International Journal of Computational Intelligence Systems, Vol. 8, Supplement 1 (2015), 39-47

GA-Based Feature Selection Method for Imbalanced Data with Application in Radio Signal Recognition

Limin Du[1,2], Yang Xu[1*], Jun Liu[3], Fangli Ma[1,4]

[1] *Intelligent Control Development Center, Southwest Jiaotong University, Chengdu，Sichuan 610031, P.R. China*
E-mail: dulimin@henu.edu.cn, xuyang@home.swjtu.edu.cn

[2] *Pharmacy College of Henan University, Kaifeng, Henan 475004, P.R. China*
E-mail: dulimin@henu.edu.cn

[3] *School of Computing and Mathematics, Ulster University, Northern Ireland, UK*
E-mail: j.liu@ulster.ac.uk

[4] *Sichuan Provincial Radio Monitoring Station, Chengdu, Sichuan 610016, P.R. China*
E-mail: scmfl@163.com

Received 8 February 2015

Accepted 27 October 2015

Abstract

This paper presents an improved genetic algorithm (GA) based feature selection method for imbalanced data classification, which is then applied to radio signal recognition of ground-air communication. The proposed method improves the fitness function while SVM is selected as the classifier due to its good classification performance. This method is firstly evaluated using several benchmark datasets and experimental results show that the proposed method outperforms the original GA-based feature selection method now that it not only reduces the feature dimension effectively, but also improves the precision of the minor class. Finally, the proposed method is applied to a real world application in radio signal recognition of ground-air communication, which again shows comparatively better performance.

Keywords: feature selection; genetic algorithm; imbalanced data; radio signal recognition; ground-air communication.

1. Introduction

Development of electronic technology (e.g., radio communication) brings great convenience to our life, in the meanwhile it also brings new challenges, e.g., radio monitoring. Radio monitoring is not only an important technical means to get radio spectrum data and to maintain the order of radio waves, but also an important part of radio management [1]. Radio recognition is an essential task of the radio signal monitoring. It can provide theoretical guidance and technical support for other parts of the radio monitoring and plays an important role in radio communications [2, 3]. In the past, the signal identification mostly depends on the operators who observed radio spectrum diagrams acquired through the broadband receivers and made intuitive judgments according to the image visual characteristics. This kind of method relies heavily on the knowledge, experience and capability of the operators and it can't meet the needs of real-time and large scale radio monitoring. Accordingly, automatic radio signal identification/classification

* Corresponding author.

technologies based on the radio monitoring software have emerged as an alternative solution and nowadays a growing number of automatic radio signal identification technologies have been developed by large research efforts and been applied in many areas such as communication industry and national defense construction [4, 5].

Radio interference of ground-air communication is a critical issue in the civil aviation. It is often accidental with a low probability of the occurrence, but it is of great harmfulness. In order to ensure flight safety, radio monitoring staffs need to identify abnormal signals rapidly and then investigate further. At present, the abnormal signals of ground-air communication are always detected based on monitoring staffs' experience combined with the application of monitoring equipments, which leads to heavy workload of the staffs [6]. Therefore, automatic radio signal identification technology for radio monitoring is very effective to deal with this type of monitoring business. Since the abnormal signals are far less than the normal signals, radio signal recognition of ground-air communication belongs to the imbalanced data problem.

The class imbalance problem refers to the issue that occurs when a data set is dominated by a major class or classes which have more instances than the other rare/small classes. Imbalanced datasets exist widely in real life, such as gene detection, text classification and fault diagnosis [7]. People take more notice of small class or classes than major class or classes generally and the cost of wrong classification of small class is usually much higher than that of major class. As for signal recognition problem, effective measures can't be taken to deal with it in time if abnormal signal is classified as normal signal, the consequence could be disastrous. For example, if jamming signals of ground-air communication can not be identified correctly, it will affect aviation safety of aircraft.

At present there are three kinds of methods to deal with the problem of imbalanced data: re-sampling [8-12]; improvement of classical classification algorithms [13-17]; and feature selection [18-20]. A detailed review of learning from imbalanced data can be referred to ref. 30.

The present work focuses on feature selection based classification method for imbalanced data with the aim at improving it by using a computational intelligence technique, i.e., genetic algorithm. Genetic algorithm (GA) is a heuristic search algorithm, using the reference of natural selection and genetic mechanism in living nature. A number of feature selection methods based on genetic algorithm have been proposed [21-25]. For imbalanced data, the features selected by these methods can improve much more the recognition rate of major class which led to improve the overall classification accuracy rate, but the recognition rate of minor class which is much more important is not improved as high as expected. The use of a fitness function based on the G-mean is mentioned in the literature 26, but it is proposed for biomedical applications and not suitable for our application background - radio signal recognition of ground-air communication.

Considering the imbalance of radio signal data and the superiority which is shown up by the support vector machine (SVM) in radio signal recognition, a new feature selection method for imbalanced data sets based on genetic algorithm is proposed through improving the fitness function while SVM is selected as the classifier due to its good classification performance in this paper. The proposed method is firstly comparatively evaluated using several benchmark datasets and then applied to radio signal recognition of ground-air communication. This article is symmetrically extended and much more elaborated version of the paper presented in FLINS2014 [27].

This paper is organized as follows. Section 2 introduces the new feature selection method for imbalanced data set by using the improved genetic algorithm. In Section 3, the proposed approach is combined with the SVM classifier for imbalanced data classification and evaluated using several benchmark datasets compared with the traditional method. The proposed approach is then applied to radio signal recognition of ground-air communication in Section 4. The paper is concluded in Section 5.

2. Feature Selection Method Based on Improved GA

GA is a kind of adaptive heuristic search algorithm to solve global optimization, which is formed in the process of simulating the genetic and evolution of bios in natural environment. GA-feature selection methods have shown to be very effective in balanced data sets[21-25]. But most of them are not suitable for imbalanced data. In this section,

an improved GA is applied for feature selection of imbalanced data sets with the aim at improving the recognition rate of minor class. The method is detailed in the subsequent sections.

2.1. Coding scheme in GA

Code scheme is the first step in using GA. In the present method, classical binary coding method is used, which is both simple and very effective. The length of the individual is the number of candidate features for any data set. For example, suppose a data set S has *n* features. A feature combination can be represented using a *n* bit string of '0' or '1', where '0' denotes that the corresponding feature has not been selected; on the other hand, '1' denotes that the corresponding feature has been selected.

2.2. Determination of fitness function in GA

The commonly used fitness function in GA-based feature selection and classification model is the total classification accuracy. However it is not suitable for the imbalanced data sets where the overall classification accuracy rate may be caused by the one from the major class, while the accuracy of the minor class may be very low. In order to enhance the accuracy of the minor class as the essential aim for handling imbalanced data classification problem, it is important to select the features which are beneficial to identify the minor class. This aim is related to the evaluation metrics to properly evaluate the effectiveness of such selection algorithms. Therefore, in the GA-based feature selection method, the fitness function may need to be modified accordingly to fulfill this requirement.

Traditionally, the most commonly used evaluation metrics is *accuracy* or *error rate*. Considering a basic two-class classification problem, a representation of classification performance can be formulated by a confusion matrix (contingency table), as shown in Table 1 below.

Table 1 Confusion matrix for performance evaluation

	Predicted Positive	Predicted Negative
Positive	True Positives (TP)	False Negatives (FN)
Negative	False Positives (FP)	True Negatives (TN)

In this paper, we assume that the minority class and the majority class are labeled positive and negative respectively. TP and TN are the samples' amount of minority class and majority class respectively under the condition of right classification. FN and FP are the samples' amount of minority class and majority class respectively under the condition of wrong classification.

Following this convention, the related concepts of Accuracy and G-mean [28] as performance evaluation measurements of classification are introduced as follows:

(1) Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy is the common evaluation standard of classification methods, however, it can be deceiving in certain situations and are highly sensitive to change in data. The accuracy metric in this case does not provide adequate information on a classifier's functionality with respect to the type of classification required [31].

(2) G-mean

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (2)$$

G-mean is the common evaluation standard for imbalanced data classification problem, it is the square root of the product of minority class's accuracy TP/(TP+FN) and majority class's accuracy TN/(TN+TP), when either of the values of product components is increasing, G-mean will be also increasing. So, G-mean can evaluate better the total classification performance, especially for the imbalanced data set classification.

According to the above discussion, G-mean is used as a basis for determining the fitness function of genetic algorithm in order to handle the unbalanced data classification while considering some other aspects as detailed below. Finally the new fitness function is defined as below:

$$f(x) = \alpha \cdot G\text{-}Mean + \beta \cdot \left(-\frac{|X|}{n} \right) \quad (3)$$

Here control parameters α, β are used to compromise the role of the number of features and the evaluation measurement G-mean played in the overall performance evaluation respectively with $\alpha + \beta = 1$. $|X|$ denotes the number of features in the selected feature subset *X* and *n* denotes the number of all the features. The first part of the right side shows that the larger the G-mean corresponding to the feature subset is, the greater the fitness function is. The second part presents that the

less the feature number is, the greater the fitness function. Users can set parameters α, β according to different problems and needs; this so-called Hurwicz approach attempts to strike a balance between the purpose of classification accuracy or feature dimension reduction by adjusting the control parameters α and β. In general, the first part is more important than the second part, so normally $\alpha > \beta$.

2.3. GA-based feature selection algorithm scheme

GA-based feature selection is described as follows and shown in Fig.1:

(1) Determine the encoding scheme and code scheme;

(2) Initialize the population;

(3) Determine the fitness function based on Eq. (3);

(4) Evaluate the fitness of the individual;

(5) Do genetic operations including selection, crossover and mutation if it does not meet the terminal condition;

(6) Repeat steps (4), (5) until the terminal condition is met.

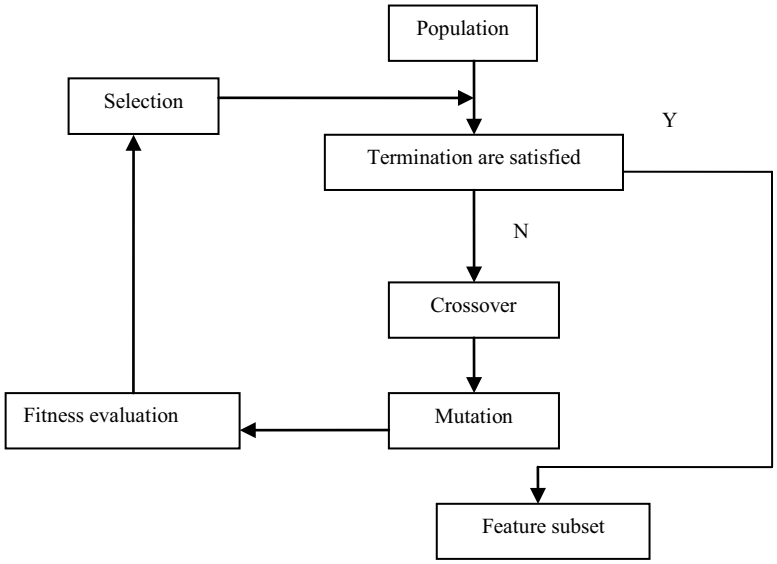

Fig.1 GA-based feature selection process

2.4. Classification using linear SVM

In this paper, the proposed feature selection algorithm will be evaluated over several binary imbalanced data sets from two aspects: feature subset size and the classification performance using linear SVM. Because the focus of this paper is on improving the feature selection so we choose the same classifier SVM for comparison purpose.

3. Evaluation Based on Benchmark Data

In the experiments below, we will compare the proposed improved GA based feature selection method with the traditional GA-feature selection method (which did not work effectively for the imbalanced data sets).

3.1. Experimental setup

In order to verify the validity of the proposed method, five imbalanced data sets from the UCI machine learning database [29] are selected. Among them, for the Satimage dataset Category 3 is assumed to be minority class, the rest samples as majority class. Specific parameters of each data set are shown in Table 2. Features (F) denotes the number of features, Sizes(S) denotes the size of the data set, Min and Max denotes the number of samples in the minority class and that in the majority class respectively, Target (T) is the class label of the minority,

and Ratio (R) denotes the ratio of the minority class to the majority class.

In the experiments, 5-fold cross validation with stratified sampling method is used in order to keep the original class distribution imbalanced. We hope to get higher classification accuracy rate in this paper, so we assume α =0.9, β = 0.1 in Eq.(3). GA and linear SVM toolboxes in the MATLAB 2012 version is used for the experimental studies. Specifically, some parameters in the GA are set as follows: the population size is 80, crossover probability is 0.7, mutation probability is 0.02 and the termination condition is that the variation of the fitness function is little in the recent 10 iterations.

In order to prevent the code of the best individual form changing or missing in the process of genetic operation such as crossover and mutation, elite reserved strategy is used.

Table 2 Summary of datasets

Data set	F	S	*Min/Max*	T	R
Satimage	36	6435	626/5809	3	9.28
Tic-tac-toe	9	958	332/626	Negative	1.89
Ionosphere	34	351	126/225	Bad	1.79
Breast	30	569	212/357	malignant	1.68
Sonar	60	208	97/111	R	1.14

3.2. Experimental results

The average of the experimental results using 5-fold cross validation with stratified sampling method are shown in Table 3 (for feature dimensionality reduction) and Table 4 (for minority accuracy), where GA denotes that the fitness function of the method shown in Eq. (4) and Improved GA (IGA) is the proposed method using the fitness function in Eq.(3).

$$f(x)=\alpha\cdot Accuracy+\beta\cdot\left(-\frac{|X|}{n}\right) \quad (4)$$

where parameters $\alpha,\beta,|X|,n$ are the same as those in Eq. (3).

Table 3 Number of features selected by GA and IGA on different data sets

Data set	Full	GA	IGA
Satimage	36	19	9
Tic-tac-toe	9	5	4
Ionosphere	34	16	14
Breast	30	14	8
Sonar	60	29	25
Average	34	17	12
Average dimensionality reduction		50%	64.71%

Table 4 Minority accuracies using the linear SVM after GA and IGA based feature selection

Data set	GA	IGA
Satimage	96.77% (±3.89%)	97.08% (±4.21%)
Tic-tac-toe	41.21% (±23.85%)	62.94% (±8.65%)
Ionosphere	88.83% (±5.27%)	92.03% (±6.35%)
Breast	98.10% (±1.99%)	99.05% (±1.30%)
Sonar	87.77% (±9.15%)	90.73% (±5.74%)
Average	82.54% (±8.83%)	88.37% (±5.25%)

Table 3 shows that average dimensionality reduction of IGA is 64.71%, but for GA is only 50%, that is to say, the improved method is better in terms of feature reduction. From Table 4 we can see that the improved

method can increase the recognition rate of the minority class effectively.

In conclusion, the proposed approach is better than the traditional GA-based feature selection method which did not consider the imbalance of data in the two aspects: the feature subset size and the recognition rate of the minority class, i.e. the proposed feature selection method can select the features which are favorable to identify the minority class.

4. Application to Radio Signal Recognition

Radio signal recognition of ground-air communication is very important to ensure flight safety. Normal signals of ground-air communication are far more than abnormal signals. So radio signal recognition of ground-air communication belongs to the imbalanced data classification problem. Correct recognition of the abnormal signals is particularly important in this problem. Feature selection is one of the key issues in signal recognition. It can improve the recognition performance by excluding redundant or irrelevant information. In the subsequent sections, the proposed feature selection method is applied to signal recognition of ground-air communication in order to select features which are favorable to identify the abnormal signals and improve the recognition rate of the abnormal signals as well.

4.1. Data sources

The sample data used in this application includes 260 normal signal samples and 30 abnormal signal samples, which are collected from the area near the airport using radio monitoring equipment with the assistance and support from Sichuan Provincial Radio Monitoring Station. According to the relationship between some statistical characteristics of the audio signal and the signal types, we extract 12 features from the speech signal. There are short-time average magnitude, short-time energy, short-time zero-crossing rate, short-time average zero-crossing rate, average energy-frequency- product, average energy-frequency- ratio, short-time average energy, normalized kurtosis, amplitude spectrum index, mean amplitude, amplitude variance, and amplitude sum. They are recorded as $f_1, f_2, \cdots, f_{12}$ respectively.

They are briefly introduced as follows[30], where $X(i)(i=1,2,\cdots,n)$ is audio signal amplitude value, n is the number of the samples.

1) Short-time average magnitude (M)

Short-time average magnitude is the parameter which measures the alteration of speech signal amplitude, its formula is

$$M = \sum_{i=1}^{n} |X(i)| \quad (5)$$

2) Short-time energy(S)

Short-time energy is defined as:

$$S = \sum_{i=1}^{n} [X(i)]^2 \quad (6)$$

3) Short-time zero-crossing rate (Z)

Short-time zero-crossing rate can be viewed as the simple measure of signal frequency. Short-time zero-crossing rate is defined as

$$Z = \frac{1}{2} \sum_{i=1}^{n} \left| \operatorname{sgn}[x(i)-] \right| \quad (7)$$

where sgn is symbolic function, see Eq. (8),

$$\operatorname{sgn}|x| = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \quad (8)$$

4) Short-time average zero-crossing rate ($\bar{Z}$)

Short-time average zero-crossing rate is defined as

$$\bar{Z} = \frac{1}{2n} \sum_{i=1}^{n} \left| \operatorname{sgn}[x(i)-] \right| \quad (9)$$

5) Short-time average energy (E_1)

Short-time average energy is defined as

$$E_1 = \frac{1}{n} \sum_{i=1}^{n} [X(i)]^2 \quad (10)$$

6) Average energy-frequency- product (P)

Average energy-frequency- product is the product of average energy and zero-crossing rate. Its formula is

$$P = E_1 \times Z \quad (11)$$

7) Average energy-frequency- ratio (R)

Average energy-frequency- ratio is the ratio of average energy to zero-crossing rate. Its formula is

$$R = E_1 / Z \quad (12)$$

8) Normalized kurtosis

All kinds of audio signals which are demodulated are divided into silence segments and sound segments. In order to well recognize silence segments and sound segments, we extract normalized kurtosis. Its formula is

$$K = \frac{\sum_{i=0}^{n} X(i)}{(\sum_{i=0}^{n} X^2(i))^2} \quad (13)$$

9) Amplitude spectrum index

Amplitude spectrum index is defined as

$$V = \frac{\sum_{i=0}^{n} X_i^4(\omega)}{\sum_{i=0}^{n} X_i^2(\omega)} \quad (14)$$

where $X_i(\omega)$ is obtained by the transformation of $X(i)$ into Fast Fourier Transformation(FFT).

10) Mean amplitude (E_2)

Mean amplitude reflects the change of audio amplitude. Its formula is

$$E_2 = \frac{\sum_{i=1}^{n} X(i)}{n} \quad (15)$$

11) Amplitude variance (D)

Amplitude variance is the stability of audio signal amplitude. Its formula is

$$D = \frac{\sum_{i=1}^{n} (E - X(i))^2}{n}, \quad i = 1, 2, \cdots, n \quad (16)$$

12) Amplitude sum (T)

Amplitude sum is the time-domain energy after eliminating the pulse signal interference. Its formula is

$$\mathrm{T} = \sum_{i=1}^{n} X(i) \quad (17)$$

4.2. Experimental results

5-fold cross validation method with stratified sampling method is used in the experiments and linear SVM is selected as the classifier. The experiment is repeated for 10 times. We select the average results as the experimental results. The proposed improved GA based feature selection method (denoted as IGA) is compared with the traditional GA based feature selection method (denoted as GA) is shown in Table 5.

Table 5 Experimental results by using GA and IGA

Algorithm	Number of features	Feature subset	Recognition rate of abnormal signal
GA	5	$f_3, f_7, f_8, f_9, f_{10}$	83.33%
IGA	2	f_8, f_{11}	100%

As illustrated from Table 5, the proposed method can effectively reduce the feature dimension and improve the recognition rate of abnormal signal significantly. It is interesting and promising for signal recognition of ground-air communication. It can not only improve the efficiency of recognition, but also can be much more conducive to the recognition of jamming signals in order to take further effective measures.

5. Conclusions

Aiming at the imbalance data classification problem, a feature selection method based on genetic algorithm is proposed through improving the fitness function. Experimental results on five UCI datasets show that the performance of the proposed method outperforms classic genetic algorithm based feature selection methods. It not only reduces the feature dimension effectively, but also improves the recognition rate of the minor class.

The proposed method was further applied to signal recognition of ground-air communication. The experimental results show that the proposed method can effectively reduce the feature dimension and improve the recognition rate of abnormal signal as well. It can provide important reference for radio monitoring personnel.

Further research is required for discussing the influence of different classifiers for the feature selection method. We only consider an abnormal interference signal in this paper, but the actual electromagnetic environment is very complex. We will consider other interfering signals in the future and also more complex data set for testing and evaluation.

Acknowledgements

This work is supported by National Science Foundation of China (Grant No. 61175055), Sichuan Key Technology Research and Development Program (Grant No. 2011FZ0051), Radio Administration Bureau of MIIT

of China (Grant No. [2011] 146), China Institution of Communications (Grant No. [2011] 051). The authors also gratefully acknowledge the helpful comments and suggestions of the teachers and students who come from Intelligent Control Development Center of Southwest Jiaotong University in China, which have improved the presentation.

References

1. Q. H. Zhu, Radio monitoring and communication investigation, *Beijing: people's posts and telecommunications publishing house*, 2005. (In Chinese)
2. D. Rivero, E. Femandez-Blanco, J. Dorado, et al. A New Signal Classification Technique by Means of Genetic Algorithms and kNN, *IEEE Conf. on Evolutionary Computation (CEC), Coruna, Spain*, 4(2011), pp. 581-586.
3. A. Ebrahimzadeh, R. Ghazalian, Blind digital modulation classification in software radio using the optimized classifier and feature subset selection, *Engineering Applications of Artificial Intelligence*, 24(2011), pp. 50-59.
4. M. Chen, Q. Zhu, Cooperative automatic modulation recognition in cognitive radio, *The Journal of China Universities of Posts and Telecommunications* , 17(2)(2010), pp. 46-52.
5. J. L. Yang, Z. Pei, L. Zou,et al. Feature extraction of radio signals using attribute reduction of formal concepts, *International Journal of Innovative Computing, Information and Control*, 7(6)(2011), pp. 3331-3343.
6. Z. H. Zhang, F. L. Ma, Z. Pei, Recognition of Aviation Interference Signal Based on K-means Clustering Algorithm, *The national conference on radio application and management*, 11(2013), pp. 106-111. (In Chinese)
7. F. Provost， T. Fawcett, Robust classification for imprecise environments, *Machine Learning*, 42(3)(2001), pp. 203-231.
8. N. Chawla, K. Bowyer, L. Hall, et al. SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, 16 (2002), pp. 321–357.
9. M. Galar, A. Fernandez, E. Barrenechea, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 99 (2011), pp. 1-22.
10. G. X. He, H. Han, W.Y. Wang, An over-sampling expert system for learning from imbalanced data sets, *International Conference on Neural Networks and Brain*, (2005), pp. 537-541.
11. H. Han, W. Y. Wang, B. H. Mao, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *International Conference on Intelligent Computing*, (2005), pp. 878-887.
12. R. Barandela, R. M. Valdovinos, J. S. Sánchez et al. The imbalanced training sample problem: under or over sampling, *Proc of International Workshops on Structural Syntactic and Statistical Pattern Recognition*, (2004), pp. 806-814.
13. L.M. Manevitz, M. Yousef, One-class SVMs for document classification, *Journal of Machine Learning Research*, 2 (2001), pp. 139–154.
14. H. J. Lee, S. Cho, The novelty detection approach for difference degrees of class imbalance, *Lecture Notes in Computer Science*, 4233 (2006), pp. 21–30.
15. G. Wu, Y. C. Edward, KBA: Kernel boundary alignment considering imbalanced data distribution, *The IEEE transactions on knowledge and data engineering*, 17(6)(2005), pp. 786-795.
16. T. Imam, K. M. Ting, J. Kamruzzaman, z-SVM: An SVM for improved classification of imbalanced data, *Australian Joint Conference on AI*, (2006), pp. 264-273.
17. H. J. Lee, S. Z. Cho, Focusing on non-respondents: Response modeling with novelty detectors, *Expert Systems with Applications*, 33(2), (2007), pp. 522-530.
18. Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, *ACM SIGKDD Explor. Newslett. (Special Issue on Learning from Imbalanced Datasets)*, 6 (1) (2004), pp. 80–89.
19. M. Wasikowski, X. W. Chen, Combating the small sample class imbalance problem using feature selection, *IEEE Transactions on Knowledge and Data Engineering*, 22(10)(2010), pp. 1388-1400.
20. R. Wang, K. Tang, Feature Selection for MAUC Oriented Classification Systems, *Neurocomputing*, 89(2012), pp. 39-54.
21. H. Frohlich， O. Chapelle, Feature selection for support vector machines by means of genetic algorithms, *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence， Sacramento， CA， USA*，(2003), pp. 142-148.
22. C. L. Huang, C. J. Wang, A GA-based feature selection and parameters optimization for support vector machines, *Expert Systems with Applications*, 31(2006), pp. 231-240.
23. Y. N. Liu, G. Wang, X. D. Zhu, et al. Feature selection based on adaptive multi-population genetic algorithm, *Journal of Jilin University(Engineering and Technology Edition)*, 41(6) (2011), pp. 1690-1693.
24. X. Zhou, Z. Pei, P. H. Liu, et al. A new method for feature selection of radio abnormal signal, *ICIC Express Letters*, 7(2)(2013), pp. 303-309.
25. Z. W. Ji, G. F. Wu, M. Hu, Feature Selection Based on Adaptive Genetic Algorithm and SVM, *Computer Engineering*, 35(14) (2011), pp. 200-202.
26. O. Soufan, D. Kleftogiannis, P. Kalnis, et al. DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm, *PloS one,* 10(2*)* (2015), pp. 1-23.
27. L. M. Du, Y. Xu, L. Q. Jin, Feature Selection for Imbalanced Datasets Based on Improved Genetic Algorithm, *Proc of the 11th International FLINS Conference on Decision Making and Soft Computing, Brazil,* (2014), pp.119-124.
28. Z. Y. Lin, Z. F. Hao, X. W. Yang, Effects of Several Evaluation Metrics on Imbalanced Data Learning, *Journal*

of South China University of Technology (Natural Science Edition), 38(4)(2010), pp. 147-155.
29. A. Asuncion， D. Newman. UCI repository of machine learning databases [DB/OL]. [2009-04-03]. http: //www. Ics. u ci. edu / ~ mlearn /MLRep-ository. Html.
30. J. Q. Han, L. Zhang, R. Tie, *Speech Signal Processing*, Beijing: Tsinghua University Press, (2004). (In Chinese)
31. H.B. He and Edwardo A. Garcia, Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, 21(9) (2009), pp. 1263-1284.