Ann. Data. Sci. (2015) 2(3):293–300
DOI 10.1007/s40745-015-0060-x

Feature Selection for Multi-Class Imbalanced Data Sets Based on Genetic Algorithm

Li-min Du[1,2] · **Yang Xu**[1] · **Hua Zhu**[1]

Received: 30 November 2015 / Revised: 8 December 2015 / Accepted: 8 December 2015 / Published online: 21 December 2015

Abstract This paper presents an improved genetic algorithm based feature selection method for multi-class imbalanced data. This method improves the fitness function through using the evaluation criterion EG-mean instead of the global classification accuracy in order to choose the features which are favorable to recognize the minor classes. The method is evaluated using several benchmark data sets, and the experimental results show that, compared with the traditional feature selection method based on genetic algorithm, the proposed method has certain advantages in the size of feature subsets and improves the precision of the minor classes for multi-class imbalanced data sets.

Keywords Feature selection · Genetic algorithm · Multi-class imbalanced data sets · Support vector machine

1 Introduction

Feature selection [1] is an important data pre-processing technique in data mining. Feature selection can be helpful when facing imbalanced data sets [2]. Many feature selection methods for two-class imbalanced data [3–5] have been proposed. However, there are not only many two-class imbalanced data sets in real-word applications, but also a lot of multi-class imbalanced data sets. For multi-class imbalanced data sets, lots of research fruits have been presented in classification [6–8], while few in feature selection [9].

✉ Li-min Du
dulimin@henu.edu.cn

[1] Intelligent Control Development Center, Southwest Jiaotong University, Chengdu 610031, China

[2] Pharmacy College of Henan University, Kaifeng 475004, China

Genetic algorithm is a heuristic search algorithm, using the reference of natural selection and genetic mechanism in living nature. Many feature selection methods based on genetic algorithm have been proposed [10–12]. These feature selection methods based on genetic algorithm are effective. But they did not consider the imbalance of data. So these methods are not suitable for imbalanced data. Du et al. [4] proposed a feature selection method for imbalanced data based on genetic algorithm. However, this method is suitable for two-class unbalanced data sets, it cannot be directly applied to multi-class problems.

In this paper, a new feature selection method for multi-class imbalanced data based on genetic algorithm is proposed through improving fitness function. We use the evaluation criterion EG-mean (G-mean's extension) for imbalanced data instead of total classification accuracy in order to select the features which are beneficial to identify small classes for improving the recognition rate of the minor classes. Support vector machine (SVM) is selected as the classifier due to its good classification performance. At last, this method is evaluated using several benchmark datasets.

The rest of this paper is organized as follows: Section 2 introduces the traditional feature selection methods and feature selection for two-class imbalanced data based on genetic algorithm. Section 3 presents the new feature selection method based on genetic algorithm for multi-class imbalanced data. The proposed approach is evaluated using several benchmark datasets in Sect. 4. The last section concludes the paper.

2 Related Work

2.1 Traditional Feature Selection Methods Based on Genetic Algorithm

The traditional feature selection methods [10,11] are the methods which do not consider the imbalance data. Their fitness function is

$$f\ (x) = \alpha \cdot Accuracy + \beta \cdot \left(-\frac{|X|}{n} \right), \tag{1}$$

where *Accuracy* is the total classification accuracy. Here the control parameters α, β are used to compromise the roles of the number of selected features and the *Accuracy* 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.

Accuracy is the common evaluation standard of classification methods. It is not an appropriate evaluation criterion for imbalanced data. For two-class imbalanced data, the samples of majority class are much more than the samples of minority class, if we classify all samples as major class, the accuracy is still very high, but the recognition rate of minor class is zero. This is obviously unreasonable.

2.2 Feature Selection for Two-Class Imbalanced Data

Du et al. [4] proposed a feature selection method for two-class imbalanced data based on genetic algorithm. This method improves the fitness function through using the evaluation criterion G-mean [13] for imbalanced data instead of total classification accuracy. Its fitness function is

$$f(x) = \alpha \cdot G - Mean + \beta \cdot \left(-\frac{|X|}{n} \right) \tag{2}$$

where α, β, $|X|$, n are the same as those of Eq. (1). G-mean is the common evaluation standard of unbalanced data set study, it is the square root of the product of minority class's accuracy TP/(TP+FN) and majority class's accuracy TN/(TN+TP), as shown in Eq. (3)

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{3}$$

where 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.

This feature selection method is effective for two-class imbalanced data, but it cannot be used for multi-class unbalanced data directly.

3 Feature Selection Method Based on Improved Genetic Algorithm for Multi-class Imbalanced Data

A new feature selection method based on genetic algorithm for multi-class imbalanced data sets is proposed through improving fitness function in this paper. This algorithm is described from the coding scheme, the determination of fitness function and the algorithm scheme respectively.

3.1 Coding Scheme

Code scheme is the first step in using the genetic algorithm. In this paper, 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 by 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.

3.2 Determination of Fitness Function

For the bi-class scenario, Kubat et al. [13] suggested the G-mean as the geometric means of recall values of two classes. Sun et al. [14] extend the G-mean definition to the geometric means of recall values of every class for multiclass imbalanced learning.

G-mean's extension is defined as follows for multiclass imbalanced learning:

$$EG - mean = \left(\prod_{i=1}^{k} A_i \right)^{1/k}, \tag{4}$$

where A_i is the ith class's accuracy.

So we define the fitness function of genetic algorithm for unbalanced data sets as Eq. (5)

$$f(x) = \alpha \cdot EG - mean + \beta \cdot \left(-\frac{|X|}{n} \right) \tag{5}$$

where α, β, $|X|$, n are the same as those of Eq. (1). The first part of the right side shows that the larger the *EG-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 is. 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 $\alpha > \beta$.

3.3 Algorithm Scheme

The algorithm is described as follows:

(1) Determine the encoding scheme and code scheme;
(2) Initialize the population;
(3) Determine the fitness function;
(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.

Feature selection process is shown in Fig. 1

4 Experiments

In this section, the proposed feature selection algorithm will be evaluated using several benchmark data sets from the UCI machine learning database [15] from two aspects: feature subset size and the classification performance using the classifier SVM. In the experiments, we will compare the proposed method in this paper with the traditional feature selection method based on genetic algorithm.

4.1 The Experimental Setup

In order to verify the validity of the proposed method, we select six benchmark datasets from the UCI machine learning database [15]. We chose two balanced data sets for

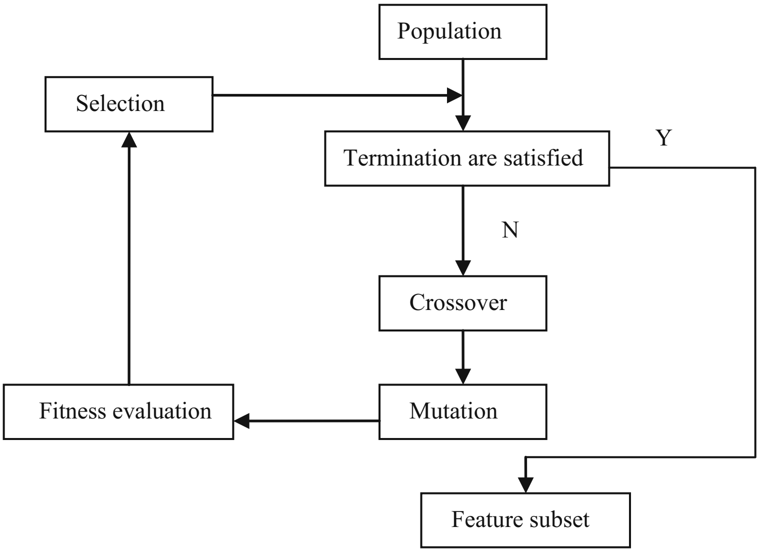

Fig. 1 Feature selection process

Table 1 Summary of data sets

Data set	Number of features	Number of samples	Number of classes	Class ratio
Wave	21	5000	3	1:1:1
Synthetic	60	600	6	1:1:1:1:1:1
Wine	13	178	3	71:59:48
Dermatology	34	366	6	111:71:60:48:48:20
Auto-mpg	7	391	3	244:79:68
Contraceptive	9	1473	3	629:511:333

comparison, the rest are imbalanced data sets. Some characteristics of these data sets are shown in Table 1.

In the experiments, 5-fold cross validation with stratified sampling method are used in order to keep the original class distribution imbalanced. We hope to get higher classification accuracy rate in this paper, so we assume $\alpha = 0.9$, $\beta = 0.1$. Making use of MATLAB language, the paper compiles the genetic algorithm program. 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 ten iterations in this paper. At the same time LIBSVM is used which are designed by Pro. Lin Chih-Jen, etc. RBF is adopted as kernel function in SVM and keep the defaults for the parameters.

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

Data set	Full	GA	IGA
Wave	21	10.0	9.4
Synthetic	60	28.6	33.2
Wine	13	3.6	3.0
Dermatology	34	11.0	10.8
Auto-mpg	7	1.4	1.4
Contraceptive	9	3.6	3.6

Table 3 Accuracies using the classifier SVM

Data set	GA	IGA
Wave(1)	0.8915	0.8886
Wave(2)	0.8555	0.8555
Wave(3)	0.7924	0.7852
Wave(global)	0.8468	0.8434
Synthetic(1)	0.4400	0.5300
Synthetic(2)	0.9300	0.7300
Synthetic(3)	0.4800	0.4900
Synthetic(4)	0.3700	0.4900
Synthetic(5)	0.6700	0.8000
Synthetic(6)	0.7200	0.6100
Synthetic(global)	0.6017	0.6083

4.2 Experimental Results and Discussions

The average of the experimental results using 5-fold cross validation with stratified sampling method are shown in Tables 2 and 3, where GA denotes that the fitness function of the method is shown as Eq. (1) and IGA is the method of this paper.

From Table 2 we can see that for two balanced datasets one set decreases in the size of feature subset, the other increases. That is, the proposed method in this paper has no obvious advantage in feature subset size for balanced datasets. For the rest four imbalanced data sets, two of them decrease, the other two did not change. In other words, the proposed method has certain advantages in the size of feature subsets for unbalanced datasets.

The per-class accuracies and the total accuracy of two balanced data sets are reported in Table 3. Wave(1),Wave(2),Wave(3) are the per-class accuracies of the wave data set respectively, Wave(global) denotes the global accuracy of the wave data set. The notation of the Synthetic data set is similar.

According to the results presented in Table 3, some of the per-class accuracies increase, while others decrease. The global accuracy of the Synthetic data set rises, while that of the wave data set reduces. Above all, the proposed approach compared with the traditional feature selection method based on genetic algorithm has no obvious advantage for balanced data sets.

Table 4 Minority accuracies using the classifier SVM

Data set	GA	IGA
Wine(small)	1.0000	1.0000
Dermatology(1)	1.0000	1.0000
Dermatology(2)	1.0000	1.0000
Dermatology(3)	0.9350	0.9550
Dermatology(4)	1.0000	1.0000
Dermatology(5)	1.0000	1.0000
Auto-mpg(small)	0.4929	0.5095
Auto-mpg(middle)	0.6200	0.6992
Contraceptive	0.3694	0.4201

The minority accuracies of imbalanced data sets are presented in Table 4. Wine(small) denotes the class of the Wine data set that has the smallest samples. Dermatology(1), Dermatology(2), Dermatology(3), Dermatology(4), Dermatology(5) are the classes that samples are 71, 60, 48, 48, 20 respectively. They are thought of as minority classes due to its relatively small samples. Auto-mpg(small) is the class of the Auto-mpg data set that has the smallest samples. Auto-mpg(middle) is the category that samples rank in the middle. Due to its relatively small samples, we see it as a small class. Contraceptive denotes the class of the Contraceptive data set that has the smallest samples.

From Table 4 we can see that some of the minority accuracies are the same and the results are 100 %. The rest of them increase. In conclusion, the proposed approach compared with the traditional feature selection has certain advantages in the minority accuracies. The proposed method can choose the features that conducive to identify the small classes.

5 Conclusions

A new feature selection method based on genetic algorithm is proposed through improving the fitness function for multi-class imbalanced data sets. Experimental results on several UCI data sets show that the proposed method has certain advantages in the size of feature subsets for imbalanced datasets. This feature selection method can select the features which are favorable to identify the minority classes.

Further research is required for discussing the problem of parameter optimization for SVM and the influence of different classifiers for the feature selection method. We will consider more complex data sets for testing and evaluation and also the practical application background in the future.

Acknowledgments This work is supported by National Science Foundation of China (Grant No. 61175055), youth science fund of China (Grant No. 61305074). 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. Guyon I, ElisseefF A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
2. Chawla NV, Japkowicz N, Kotcz A (2004) Editorial: special issue on learning form imbalanced data sets. SIGKDD Explor 6(1):1–6
3. Maldonado S, Weber R, Famili F (2014) Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Inform Sci 286:228–246
4. Du LM, Xu Y, Jin LQ (2014) Feature selection for imbalanced datasets based on improved genetic algorithm. In: Proc of the 11th International FLINS conference on decision making and soft computing, Brazil, pp 119–124
5. Yin LZ, Ge Y, Xiao KL et al (2013) Feature selection for high-dimensional imbalanced data. Neurocomputing 105:3–11
6. Cerf L, Gay D, Selmaoui-Folcher N et al (2013) Parameter-free classification in multi-class imbalanced data sets. Data Knowl Eng 87:109–129
7. Fernández A, López V, Galar M et al (2013) Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. Knowl Based Syst 42:97–110
8. Tang K, Wang R, Chen T (2011) Towards maximizing the area under the ROC curve for multi-class classification problems. In: Proceedings of the 25th AAAI conference on artificial intelligence (AAAI 2011), San Francisco, pp 483–488
9. Wang R, Tang K (2012) Feature selection for MAUC oriented classification systems. Neurocomputing 89:39–54
10. Frohlich H, Chapelle O (2003) Feature selection for support vector machines by means of genetic algorithms. In: Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, pp 142–148
11. Huang CL, Wang CJ (2006) A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl 31:231–240
12. Zhou X, Pei Z, Liu PH et al (2013) A new method for feature selection of radio abnormal signal. ICIC Express Lett 7(2):303–309
13. Kubat M, Holte R, Matwin S (1998) Machine learning for the detection of oil spills in satellite radar images. Mach Learn 30:195–215
14. Sun Y, Kamel MS, Wang Y (2006) Boosting for learning multiple classes with imbalanced class distribution. In: Proceedings of the international conference on data mining, pp 592–602
15. Asuncion A, Newman D (2007) UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html