In SRU-SMOTE algorithm, the ratio of removed samples number in majority class to difference between majority samples number and minority samples number is 1 to 2. In Kmeans-SMOTE algorithm, let the number of samples for each class reach to 350. The experimental results are shown in Tables 4 and 5.

Table 4 lists the difference of sample number before and after sampling with different algorithms. In this table, N_s stands for the number of sand-dust storm samples, N_{ns} stands for the number of non-sand-dust storm samples, N_{As} is the number of sand-dust storm samples after sampling, and N_{Asn} is the number of non-sand-dust storm samples after sampling. Comparing with the traditional SVM algorithm, the other three algorithms combine different undersampling and oversampling separately and balance the original data in different levels. The experimental results show that after resampling the ratio of samples number in non-sand-dust storm class to samples number in sand-dust storm class reduces from 6.25 and 7.14 approaches to 1. For self-adaptation in RU-ABSMOTE, it can insert different number of virtual samples into minority class around boundary based on the distribution of samples in majority class, which can make the number of samples in these two classes similar. Table 5 shows the different performances of these four algorithms. First, the correct classification of resampling algorithm is higher than that of the traditional SVM algorithm by 30% to 40% in every index. RU-SMOTE algorithm performs undersampling with all samples in majority class randomly. Because of some blindness of undersampling, some important data will be removed together with the deletion of noise and redundant data; therefore, the parameter F in RU-SMOTE algorithm is lowest. Comparing with the other two algorithms, no matter Yanchi or Tongxin stations the SRU-AIBSMOTE algorithm shows the highest values in three indexes, for example, the parameter F reaches 87.21% which outperform the other two algorithms by almost 7%. The results validate the good classification performance of our proposed method.

4. Conclusion

This paper proposes the SRU-AIBSMOTE algorithm which combined random undersampling with hybrid self-adaptive space sampling for sand-dust storm data classification. In terms of the oversampling, it can adaptively adjust neighbor selection strategy based on the internal distribution of the sample sets and randomized interpolation in the sphere space rather than simple linear interpolation on the minority class, while the undersampling only sample on the safe set of majority class for retaining the boundary samples. The experimental results on the real data sets from Yanchi and Tongxin district in Ningxia show that the SRU-AIBSMOTE method improves the overall performance of sand-dust storm data classification.

Although this paper has made certain progress in the study of SVM for sand-dust storm warning, but the future research work is needed from the following directions: firstly, in the forecast area, this paper proposed a warning method for single station, but the occurrence of dust storms is regional. Taking the relevant meteorological factors of the around stations into account, the warning model will be more complex. Therefore, how to establish a nationwide sand-dust storm warning system is one of future research direction. Secondly, the forecasting model in this paper is a qualitative prediction. Therefore, how to upgrade the qualitative forecasts for the quantitative prediction is also research direction in the future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by Special Fund for Meteorological Research in the Public Interest (GYHY201306015) and National Natural Science Foundation of China (61375030).

References

[1] Z. Y. Lu, Q. M. Zhang, and Z. C. Zhao, "SVM in the sand-dust storm forecasting," in *Proceedings of the 5th International*

Conference on Machine Learning and Cybernetics, pp. 3677–3681, August 2006.

[2] M. Akhlaq, T. R. Sheltami, and H. T. Mouftah, "A review of techniques and technologies for sand and dust storm detection," *Reviews in Environmental Science and Biotechnology*, vol. 11, no. 3, pp. 305–322, 2012.

[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[4] K. El-Tawil and A. A. Jaoude, "Stochastic and nonlinear-based prognostic model," *Systems Science & Control Engineering*, vol. 1, no. 1, pp. 66–81, 2013.

[5] J. van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 935–942, ACM, New York, NY, USA, 2007.

[6] M. M. Arefi, J. Zarei, and H. R. Karimi, "Observer-based adaptive stabilization of a class of uncertain nonlinear systems," *Systems Science and Control Engineering*, vol. 2, no. 1, pp. 362–367, 2014.

[7] G. Wu and Y. Chang E, "Class-boundary alignment for imbalanced dataset learning," in *Proceedings of the Workshop on Learning from Imbalanced Data Sets II (ICML '03)*, pp. 49–56, Washington, DC, USA, August 2003.

[8] Y. Yong, "The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm," *Energy Procedia*, vol. 17, pp. 164–170, 2012.

[9] S. Y. Lin, C. H. Li, Y. Jiang, C. Lin, and Q. Zou, "Under-sampling method research in class-imbalanced data," *Journal of Computer Research and Development*, vol. 48, no. 2, pp. 47–53, 2011.

[10] X. M. Tao, Z. J. Tong, Y. Liu, and D. D. Fu, "SVM classifier for unbalanced data based on combination of ODR and BSMOTE," *Control and Decision*, vol. 26, no. 10, pp. 1535–1541, 2011.

[11] X. Dong and L. J. Cai, "S-SMOTE method in class imbalance data sets," *Computer Simulation*, vol. 29, no. 12, pp. 175–179, 2012.

[12] D. D. Xu, Y. Wang, and L. J. Cai, "ISMOTE algorithm for imbalanced data sets," *Journal of Computer Applications*, vol. 31, no. 9, pp. 2399–2401, 2011.

[13] P. Cao, B. Li, and D. Z. Zhao, "Imbalanced data learning based on particle swarm optimization," *Journal of Computer Applications*, vol. 33, no. 3, pp. 789–792, 2013.

[14] Z. Li, W. Wang, and H. Guo, "SVM classification algorithm for solving multi-class imbalance data," *Computer Engineering and Design*, vol. 35, no. 7, pp. 2499–2503, 2014.

[15] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, pp. 878–887, Springer, Berlin, Germany, 2005.

[16] China Meteorological Data Sharing Service System [EB/OL], 2013, http://data.cma.gov.cn/.

[17] Earth system research laboratory [EB/OL], 2013, http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

[18] M. Zhu and X. M. Tao, "The SVM classifier for unbalanced data based on combination of RU-undersample and SMOTE," *Information Technology*, no. 1, pp. 39–43, 2012.